Multi-modal Analysis for the Automatic Evaluation of Epilepsy

BSc. MSc. David Esteban Ahmedt Aristizabal

This dissertation was submitted in fulfilment of
the requirement for the degree of
Doctor of Philosophy

School of Electrical Engineering and Computer Science
Science and Engineering Faculty
Queensland University of Technology

2019

Statement of Original Authorship

In accordance with the requirements of the degree of Doctor of Philosophy in the School of Electrical Engineering and Computer Science, I present the following thesis entitled,

Multi-modal Analysis for the Automatic Evaluation of Epilepsy

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

QUT Verified Signature

BSc. MSc. David Esteban Ahmedt Arizstabal

30/08/19

Acknowledgements

I would like to express my gratefulness and love to my wife, Estefania Vargas, without her continuous support this work would not have been possible. I am also very grateful to my parents and brother, who have shown me the value and power of patience and perseverance.

I am very grateful to my advisors, Prof. Clinton Fookes, Dr. Sasha Dionisio, Dr. Kien Nguyen, Prof. Sridha Sridharan and Dr. Simon Denman. They gave me their support to address this work and their contributions helped to improve this thesis. I also want to thank Prof Stiefelhagen because he opened me the doors of the Computer Vision for Human-Computer Interaction Laboratory at Karlsruhe Institute of Technology. Dr Saquib Sarfraz and Prof Stiefelhagen support and advice have been highly important to me for a research and development career. I am also very grateful with the people (clinicians, patients and volunteers) from Mater Hospital. Without their help and disposition to collaborate by providing their recording sessions, the development of this thesis would be impossible.

Finally, I want to thank the Queensland University of Technology for its financial support during the development of this thesis through the scholarship QUT Postgraduate Research Award.

Abstract

Epilepsy is one of the most prevalent neurological disorders affecting 65 million people worldwide. Almost 30 per cent of people with epilepsy do not respond to medication, and surgery provides a chance for seizure freedom. During seizures, patients with epilepsy may exhibit stereotypical behaviour or motor manifestations. There are multiple types of epilepsy that all have different symptomatology and many forms of epilepsy have characteristic movements during a seizure. Seizure semiology refers to the analysis of these clinical signs that is reflective of the integration of connected brain neural networks. Along with the brain electrical activity and neuroimaging recordings, seizure semiology provides localising information of the brain networks affected, which in turn allows for a successful surgery.

Epileptologists often spend a lot of time analysing videos and EEGs to unravel the underlying epileptic network. Seizure semiology is still widely interpreted by visual inspection, and years of training and experience are required. The incorporation of quantitative methods would assist in developing and formulating a diagnosis in situations where clinical expertise is unavailable. Automated analysis of semiology can also provide standard assessment among evaluators, reducing the subjective nature of clinical decisions. Vision-based systems are still largely unexplored, in which most of them lack integrated approaches that analyse simultaneously semiology arising from multiple body parts, because of many complex natural clinical settings. Examples include attempts to analyse physical movements when the patient is covered and room lighting is inadequate. Recent advances in artificial intelligence (AI), especially in the burgeoning area of deep learning, offer an exciting avenue to overcome these challenges. However, research in evaluating patients' behaviour during a seizure has not been addressed.

This research aims to increase the semiology understanding by developing automated methodologies which test motion features and constitute a reliable source of information for early diagnosis and support of physicians' clinical criteria. The automated analysis of epilepsy is a greatly needed topic which aims to evaluate advanced computer vision and machine learning algorithms for the purpose of presurgical assessment of epilepsy. These techniques can be used to support neurologists to identify both types of epilepsy, as well as better understand the temporal evolution of seizures, from their onset through to termination. The robustness and flexibility of each methodology proposed to assess semiology exploits the existing camera monitoring infrastructure to modernise Epilepsy Monitoring Units without incurring an additional cost. Additionally, techniques we have

developed for computer-aided diagnosis can also be potentially useful in the evaluation of broader neurological diseases that experience movement disorders such as stroke and dementia.

The overarching innovation of this thesis is the investigation of how deep learning can be exploited in the presence of limited data and complex clinical conditions to support the diagnosis of epilepsy. The following are the main original contributions of this research: (1) A new modelling approach to quantify facial semiology; (2) The first effort on multi-modal approaches to quantify and classify seizures using clinical manifestations from the face, head, upper limbs, hands and finger movements; (3) A novel method to quantify and classify mouth semiology based on detailed information from 3D face reconstruction in the epilepsy scenario; (4) The design of a robust architecture capable of modelling two relevant diagnoses: an identification system of aberrant epileptic seizures and the first marker-free system that differentiate overlaying clinical features during epileptic and non-epileptic seizures; (5) The first of its kind computer-aided system that captures the stepwise progression of semiology as a flow of signals enabling visualisation and information from signal processing techniques; and (6) Preliminary investigations towards the development of a light-weight deep network to effectively learn and model discriminative temporal patterns from EEG sequential data.

This research is unique and provides important supplementary and unbiased data to assess semiology. It is a vital complementary resource in the era of seizure-based detection through electrophysiological data. This thesis successfully demonstrates a basis for ongoing significant breakthroughs in the field of epilepsy.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms / Abbreviations

CNN	Convolution Neural Networks
EEG	Electroencephalography or scalp EEG
EMU	Epilepsy Monitoring Units
EZ	Epileptogenic zone. Region of cortex that can generate epileptic seizures
ETLE	Extra-Temporal Lobe Epilepsy
Ictal	Refers to recordings during a seizure
Interictal	Refers to the period between seizures
LOSO-CV	Leave-one-subject-out Cross-Validation
LSTM	Long-Short Term Memory Network
MTLE	Mesial Temporal Lobe Epilepsy
RNN	Recurrent Neural Networks
SEEG	Stereotactic electroencephalography or depth EEG
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Motivation

Neurological diseases affect millions of people around the world. In particular, epilepsy is among the most common serious brain disorders and influences approximately 65 million people worldwide, with 2.4 million new cases diagnosed each year globally. Epilepsy is a disorder of the brain characterised predominantly by a paroxysmal disturbance of brain electrical activity, *i.e.*, transient manifestations of abnormal, excessive or synchronous neuronal activity in the brain [Fisher et al., 2005]. Epilepsy can occur at all ages and is characterised by a variety of presentations and causes. Epilepsy can have negative impacts on the social, psychological, and physical interaction of the patients. Thus, the early diagnosis of epilepsy and its monitoring are highly important and necessary to improve the patient's quality of life.

Epileptic seizures are recognisable by a transient change in clinical state, *e.g.* behaviour modification and motor signs, known as semiology [Chauvel and McGonigal, 2014]. There are multiple types of epilepsy which all have different symptomatology and many forms of epilepsy have characteristic movements during a seizure, allowing an understanding of the underlying brain networks. The variety of signs are considered as discriminative features of each epilepsy type such as jerking, spasm or posturing, head turning, facial expressions and hand movements. Figure 1.1 illustrates selected examples of clinical manifestation during a seizure.

With almost 30–40% of patients experiencing partial epilepsy being nonresponsive to medication, epilepsy surgery provides a chance for seizure freedom [Wiebe et al., 2001]. Along with the electrophysiological and neuroimaging recordings, seizure semiology constitutes a crucial set of clues, which provides localising information of the brain networks affected, enabling the progression to successful surgery in patients who are drug-resistant [Noachtar and Peters, 2009]. Although semiology is a major component of epilepsy evaluation, the understanding of the neural bases of semiology has advanced surprisingly insufficiently over recent years [Chauvel and McGonigal, 2014]. The study of video monitoring recordings is, to a certain extent, dependent on the experience and training of the clinician and the interpretation can differ from physician to physician, and between

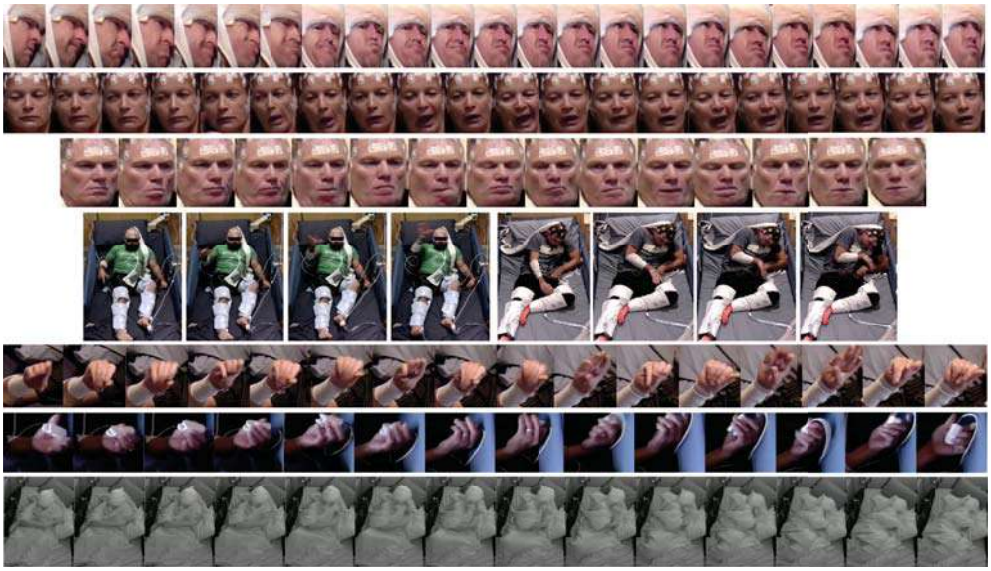


Figure 1.1 Selected video sequences that capture motor seizures from patients with epilepsy. Motions in the face, the head and upper limbs, hand and fingers, and body turning recorded in the epilepsy research dataset.

cases [Tufenkjian and Lüders, 2012]. Pre-surgery misdiagnosis takes place due to the inherent complexity and difficulty in characterising seizure patterns and to electrode implantation that could mislocalise the precise regions affected [O'Muircheartaigh and Richardson, 2012]. This produces particular challenges in the clinical evaluation of epilepsy, *e.g.*, the acknowledged difficulty in some cases of distinguishing between temporal and extra-temporal lobe seizures [O'Brien et al., 2008]. Semiology is a highly valuable tool, and significant advancements are essential to improve the accuracy and repeatability of its use. Because surgery misdiagnosis reaches a rate of 30%, and more than one-third of all epilepsies are poorly understood [De Tisi et al., 2011], there is an interest worldwide to develop techniques that can support the diagnostic precision by enabling the capture and assessment of objective information from video recordings.

In the light of the prevalence of motor and behavioural seizure manifestations in epilepsy, it is necessary to develop methodologies for the automatic evaluation of these clinical signs. Considering the recent breakthroughs in computer vision and machine learning, the visual inspection of semiology has the potential to be significantly automated. Automated analysis of semiology, *i.e.*, detection, quantification, and recognition of clinical signs, could help to increase the diagnostic precision by standardising the assessment evaluation among evaluators and identifying features that are unambiguous [Cunha et al., 2013, Bonini et al., 2014]. Vision-based approaches have shown promise in extracting information from seizure semiology to help define seizures [Pediaditis et al., 2012b]. However, this is a challenging and largely underdeveloped field due to a lack of datasets and the highly complex natural clinical settings [Pediaditis et al., 2012b].

Conventional automated assessments for medical conditions, including semiology, are heavily reliant on hand-crafted techniques where the performance is severely affected by uncontrolled hospital conditions such as patient position, illumination changes and motion discontinuities. Motivated by recent advances in human motion understanding based on computer vision and deep learning techniques [LeCun et al., 2015, Razzak et al., 2018, Bulat and Tzimiropoulos, 2017b, Cao et al., 2017, Simon et al., 2017, Feng et al., 2018a], *this thesis addresses the problem of modelling the mechanisms of semiological differences in epileptic patients with objective and quantitative motion analysis*. It is expected that these techniques could solve the current limitations of automated analysis of semiology.

Semiology is defined as the manifestations of epilepsy and the study of semiology analyses the progression of clinical symptoms in a stepwise/temporal progression. Thus, a single sign in isolation is far less informative [Bonini et al., 2014]. Therefore, this thesis argues that multiple clinical signs should be considered in combination from a *multi-modal perspective* to support clinical experts with the assessment criteria. Among all signs and symptoms studied during a seizure, the patient's facial expression, head and upper limbs motions, along with hand and fingers movements seem to be a promising source of information and can serve as a potentially useful tool for the patient evaluation as discussed with the co-investigators from the Mater Advanced Epilepsy Unit, Brisbane, Australia.

This research is a significant step forward towards the development of computer-based methodologies for monitoring and diagnosis of patients with epilepsy by considering multiple types of semiology. This research is unique and provides important supplementary and unbiased data to assess semiology. It is a vital complementary resource in the era of seizure-based detection through electrophysiological data. The computer-aided-diagnosis will enable us to identify consistent features that might help to reduce the failure rate of epilepsy surgery. The system will save time, improve patient safety and provide objective clinical analysis to assist with clinical decision making. Additionally, the technology could also be potentially useful in the evaluation of broader neurological diseases that experience movement disorders such as stroke and dementia.

1.2 Research aims

Diagnosis of epilepsy depends on the observation of symptoms which requires objective neurological measures. *This thesis aims to increase semiology understanding by developing automated methodologies which evaluate motion features and constitute a reliable source of information for early diagnosis and support of physicians' clinical criteria.* There is a great interest to understand and model different phenomena and symptoms from epileptic patients. This research develops vision-based approaches relying on the existing camera monitoring infrastructure. The epilepsy evaluation uses the characterisation of human motion patterns captured through videos (a current fixed high definition single camera in the hospital room) recorded during non-invasive (scalp electro-encephalography EEG) and surgical monitoring (stereo-electro-encephalography SEEG) at the Mater

Centre for Neuroscience, Brisbane, Australia. To contribute to this main aim, the ambitious specific aims of this thesis are recognised by the following five (5) research aims illustrated in Figure 1.2.

Aim 1: Analysis and identification of facial semiology

Analysis and development of quantitative methods that characterise and identify facial semiology. We address limitations of existing computer-based analytical approaches of epilepsy monitoring, where facial movements have largely been ignored. This an area that has seen limited advances in the literature. This aim addresses the challenge of capturing and distinguishing between facial modification during seizures and natural/random facial expressions during the patient's monitoring.

Aim 2: Multi-modal analysis of semiology and classification of epilepsy types

Analysis and evaluation of multi-modal approaches that quantify different motor manifestations simultaneously and classify epilepsy type. A single semiology sign in isolation may not be helpful, but rather the integration of various clinical manifestations would support clinical decision-making. Semiology in the form of facial expressions, body, hand and finger movements are considered to evaluate the performance of a fusion and a hierarchical approach. This research aim addresses the clinical hypothesis that similar semiology involves neuronal activity within the same specific brain networks and are sufficient to categorise seizures.

Aim 3: Analysis of behaviour of interest

Evaluation and exploration of automated strategies to analyse the key isolated sign of mouth semiology. The evaluation of one clinical sign such as mouth semiology is a necessity to reduce the problem dimensionality, as more than 40 possible facial descriptions can be linked to epilepsy. However, current computer vision based techniques are unable to accurately quantify mouth motions, which are heavily examined by neurologists to distinguish seizure types. This process will make the analysis of this behaviour of interest objective, thus facilitating the disease diagnosis and treatment.

Aim 4: Analysis and identification of aberrant epileptic seizures

Investigation and identification of aberrant or unusual epileptic behaviours. Aberrant identification methods can be very useful in identifying interesting, concerning, or unknown events that deviate from the majority of examples recorded in the hospital. The identification of unusual semiology will alert clinicians to consider different diagnostic choices and to progressively update automated identification systems with new semiologies. We attempt to group epileptic seizures into a best-fit model, using a template of known semiology or stereotypical behaviours in a form of libraries. The libraries store feature representations of the motion exhibited by the patient during the recorded seizure. A new patient presented to these libraries, where the semiology does not fit the status-quo of learned information, would be considered as having dissimilar findings or aberrant semiology.

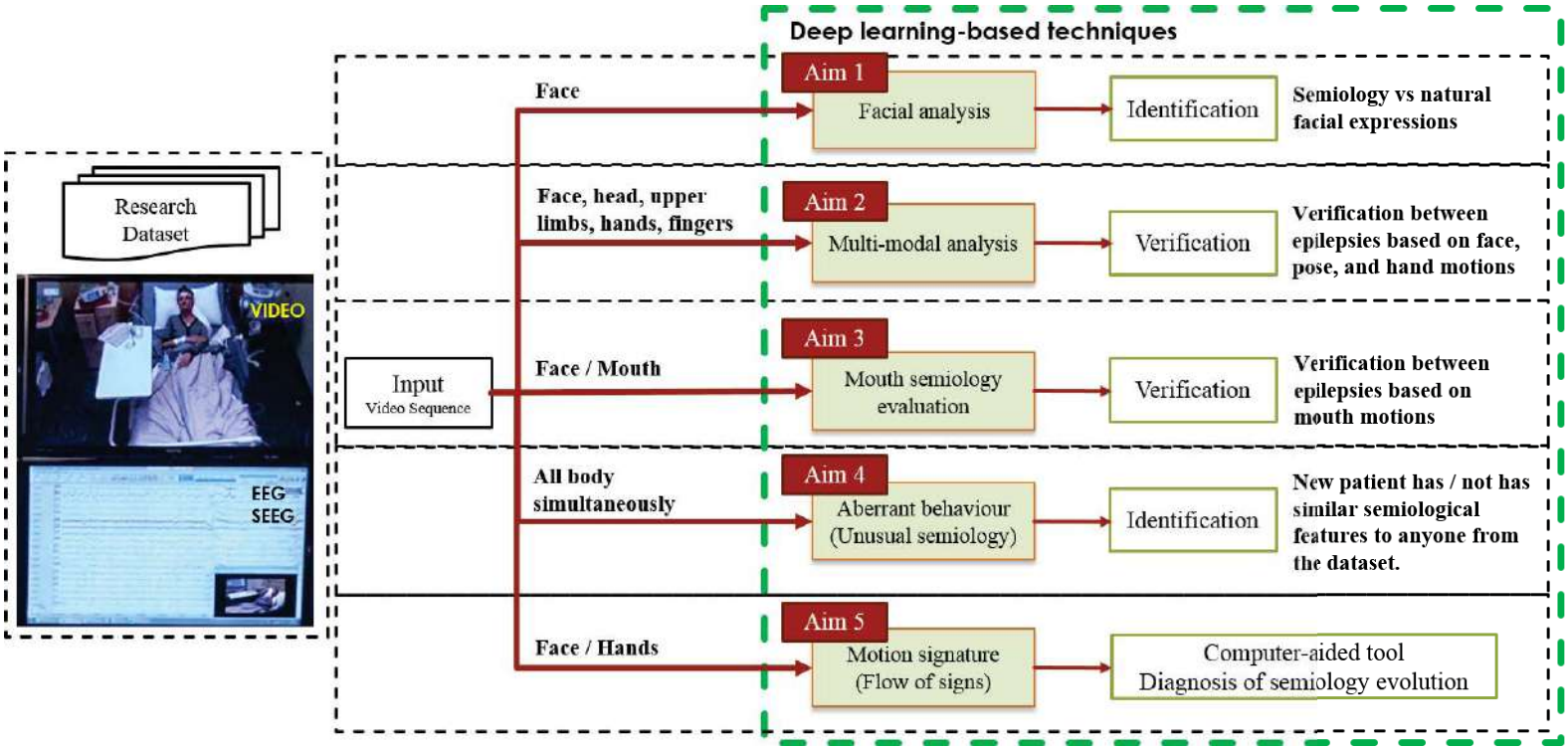


Figure 1.2 Research aims and their interconnections. A new research dataset of semiology is created. From the video corpus, we address five (5) research aims based on computer vision and deep learning architectures: (1) identification of facial semiology; (2) multi-modal classification of epilepsies; (3) mouth semiology evaluation; (4) identification of aberrant semiology seizures; and (5) analysis of the step wise progression of semiology.

Aim 5: Analysis of seizure evolution through motion signatures

Development of a computer-aided system capable of capturing the dynamic changes in semiology over an entire seizure which we call a motion signature. The analysis of seizure evolution aims to identify the presence or absence of certain clinical features. Nevertheless, the automated representation of the evolution of semiology has not been properly investigated. This system allows the analysis of the stepwise progression of stereotypical manifestations from the representation of semiology as a flow of signs and the correlation between body parts. We show how this representation can be used by providing quantitative information from signal processing techniques such as the periodicity, speed of the motion, dominant signs and the order of signs.

1.3 Structure of this work

The chapters of this thesis are structured according to the interaction of the proposed methodologies to support the evaluation of epilepsy by addressing each research aim and providing quantitative information from video recordings. This organisation is illustrated in Figure 1.3. The left side of Figure 1.3 shows the process conducted by clinical experts for the pre-surgical evaluation of epilepsy by understanding the brain functions and pathophysiological processes using non-invasive (EEG) and surgical monitoring (SEEG) techniques. Video recordings from EEG and SEEG monitoring are used to develop the research epilepsy dataset to assess our methodologies (Chapter 3), which mainly involves the quantification and identification of clinical manifestations between epilepsies (Chapters 4,5,6,7), the verification between seizure disorders (epileptic and non-epileptic seizures) (Chapter 7), and the analysis of seizure evolution (Chapter 8). In this chapter, we also provide experiments for analysing brain electrical activity based on deep learning. Each chapter includes details of the methods and algorithms used to address each research aim including an extensive discussion about the experiments and results. The specific information of the chapters is explained in detail as follows:

Chapter 2 provides a basic knowledge of epilepsy and analyses the state-of-the-art of automated approaches proposed to support the diagnosis of semiology (quantification of clinical manifestation and its application to assess epilepsy types). This chapter also describes information on traditional machine learning and computer vision techniques and introduces deep learning concepts and its successful preliminary implementation in healthcare and epilepsy.

Chapter 3 introduces the epilepsy dataset developed to support this thesis, including the technical description of the data from video-EEG and video-SEEG monitoring. Details of public human motion analysis datasets considered in the research experiments are also included.

Chapter 4 presents a methodology that provides quantitative motion information from facial semiology to support the assessment of epilepsy by distinguishing between facial expressions during a seizure and random facial expressions during a patient's monitoring (*Research Aim 1*).

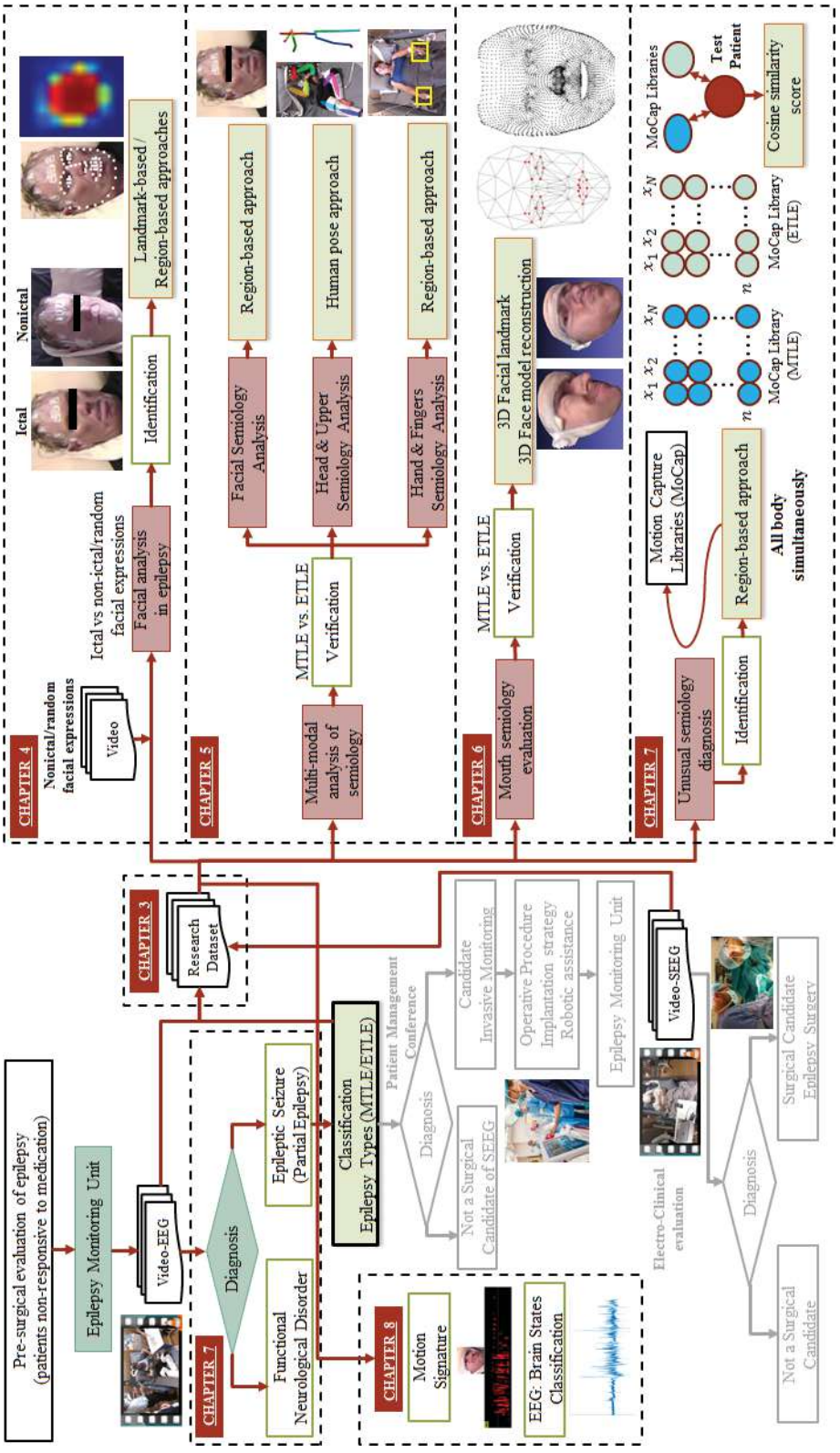


Figure 1.3 Chapters and their interconnections in the clinical evaluation of epileptic patients.

Chapter 5 introduces multi-modal strategies that assess semiology from the head, upper limbs, hands and fingers movements in order to quantify and classify epilepsy types. Multiple semiology signs are considered in combination to support clinical experts with the assessment criteria (*Research Aim 2*).

Chapter 6 provides a methodology to capture and assess the variety of mouth semiology, which is useful to distinguish epilepsy types (*Research Aim 3*).

Chapter 7 proposes an architecture flexible to analyse two specific and relevant diagnoses: identification of aberrant or unusual epileptic seizures (*Research Aim 4*) and recognition of functional neurological disorders and epileptic seizures which are often misdiagnosed because of their over-reliance on specific clinical features.

Chapter 8 suggests an efficient, in both computation and architecture, computer vision approach to capture motion signatures of face and hand semiology, to provide a diagnostic tool to clinicians to evaluate the evolution of clinical manifestations in patients with epilepsy (*Research Aim 5*). This chapter also reveals preliminary evidence of the benefits of deep networks to learn and model discriminative temporal patterns from EEG sequential data.

Chapter 9 summarises and analyses aspects of semiology to support the evaluation of epilepsy, the methodologies proposed and the main experimental results. This chapter also presents an outlook on future research in the area of epilepsy including the electro-clinical analysis, *i.e.*, methodologies that integrate both semiology and electro-graphic patterns.

The appendix provides information about the ethical clearance application and introduces relevant collaborative work adapting methodologies proposed in this thesis to support medical research.

1.4 Contributions of this research

The overarching innovation of this thesis is the investigation of how deep learning can be exploited in the presence of limited data and complex conditions to support the diagnosis of epilepsy. The development of artificial intelligence techniques enable the improvement of the automated assessment of semiology and provides well-informed and timely decisions. The diagnosis will support clinical decisions in critical cases of epilepsy assessment. This research effectively exploits existing camera infrastructure to modernise epilepsy monitoring units without incurring an additional cost, by automating manual analysis tasks. The use of vision-based systems in clinical practice is still not widespread, and the outcomes reported in this thesis will encourage the research community to improve the current understanding of epilepsy from automated methodologies. New pathways will be created to address the development of automated methodologies that can jointly learn across electro-clinical data to assess epilepsy by attributing subsets of brain networks related to seizure events.

The following are the main contributions presented in this thesis:

- A novel quantification process based on deep learning is proposed as a new modelling approach to discriminate facial semiology from spontaneous expressions during routing monitoring, which is an area that has seen limited advances in the literature.
- The first effort on a fusion strategy of spatio-temporal features from face, head and upper limbs motions to classify patients with mesial temporal (MTLE) and extra-temporal lobe (ETLE) epilepsy.
- We propose the first hierarchical multi-modal approach to assess seizures using clinical manifestation from the face, head, upper limbs, hands and fingers motions where each semiology is solved independently and the results are analysed with clinical experts to evaluate a general condition of the symptomatology of the patient. The most obvious distinction of our contribution is the modularity and flexibility of the system, allowing to enhance the performance by adapting each module of human motion quantification with new, more accurate and robust computer vision approaches.
- A novel vision-based system to analyse the isolated mouth semiology is proposed. The system exploits the detailed 3D face reconstruction from sequences of 2D images, making this, the first-of-its-kind experiment in the epilepsy research area. The introduction of a 3D perspective allows capturing rich information to analyse diverse mouth motions, which are challenging to model with traditional facial expression systems.
- The first detection system for aberrant or unusual epileptic seizures is introduced, using a simple strategy of motion capture (MoCap) libraries and average similarities to pre-learned semiology (past patient cases stored in health records with stereotypical behaviour). This contribution enhanced the analysis of semiology by enabling active learning to progressively update the quantification strategy of seizures with new semiology features detected within the system.
- A design architecture which captures all body motion simultaneously is used to investigate the first application of vision-based techniques to differentiate overlaying clinical features during epileptic and non-epileptic seizures. This study provides preliminary evidence that advances in video analytics and human action understanding can support the analysis of epileptic seizures and manifestations of psychological distress.
- The first computer vision approach is presented to capture the stepwise progression of semiology as a flow of signals enabling visualisation and information from signal processing techniques. The compact image representation illustrates the dynamic changes in semiology (spatial and temporal) over an entire seizure which we call a motion signature. The motion perspective provides relevant features to the physician and a way to intuitively assess the patient's movement, which is helpful for proper disease management.
- A light-weight deep network is developed to provide insights into the benefits of recurrent models to analyse electrographic data such as temporal patterns of EEG sequential data.

The experiments addressed in this thesis indicate that it is possible to design computational systems to assess semiology, enabling the easy and fast deployment of this technology in several hospitals.

1.5 Publications emerging from the development of this research

1.5.1 Journal Publications

The journal ranking (Q1, Q2, Q3 and Q4) is a measure of scientific influence of scholarly journal that accounts for both the number of citations received by a journal and the importance or prestige of the journal where such citations come from. This indicator is based on the Scimago Journal Rank (https://www.scimagojr.com/). The journal impact factor (IF) is a measure reflecting the yearly average number of citations to recent articles published in that journal, which can be found in the website of each journal. The h-index (H) expresses the journal's number of articles (h) that have received at least h citations. It quantifies both journal scientific productivity and scientific impact. This information is also available in the Scimago Journal Rank website. The h5-index (H5) is the h-index for articles published in the last five complete years, and the information is available on the metrics of Google Scholar website.

Epilepsia [Q1(2018), IF: 5.067, H: 174, H5: 73] is the world's leading journal of original scientific research in epileptology focusing on all aspects of the multidisciplinary field of epilepsy and is the official journal of the International League Against Epilepsy (ILAE).

IEEE Journal of Biomedical and Health Informatics [Q1(2018), IF: 3.85, H: 104, H5: 56] publishes the latest advances and practical applications of biomedicine where computer technologies intersect with health and healthcare.

Seizure [Q1(2018), IF: 2.839, H: 76, H5: 36] is the European Journal of Epilepsy owned by Epilepsy Action (the largest member-led epilepsy organisation in the UK), which aims to share and disseminate knowledge between all disciplines that work in the field of epilepsy.

Epilepsy & Behavior [Q1(2018), IF: 2.600, H: 91, H5: 43] is the fastest-growing international journal uniquely devoted to the rapid dissemination of the most current information available on the behavioural aspects of seizures and epilepsy.

- **D. Ahmedt-Aristizabal**, C. Fookes, S. Dionisio, K. Nguyen, J.P.S. Cunha, S. Sridharan, Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey, *Epilepsia*, 58 (2017) 1817-1831.
- **D. Ahmedt-Aristizabal**, C. Fookes, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, Deep facial analysis: A new phase I epilepsy evaluation using computer vision, *Epilepsy & Behavior*, 82 (2018) 17-24.

- **D. Ahmedt-Aristizabal**, C. Fookes, S. Denman, K. Nguyen, T. Fernando, S. Sridharan, S. Dionisio, A Hierarchical Multi-modal System for Motion Analysis in Epileptic Patients, *Epilepsy & Behaviour*, 87 (2018) 46-58.
- **D. Ahmedt-Aristizabal**, C. Fookes, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, Aberrant Epileptic Seizure Identification: A Computer Vision Perspective, *Seizure*, 65 (2019) 65-71.
- **D. Ahmedt-Aristizabal**, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, C. Fookes, Understanding Patients' Behaviour: Vision-based Analysis of Seizure Disorders, *IEEE Journal of Biomedical and Health Informatics*, (2019). Accepted (DOI: 10.1109/JBHI.2019.2895855).

1.5.2 International Conference Publications

EMBC conference is the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS) which is the world's largest annual Biomedical Engineering forum with over 3000 attendees from over 73 countries. The conference covers a broad range of topics from cutting-edge biomedical and healthcare technology research and development to clinical applications, biomedical education and entrepreneurship.
ISBI conference is the IEEE International Symposium on Biomedical Imaging, which is a joint initiative from the IEEE Signal Processing Society (SPS) and the IEEE Engineering in Medicine and Biology Society (EMBS) dedicated to mathematical, algorithmic, and computational aspects of biological and biomedical imaging, across all scales of observation.
ESA conference is the Annual Scientific meeting of the Epilepsy Society of Australia, which is the most important conference in Epilepsy in Australia and South East Asia.

- **D. Ahmedt-Aristizabal**, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, C. Fookes, Deep Motion Analysis for Epileptic Seizure Classification, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2018, 3578-3581. (Oral Presentation).
- **D. Ahmedt-Aristizabal**, K. Nguyen, S. Sridharan, C. Fookes, Deep Classification of Epileptic Signals, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2018, 332-335. (Poster Presentation).
- A. Pemasiri, **D. Ahmedt-Aristizabal**, K. Nguyen, S. Sridharan, S. Dionisio, C. Fookes, Semantic Segmentation of Hands in Multimodal Images: A New Region-based CNN Approach, *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019. (Poster Presentation).
- **D. Ahmedt-Aristizabal**, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, C. Fookes. Vision-Based Mouth Motion Analysis in Epilepsy: A 3D Perspective, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2019. (Oral Presentation).

- **D. Ahmedt-Aristizabal**, M. Saquib Sarfraz, S. Denman, C. Fookes, S. Dionisio, R. Stiefelhagen, Motion Signature for the Analysis of Seizure Evolution in Epilepsy, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2019. (Oral Presentation).
- **D. Ahmedt-Aristizabal**, C. Fookes, K. Nguyen, S. Dionisio, Deep Facial Semiology Analysis For Mesial Temporal Lobe Epilepsy Evaluation, *Annual Scientific Meeting, Epilepsy Society of Australia (ESA)*, 2017. (Abstract submission: Poster Presentation).
- **D. Ahmedt-Aristizabal**, C. Fookes, K. Nguyen, S. Dionisio, Analysing body movement in seizures and dissociative attacks: A computer vision perspective, *Annual Scientific Meeting, Epilepsy Society of Australia (ESA)*, 2018. (Abstract submission: Poster Presentation).

1.6 Publications emerging from collaborative research

Scientific Reports [Q1(2018), IF: 4.122, H: 149, H5: 151].

- S. Sonkusare, **D. Ahmedt-Aristizabal**, M. Aburn, V. Nguyen, T. Pang, S. Frydman, S. Denman, C. Fookes, M. Breakspear, C. Guo, Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking, *Scientific Reports*, 9 (2019) 4729.

EMBC conference [H5: 31].

- M. Martinez, **D. Ahmedt-Aristizabal**, T. Väth, C. Fookes, A. Benz, R. Stiefelhagen, A Vision-based System for Breathing Condition Identification: A Deep Learning Perspective, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2019. (Oral Presentation).

Chapter 2

Literature review

2.1 Overview

The key contribution of this thesis is the development of automated frameworks that better capture and model semiology to support the analysis of epilepsy. Hence, this chapter reviews the literature on automatic detection and quantification of clinical manifestation from epileptic seizures and outlines its application in classifying epilepsy types. Major challenges and their impacts on the performance of epilepsy evaluation are discussed in the literature review.

This chapter first introduces basic knowledge of epilepsy and motor manifestations of specific epilepsy types considered in this research. Then, the chapter provides the analysis of traditional machine learning and computer vision approaches in the epilepsy context with their strengths and weaknesses. This section is divided according to the clinical sign that has been evaluated. Later, this chapter proceeds to introduce deep learning and its related techniques including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Considering the limited application of modern deep learning techniques in the epilepsy context, this chapter extends the review to applications of them in a broader context of human motion analysis and healthcare.

Machine learning is a field studying data analysis, which finds hidden data insights without being explicitly programmed with the ability to manipulate multi-modal and potentially contradictory data. Deep learning is a subset of the machine learning family that simulates operations of a human brain through a hierarchical multiple-layer signal representation coupled with advanced training algorithms. The study of this chapter seeks to reveal the lack of automated methodologies that assess some clinical manifestations during epileptic seizures. The application-specific literature will be introduced in later relevant chapters.

This chapter is supported by the following published manuscript:

- **D. Ahmedt-Aristizabal**, C. Fookes, S. Dionisio, K. Nguyen, J.P.S. Cunha, S. Sridharan, Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey, *Epilepsia*, 58 (2017) 1817-1831.

2.2 Contextualization of epilepsy and semiology

Epilepsy is a brain disorder characterised by recurrent and unpredictable interruptions of normal brain function, called epileptic seizures [Berg et al., 2010, Fisher et al., 2005]. People with epilepsy can face social stigma and exclusion [Shih et al., 2018]. Epilepsy is linked with an increased risk of death, up to 2–3 times the general population, which may be related to prolonged seizures or accidents by seizures in dangerous circumstances leading to drowning, burns or head injury [Australia, 2016]. Each year, Americans spend more than US$15.5 billion caring for and treating epilepsy. Direct care costs per patient can range from US$10,192 to US$47,862 annually [Holland, 2018]. Epilepsy-specific costs per year can cost upwards of US$20,000 [Holland, 2018]. Therefore, the aim of epilepsy treatment is to suppress seizures, to avoid social consequences of epilepsy and secondary handicap, to reduce mortality and morbidity, and to decrease the economic burden to the health system.

During seizures, patients with epilepsy may exhibit stereotypical behaviour. The analysis of such signs is termed *semiology*. The signs of epileptic seizures could include motor and non-motor manifestations [Noachtar and Peters, 2009]. In this research, we are interested in evaluating and developing quantitative methods that characterise motor manifestations such as facial modifications (*e.g.*, blinking, chewing, smacking), limb automatisms, ictal head turning and hand movements (*e.g.*, hand dystonia, tapping, grabbing) [Chauvel and McGonigal, 2014]. These symptoms depend on the part of the brain involved in the epileptic neuronal discharge, and the intensity of the discharge.

Over recent decades, the value of seizure semiology has been rather dichotomous, as evidenced by often contradictory contributions in the literature. On one side, seizure semiology is considered as the key to the localisation of dysfunction within the nervous system, in the same way as for clinical signs in other neurological diagnoses [Adams et al., 1997]. It is widely accepted that the analysis of clinical signs is essential for the diagnosis of epilepsy and that this process offers clues to underlying anatomical localisation of a possible epileptogenic network and pathophysiology, *i.e.*, the cerebral networks underpinning the epileptic episode [Lüders et al., 1998]. However, in contrast, some signs and patterns are considered of uncertain localising value or even potentially mislocalising [So, 2006]. The notion of misleading localising information comes particularly from observations in which similar clinical expressions have been found to occur in seizures arising from different brain localisations [Ryvlin et al., 2006]. This lack of certainty produces particular challenges in the clinical context and rely on the observational skills and experience of the expert.

Despite limitations as discussed above, detailed observations of multiple epileptic seizures show that similarities in semiology could involve neuronal activities within the same specific brain networks and could be sufficiently reliable to categorise patients with a specific type of epilepsy [Bonini et al., 2014]. Therefore, semiology is a vital resource in the pre-surgical evaluation in addition to neurophysiological and imaging data [Chauvel and McGonigal, 2014, Tufenkjian and Lüders, 2012].

One-third of patients with epilepsy will not respond to standard anti-epileptic drugs (AEDs) [Wiebe et al., 2001], where resective surgery can be envisaged if the epileptic seizure is

proved to come from a relatively localised area in the brain compared with others options such as laser ablation therapies, repetitive neurostimulation, deep brain stimulation, and vagus nerve stimulation [Wiebe et al., 2001]. Once the cortical region giving rise to seizure onset and the sites of early seizure propagation are identified, the patient has the recourse of surgical resection of the culprit brain region, which has been shown that the disconnection of the epileptogenic network is necessary and sufficient for seizure freedom [Rosenow and Lüders, 2001]. The goal of the pre-surgical evaluation is to delineate this network accurately. However, this network could be composed of other relevant networks, which are involved in originating interictal epileptiform discharges and producing the first clinical manifestation of a seizure. These networks could be localised on different brain areas, where identical symptoms may arise from various cortical regions. The misjudgement of the location of this brain network causes ineffective pre-surgical decisions [Thornton et al., 2010], so that follow-up of patients who have undergone surgery shows that only 66-70% and 41-79% of patients are rendered seizure-free after short-term and long-term evaluation, respectively [Spencer and Huh, 2008, De Tisi et al., 2011]. Some clinical signs or seizure semiology are particularly difficult to analyse because of their complexity [Williamson et al., 1985], which may result in mislocalisation of the seizure onset. Despite the establishment of several automatic methods that help to analyse brain electrical activity and neuroimaging in epilepsy, seizure semiology is still widely interpreted by visual inspection. This time-consuming process is based on neurologists' subjective interpretation and is prone to considerable inter-observer variability [Tufenkjian and Lüders, 2012].

There are different diagnostic tools to identify the epileptogenic network, including non-invasive and surgical methods. Non-invasive presurgical tools include scalp electroencephalography (EEG), magnetic resonance imaging (MRI), functional MRI, magnetoencephalography (MEG), single photon emission computed tomography (SPECT), and positron emission tomography (PET) [Gelziniene et al., 2008]. Certain clinical cases call for surgical monitoring, defined as recording methods employing electrodes placed surgically. These methods can be performed either using a subdural grid or electrocorticography (ECoG), or by depth electrodes or stereo-EEG (SEEG). This research focuses on SEEG recordings, because our work is solely in SEEG, with infrequent ECoG utilisation. SEEG as a methodology relies more heavily on semiology, as one of the founding principles is clinical–electrical–anatomical correlation. SEEG methodology is challenging to interpret, and training in anatomo-electro-clinical correlation is required [Serletis et al., 2014]. This correlation indicates the comprehension of the brain electrical activity associated with the semiology. In video-EEG or video-SEEG, patients are videotaped at the same time as their electrophysiology is recorded. The recording is carried out for a long period, often several days. In this way, epileptologists can analyse precisely how the behaviour of the patients during seizures is related to the electrical activity in the brain.

Temporal Lobe Epilepsy (TLE) is considered the most frequent cause of partial epilepsy in adults and many patients with MTLE have seizures resistant to antiepileptic drugs. This research chose this epilepsy type for analysis because it is one of the most common types of chronic medically

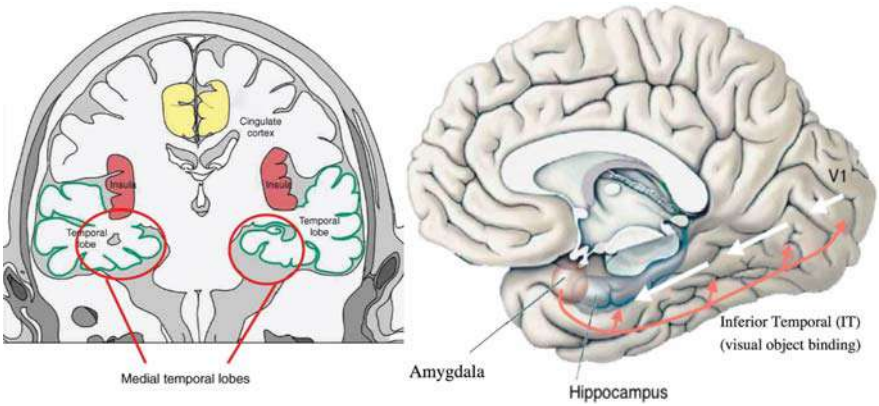


Figure 2.1 Schematic of predicted location of mesial temporal lobe epilepsy (MTLE). Image adapted from [Baars and Gage, 2012].

intractable epilepsy, which increases the probability to have sufficient information for experiments. Furthermore, characteristic clinical features of MTLE are well described and are reproducible in the majority of patients, which support the clinical hypothesis of the well-known proofs of characteristic motor symptoms that differ types of epilepsy quantitatively (see Section 3.2).

MTLE arise in the hippocampus, the parahippocampal gyrus and the amygdala located in the inner temporal lobe displayed in Figure 2.1. TLE is a group of disorders that predominately involves dysregulation of hippocampal function and hippocampal onset accounts for at least 80% of all temporal lobe seizures [Tatum IV, 2012]. The recommended treatment for patients with pharmacoresistant MTLE is surgery to remove the brain area(s) associated with the epileptogenic zone. The causes of TLE include mesial temporal sclerosis, traumatic brain injury, brain infections, such as encephalitis and meningitis, hypoxic brain injury, stroke, cerebral tumours, and genetic syndromes. The symptoms and the patient's behaviour depend on the precise location of its point of origin. Seizures generally tend to last only 1-2 minutes and there is some period of recovery in which neurological function is altered after each seizure. Focal seizures with impaired consciousness in MTLE are the primary seizure type. Reliable lateralising signs may include motionless staring, asymmetric tonic limb posturing, dystonic posturing, hand automatism, movements of the mouth such as chewing and fear [So, 2006]. Temporal lobe epilepsy is a common cause of EEGs containing inter-ictal epileptiform discharges (IEDs) and MTLE may also be associated with slow waves that have localising value. Interictal EEG findings in patients with MTLE typically include unilateral or bilaterally-independent mesial temporal spikes. Ictal EEG recordings usually reveal ictal onset consisting of rhythmic 5 to 7 Hz activity, but there are may be variations in this pattern [Cendes, 2005, Kibler and Durand, 2011].

2.3 Traditional automatic motion analysis in epilepsy

There have been numerous studies aiming to find a useful seizure detection device based on patient characteristics and seizure type [Ulate-Campos et al., 2016]. These detection devises are

assessed in terms of the effectiveness to capture diverse semiologies such as automatism, language, and motor abnormalities [Ulate-Campos et al., 2016, Ramgopal et al., 2014, Van de Vel et al., 2016]. To evaluate motor manifestations, movements can be analysed using accelerometers, surface electromyography, video detection systems, mattress sensor, and seizure-alert dogs. A disadvantage of each of these methods, however, is that they are limited to a subset of epileptic symptoms such as speech disturbances or motor signs [Ulate-Campos et al., 2016]. Nevertheless, video monitoring has been considered the gold standard to diagnose motor manifestations [Noachtar and Peters, 2009]. Movement-recording systems are needed to help neurologists to distinguish between partial and generalised seizures when selecting an initial antiepileptic drug, *i.e.*, they are vital to defining the accurate monitoring and treatment procedure [Blume et al., 2001].

An extensive literature analysis shows that researchers have devised diverse automatic applications in epilepsy for the tasks of seizure detection and classification, seizure prediction, continuous monitoring and pre-surgery assessment [Supratak et al., 2016, Orosco et al., 2013]. Researchers have implemented machine learning algorithms and computer vision techniques to increase the performance of epilepsy evaluation [Ulate-Campos et al., 2016, van Andel et al., 2016]. To assess seizures with motor phenomena, motion trajectories of markers and computer vision techniques have been used as feasible methods in recognising kinematic patterns of seizure. Computer vision and machine learning are concerned with the automatic extraction, analysis, and understanding of useful information from a single image or a sequence of images.

Clinical features that refer to specific cyclic motions of the body and its constituent parts such as the head, limbs, trunk, facial gestures, and posturing, can relate and characterise the epileptogenic network and involved anatomical areas [Chauvel and McGonigal, 2014]. Video monitoring is still the ideal detection sensor to assess seizures characterised by motor phenomena as is shown in Figure 2.2. Different types of cameras have been used in patients monitoring, including single, stereo, and depth cameras (*e.g.*, Microsoft Kinect camera). However, the most common technology used in Epilepsy Monitoring Units (EMUs) around the globe are single cameras. Noncamera approaches for motion quantification and classification among different epilepsy types have been investigated using devices such as accelerometers and gyroscopes; however, any sudden movement can be registered as a seizure event and they are not suitable for all kinds of semiology such as facial expressions [Ramgopal et al., 2014, Cunha et al., 2016a].

Pediaditis et al. [Pediaditis et al., 2012b] in their thorough review of vision-based motion analysis of epileptic seizures, evaluated different methodologies using marker-based and marker-free systems. Marker-based techniques must employ reference markers such as reflective materials attached to the key body parts to track the trajectory of the movements. Alternatively, marker-free systems eliminate the markers by exploiting computer vision techniques and new sensors developments for this task by calculating the displacement of flow vectors or relevant key points. Pediaditis et al. [Pediaditis et al., 2012b] revealed that the quantification of motion patterns in epilepsy is still under development and requires significant ongoing research and developments to address multisensorial approaches applied in the decision support system for epilepsy. Similarly,

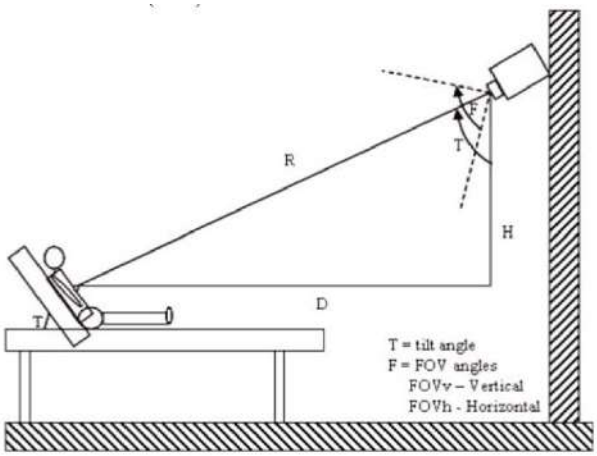


Figure 2.2 Schematic of traditional video set-up in epilepsy monitoring room. Image adapted from [Li et al., 2002].

a meaningful review of movement quantification in neurological diseases developed by do Carmo Vilas-Boas and Cunha [do Carmo Vilas-Boas and Cunha, 2016] summarised the literature of early automatic applications in extracting seizure motion patterns and recent computer vision approaches. The authors identified that clinical qualitative scales are considered reliable for medical practices but are still incomplete without effective quantitative motion capture solutions. The analysis of seizure semiology is still in the process of being clinically established in the evaluation of patients considered for epilepsy surgery, which provides lateralising information [Chauvel and McGonigal, 2014, do Carmo Vilas-Boas and Cunha, 2016]. Visual and painful auras (a perceptual disturbance experienced before the seizure begins and reflect the initial seizure discharge), as well as ictal motor manifestations such as dystonic posturing and emotional facial asymmetry, have shown lateralising value [Loddenkemper and Kotagal, 2005].

2.3.1 Quantification of semiology from head and upper limbs movements

The problem of motion based on a monitoring camera has been an active field of research in epilepsy for the past 15 years. Movement of the joints provides the key for the estimation of motion and recognition of the complete figure. Motion analysis and pose estimation is the process of estimating the configuration of the underlying kinematic or skeletal articulation structure of the patient, using motion capture and detection by joint skeleton tracking. The model usually is represented by a vector of joint angles [Pons-Moll and Rosenhahn, 2011]. Measurements are typically abstracted in salient points of interest or key points that are represented by feature vectors and tracked across consecutive image frames [Shi and Tomasi, 1994]. To conduct an accurate analysis of movement, it is essential to have a good follow-up of the human body movement. Body tracking is based on previously set body points defined, where motion information, such as position and velocity incorporated with intensity values, is employed to establish matching between consecutive frames. Once the movement detection

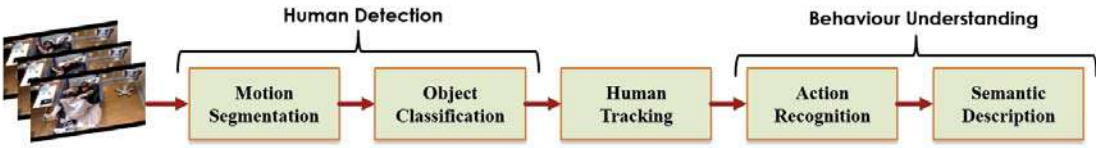


Figure 2.3 Traditional human motion analysis scheme. Image based on [Aggarwal and Cai, 1999].

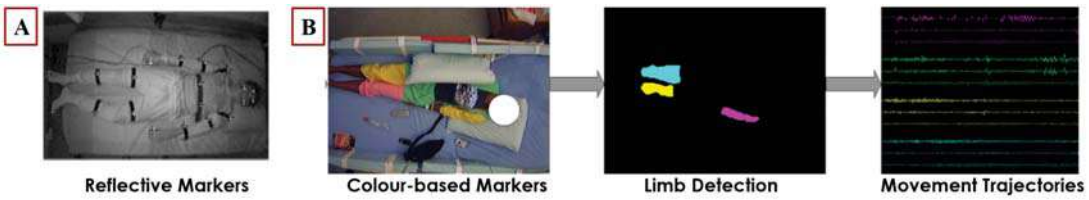


Figure 2.4 Selected samples of marker-based quantification. **A.** Reflective markers. Image adapted from [Li et al., 2002]. **B.** Colour-based markers. Image adapted from [Lu et al., 2013].

and tracking are solved, the next step is the understanding of the behaviour of these features throughout the image sequence. The sequence of human motion recognition is illustrated in Figure 2.3.

Countless automatic movement quantification approaches calculated features by analysing velocity, amplitude, duration, rotation, the direction of motion, and angular speed of the body using marker-based or marker-free systems. Techniques that use reference markers such as reflective material are placed on the head, trunk or extremities. Li et al. [Li et al., 2002], conducted a pioneering quantitative analysis of motion trajectories in human body parts in two-dimensional (2D) video recordings, using markers at landmark points of a patient's body covered with infrared reflective material. Following this pioneering work, different studies have shown suitability in interpreting semiology signs from limbs motion [Cunha et al., 2003, Lu et al., 2013] and head movements [O'Dwyer et al., 2004, Wagner et al., 2004]. Figure 2.4 displays selected samples of attached markers in the body. Although positive results have been achieved in motion quantification using marker-based methods, they lack precision, rely on manual marking and the attached markers can be uncomfortable or dislocated over time by violent movements during seizures.

Based on the evolution in the field of computer vision, maker-free systems have started to be used. Different techniques to detect and to track motion including optical flow and shift clustering analysis [Karayiannis and Tao, 2003, Karayiannis et al., 2006, Cuppens et al., 2010, Kalitzin et al., 2012, Cuppens et al., 2012a], spatio-temporal interest point detectors (SITPs) and histograms-of-flow features [Cuppens et al., 2012b, Mandal et al., 2012] have been used with marker-free systems. Selected samples of seizure quantification using these techniques are shown in Figure 2.5.

The optical flow method calculates the displacement flow vectors from a video sequence and estimates a motion vector in each pixel of the video according to the estimated movement, based on the change in the pixel intensities [Barron et al., 1994, Hu et al., 2004]. This method provides a two-dimensional representation of three-dimensional motion. The most common implementation of the optical flow method is the traditional Horn–Schunk algorithm [Horn and Schunck, 1981]. This

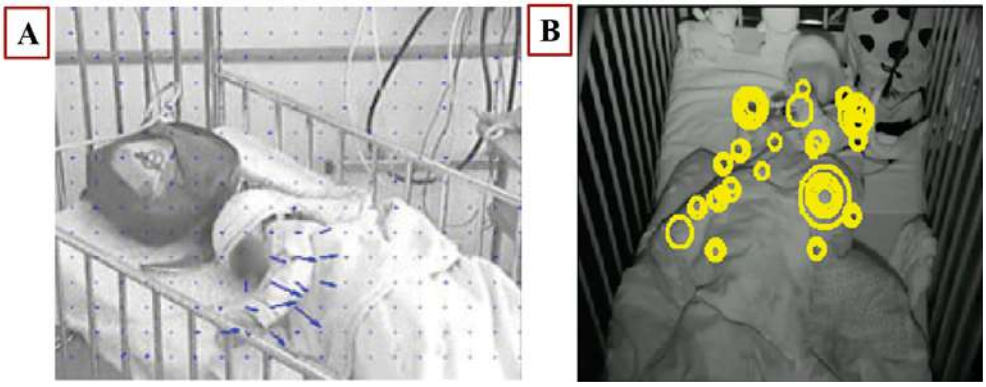


Figure 2.5 Selected samples of marker-free quantification. **A.** Optical flow image. Image adapted from [Cuppens et al., 2010]. STIPs detected within one frame. Image adapted from [Cuppens et al., 2012b].

implementation uses a smoothness constraint to solve the aperture problem and the Lucas–Kanade method [Lucas and Kanade, 1981] to assumes that the flow in small sections of pixels is the same. Spatio-temporal interest point detector measures relevant key points inside a spatio-temporal window [Laptev, 2005]. This technique searches for corners in 3D video space (two spatial and one temporal dimension) based on the Harris corner detection.

Optical flow, spatio-temporal interest point and clustering analysis have shown limited success because they are based on handcrafted features. These approaches often require the presence of specific motion frequencies during seizures and they are dependent on detecting a sufficiently large amount of key points [Pediaditis et al., 2012b, Achilles et al., 2016c]. Additionally, these techniques implemented in 2D videos have limitations regarding the area of the patient's body that can be observed, where the patient must be visible without occlusion by other objects, bed linens, or humans such as clinical staff. Current techniques are also heavily constrained, as they can only recognise seizures with significant movements, and they fail to capture subtle and fine-grained changes and the quantification of movements that are not in view of the camera.

In order to solve these challenges of motion quantification with single cameras, an approach was proposed based on multi-cameras. Cunha et al. [Cunha et al., 2012], developed the first 3D application in epilepsy using four high-speed infrared motion-tracking cameras and spherical infrared reflective markers attached to anatomical points (head, trunk, arms, and legs) as is illustrated in Figure 2.6. However, this proposal is an expensive multi-camera optical system, and still requires manual marking of the human body and references to objects located in the patient room.

2.3.2 Quantification of semiology from facial expressions

Facial expressions are a significant source of information while monitoring the wellness of a patient, including pain, depression, drowsiness and stress monitoring [Sathyanarayana et al., 2015, Miyazaki et al., 2000]. Clinical studies have demonstrated the usefulness of facial expression

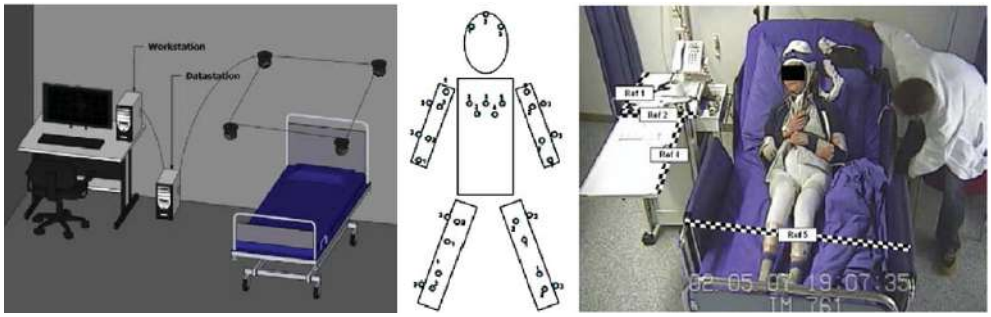


Figure 2.6 Selected sample of marker-based quantification with multi-cameras. System setup with 4 high-speed SVCams for the 3D motion tracking, the setup of the reflective markers and the room position references to quantify motions. Image adapted from [Cunha et al., 2012].

analysis as a lateralizing information in epilepsy. These expressions include unilateral blinking and eyeball deviation [Benbadis et al., 1996, Henkel et al., 2000], chewing movements [Gil-Nagel and Risinger, 1997, Kotagal et al., 1995, Aupy et al., 2018], post ictal nose wiping [Ataoğlu et al., 2015], motion in the mouth area in the form of lip licking, and turned-down mouth [Souirti et al., 2014] as illustrated in Figure 2.7. Szaflarski et al. [Szaflarski et al., 2018] demonstrated the presence of differences in facial emotions between motor functional neurological disorders and epileptic seizures, *i.e.*, patients with similar clinical semiologies but with different etiologies. Baysal-Kirac et al. [Baysal-Kirac et al., 2015] indicated that eye and head movement directions distinguish ictal ipsi and contralateral turns. Although clinical studies have used analytical methods to identify differences between seizures, the movement detection and quantification were assessed manually with symptoms scales and chart review.

Recent contributions in the engineering field have been mainly focused on the extraction of features from facial motions. Maurel et al. [Maurel et al., 2008], one of the earliest approaches in the studies from computer vision area, performed a facial expression analysis in epileptic patients during seizure events by fitting a 3D active appearance model illustrated in Figure 2.8. This architecture uses a modified version of the Candide 3D face model [Ahlberg, 2001], which contains 113 vertices and 184 triangles. The facial expression was represented as the variation of the animation parameters between a neutral image and the expressive one. Although it is a novel face representation, this model is simplistic, unable to adequately represent the facial expressions and does not handle significant changes in head orientations. Pediaditis et al. [Pediaditis et al., 2011], pioneered a method for facial motion analysis in epilepsy using an averaging background method and dense optical flow for the automatic feature extraction in video sequences. The features were used to discriminate between facial motion classes, including eye and mouth. Later, the authors improved the previous methodology using a region-based technique based on the Viola-Jones algorithm for face and eyes detection to study facial expressions from patients with absence seizures [Pediaditis et al., 2012a]. Sathyanarayana et al. [Sathyanarayana et al., 2015] proposed a template-based method to detect eye movements in the occurrence of absence seizure, where the technique is based on the weighted accumulation of

(a) Patients with ictal pouting ("*chapeau de gendarme*"). Image adapted from [Souirti et al., 2014].

(b) Patients with eye deviation during head turning. Image adapted from [Baysal-Kirac et al., 2015, Zhang et al., 2017c].

Figure 2.7 Selected samples of facial semiology.

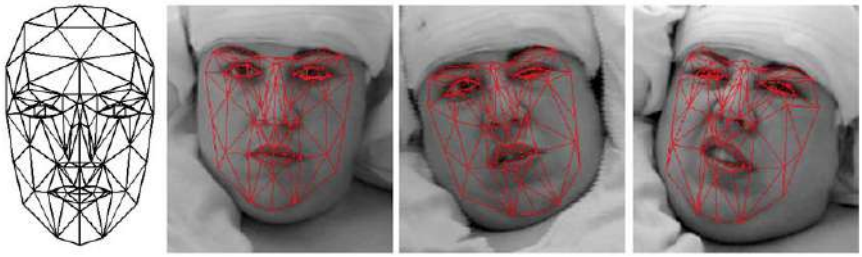


Figure 2.8 Selected sample of quantification of facial expressions in epilepsy. Candide technique for 3D model fitting. Image adapted from [Maurel et al., 2008].

intensities and gradient magnitudes in the eye region. Overall, automated quantification of semiology in facial expressions has advanced insufficiently and occlusion and head pose are still a major challenge.

2.3.3 Automatic classification of epilepsy types based on semiology

Patients with refractory epilepsy show lateralising phenomena in head movement, eye deviations, and dystonic hand posturing [Noachtar and Peters, 2009, McGonigal and Chauvel, 2004], as illustrated in Figure 2.9. Diverse studies have compared semiologic features through quantitative movement analysis that could allow differentiation between epilepsy types. For instance, temporal lobe (TLE) and frontal lobe (FLE) epilepsies are types of epilepsy that are localisation-related and are the most common epilepsy syndromes in patients considered for epilepsy surgery [Noachtar et al., 2003].

Figure 2.9 Selected sample of semiology from upper limbs movements. Asymmetric tonic posturing. Image adapted from [McGonigal and Chauvel, 2004].

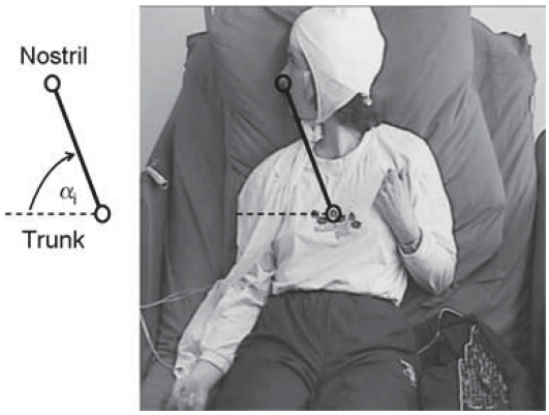


Figure 2.10 Selected sample of the quantification of the head movement tracked at the left nostril in relation to a reference point on the patient's shirt. Image adapted from [O'Dwyer et al., 2007].

Preliminary studies on wrists and trunk movements demonstrated the advantages of the quantitative analysis to classify hypermotor and automotor seizures [Ulowetz et al., 2005, Meier et al., 2004]. O'Dwyer et al. [O'Dwyer et al., 2007] validated the high lateralising value of ictal lateral head movements in TLE and analysed the relationship between the nose and a defined point on the thorax as shown in Figure 2.10. They focused on seizures in which bilateral head movements occurred. The motion tracking computed the duration and the angular speed at which the head movements occurred, and the direction of the ictal head movements was defined by the ictal EEG analysis performed by the experts manually. Rémi et al. [Rémi et al., 2011b] evaluated significant differences between patients with FLE and TLE through quantitative analysis of head turning, where these changes likely represent differences in the spread of epileptic activity. Head movements were quantified using infrared markers placed on the head and the trunk of the patients. The time of onset, duration, and angular speed of the head movements were calculated. However, the ictal activity was provided by the experts and the method could only process seizures for which the camera position was perpendicular to the patient's coronal plane.

(a) Hypermotor seizure. (b) Automotor seizure.

Figure 2.11 Trajectory detection of upper limbs using an infrared reflective marker strategy. Image adapted from [Cunha et al., 2009].

Understanding the significance of analysing upper limbs movements, and using infrared reflective markers attached to the patient's body, researchers have classified types of epilepsy [Mirzadjanova et al., 2010]. Cunha et al. [Cunha et al., 2009] classified automotor and hypermotor seizure using quantitative metrics such as the average and maximum speed of the wrist movements from reflective markers as displayed in Figure 2.11. Chen et al. [Chen et al., 2009] discriminated four types of supplementary motor area seizures: hyperkinetic, tonic posturing, fencing posture, and tonic head turning. Rémi et al. [Rémi et al., 2011a] differentiated hyperkinetic and automotor seizures by noninvasive video-EEG monitoring. The trunk and wrist movement extent and speed were analysed. Furthermore, Cunha et al. [Cunha et al., 2013] developed an automatic analysis that differentiates FLE from TLE through ictal upper limb automatisms. To measure the seizure duration, the authors used clinical onset, as it preceded the EEG seizure onset. In general, these studies have the limitation of the camera position, and a further correlation with the electrical activity for epileptogenic network localization was not explored.

In order to improve the limitations of two-dimensional marker-based and marker-free approaches, the easy access to depth cameras is unlocking new computer vision applications. Cunha et al. [Cunha et al., 2016a] implemented a low-cost 3D video-EEG system to perform a kinematic analysis and classify TLE and extra-temporal lobe epilepsies (ETLE). The hardware was based on a red-green-blue–depth (RGB-D) sensor (Microsoft Kinect camera) connected to a custom acquisition software tool called Kinect-Tracker. Figure 2.12 shows the kinematic analysis of the patient. Seizure groups were discriminated by different movements of interest, 19 arising from TLE and 23 from EXTE, marked by clinical experts. The tracking process was computed using the Horn-Schunk optical flow method over RGB images and depth maps provided by the Kinect, which made it possible to identify the relevant body parts during the seizure. Different metrics such as velocity, acceleration, movement displacement, and covered distance were calculated to perform the 3D quantification. Although the proposal is a novel approach, the 3D joint tracking carried out is not accurate enough in different scenarios due to the limitations of depth cameras, including structured light, time of

Figure 2.12 3D tracking of the body joints movements. These methodologies are still undergoing improvement; for example, this image illustrates that the fitting of the skeleton model has not been completely accurate because one of the two arms is detected entirely incorrectly. Image adapted from [Cunha et al., 2016a].

flight, depth discontinuities, brightness, multiple reflections, and viewpoint [Lun and Zhao, 2015]. No further automation method using electro-clinical correlation in the presurgical evaluation was covered in this research. This contribution also supported the assessment of neurological diseases characterised by abnormal and/or involuntary movements such as Parkinson Disease [Cunha et al., 2016b] using the 3D position of several body joints. Additionally, the system was used to provide an on the fly quantitative report of the gait analysis [Rodrigues et al., 2018]. Pereira et al. [Pereira et al., 2018], expanded the single-bed system documented in [Cunha et al., 2016a], to a multi-bed that analyse three patients simultaneously. Although the detection and tracking strategy used an improved detector (Microsoft Kinect v2), the methodology implemented still has the limitations mentioned of depth sensors.

The literature review has revealed that despite the potential of motion analysis, the use of automated techniques for assessing epilepsy is still relatively rare in clinical practice. Facial expression quantification and its implementation to classify epilepsy types have not been reported. Table 2.1 summarises engineering contributions to classify epilepsy types. Until now, the research community has not sufficiently addressed the development of automated multi-modal methodologies that can jointly learn across different clinical manifestations to assess epilepsy. Although traditional machine learning and computer vision techniques have been used successfully in the analysis of epilepsy, it is still very challenging to reach human-level accuracy in evaluating specific motor phenomena. The main problem is the selection and extraction of the appropriate features to capture motion. Table 2.2 summarises the most significant limitations of traditional techniques, which can be addressed in many ways by implementing deep learning methodologies in epilepsy as discussed in the following section.

Table 2.1 Summary of contributions from engineering in the classification of epilepsy types.

Author	Epilepsy Type	Method	Dimension	Motion
[O'Dwyer et al., 2007]	TLE (Ipsilateral, Contralateral)	Marker-based	2D	Head
[Cunha et al., 2009]	Automotor, Hypermotor	Marker-based	2D	Upper limbs
[Chen et al., 2009]	Hyperkinetic, Tonic Posturing, Fencing posture, Tonic head turning	Marker-based	2D	Upper limbs
[Rémi et al., 2011a]	FLE, TLE	Marker-based	2D	Head
[Rémi et al., 2011a]	Hyperkinetic, Automotor	Marker-based	2D	Upper limbs, Trunk
[Cunha et al., 2013]	FLE, TLE	Marker-based	2D	Upper limbs
[Cunha et al., 2016a, Pereira et al., 2018]	FLE, ETLE	Marker-free	3D	Upper limbs

Table 2.2 Summary of disadvantages of traditional techniques to quantify epilepsy.

Methodology	Limitations
Marker-based systems	
Applications: 2D/3D; Key areas: head, upper limbs, trunk; Techniques: thresholding, ANN, probabilistic methods	Manual marking labelling. Markers can be uncomfortable or dislocated over time by violent movements during seizures. Unable to detect seizures at night monitoring. 3D approaches with multiple cameras are expensive optical systems. Camera position
Marker-free systems	
Applications: 2D/3D; Key areas: upper limbs, face;	*2D (single cameras)*: No automatic inference of specific body parts. Variation in the head pose. Changes in illumination. Vulnerability to image noise. Sensitivity to motion discontinuities.
Techniques: Optical flow & Lucas-Kanade tracking, Spatiotemporal interest points, Histograms-of-flow features, Cascade-based and deformable models	*3D (depth sensors)*: Pose estimation and tracking are still under development. General limitations of depth sensors such as structured light, time of flight, depth discontinuities, brightness, multiple reflections and viewpoint.
Traditional computer vision analysis	
Marker-based and marker-free systems. Classification/prediction: Support vector machines, neural networks	Hand-crafted feature extraction and relies on assumptions. Controlled test situations. Using 2D videos fails to capture movements that are not included in the field of view of the camera; only processes seizures when the camera position is perpendicular to the patient's coronal plane. Algorithms required the presence of specific motions frequencies during seizures. Heavily constrained as they can only recognise seizures with significant movements and fail to capture subtle and fine-grained changes during motion seizures. Dependency on a sufficiently large amount of detected key points. Occlusion by objects, bed linens, or humans (family members and clinical staff).

2.4 Modern deep learning for epilepsy

2.4.1 Deep learning

Deep learning is a family of machine learning that allows computational architectures that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction and using non-linear transformations [Bengio et al., 2013]. Deep learning methods simulate the operation of a human brain through hierarchical multiple-layers coupled with advanced training algorithms. The major difference between traditional machine learning and deep learning techniques is in feature engineering.

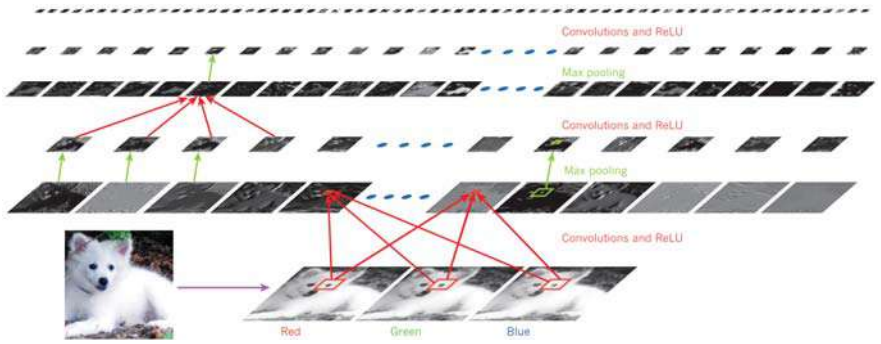


Figure 2.13 Representation of a convolutional neural network architecture. Each rectangular image is a feature map corresponding to the output for one of the learned features, detected at each of the image positions. Information flows bottom up, with lower-level features acting as oriented edge detectors, and a score is computed for each image class in output. Image adapted from [LeCun et al., 2015].

In deep learning algorithms, *the feature representation is automatically learned from the training data, not from human assumption like traditional hand-crafted approaches* [LeCun et al., 2015]. This could be extremely useful when there is a limited understanding of the relation between the input and the output, which might prevent engineering effective features [Arel et al., 2010]. When modelling data, most traditional kernel-based methods such as Support Vector Machines, Gaussian Processes and Bayesian models utilise numerous hyper-parameter that are tuned to specifically to the task at hand, hence expected the test set to be closer to the training set [Bengio et al., 2013]. There is a no clear definition on the number of layers that separates deep architectures from shallow architectures, yet it is widely believed that any neural network that has more than two hidden layers should be considered deep.

Deep learning techniques have become the mainstream in computer vision and signal processing in the past several years, outperforming traditional approaches in many tasks including object detection, segmentation, recognition, natural language processing, and sequence learning [LeCun et al., 2015, Schmidhuber, 2015]. These representation-learning methods have won many contests in pattern recognition and machine learning and proven themselves successful in surpassing human abilities in image recognition [Szegedy et al., 2015], time series analysis [Längkvist et al., 2014], neuroimaging [Plis et al., 2014], and biological applications [Sønderby et al., 2015, Angermueller et al., 2016].

Notably, there are several architectures of deep neural networks, including convolutional neural networks, and recurrent neural networks. The following subsections present an overview of these architectures that have been implemented for image and signal analysis.

Convolutional neural networks

Convolutional Neural Networks (CNNs) are a special branch of feedforward neural networks, which are specifically designed for analysing visual imagery. Their architecture and operations were inspired

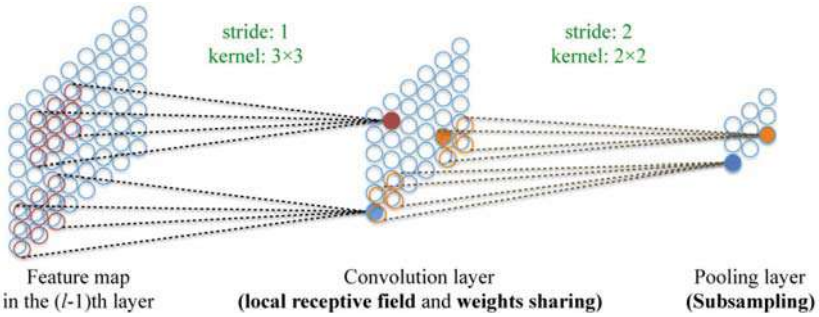


Figure 2.14 Schematic of three key mechanisms in convolutional neural networks: local receptive field, weights sharing, and subsampling. Image adapted from [Shen et al., 2017].

by biological processes in that the connectivity pattern between neurons resembles the organisation of the animal visual cortex. Compared with traditional feed-forward neural networks, CNNs exploit spatial locality by enforcing a local connectivity and parameter sharing [LeCun et al., 2015]. CNNs posses the most straightforward learning procedure which directly back-propagates the classification error. CNNs recognise small patterns at each layer and generalising them (detecting higher order, more complex patterns) in subsequent layers. They use convolution instead of a general matrix multiplication, which allows dealing with the input of variable size. A typical operator used together with convolution is pooling, which combines nearby values in input or feature space through a max, average or histogram operator. The purpose of pooling is to achieve invariance to small local distortions and reduce the dimensionality of the feature space [Längkvist et al., 2014, Cireşan et al., 2012]. Figure 2.13 illustrates an example of the architecture and detection of patterns on CNNs. Traditionally, CNNs have convolutional layers interspersed with pooling layers, followed by fully connected layers as in a standard multi-layer neural network. CNNs exploit three mechanisms of local receptive fields, weights sharing, and subsampling that help greatly reduce the degrees of freedom of a model as it is illustrated in Figure 2.14.

There has been a considerable advancement in the architecture design of CNNs to improve the analysis of images. The ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al., 2015] has been considered a representative competition to compare deep learning technologies to see which one is how much better than the other one. This is an annual contest that started in 2010. Since 2012, variants of CNNs have dominated the ILSVRC and have surpassed the level of human accuracy. Deep Learning models for image classification have achieved an exponential decline in error rate over in the last few years. Figure 2.15 illustrates this performance.

The most famous CNNs architectures are displayed in Figure 2.16. AlexNet [Krizhevsky et al., 2012] was the first famous CNN (Figure 2.16(b)) and the one that started it all. However, some researchers think that it was LeCun et. al [LeCun et al., 1998] with the architecture depicted in Figure 2.16(a), to recognise handwriting. AlexNet is one of the most influential architectures in the field which won the 2012 ILSVRC. The network was made up of 5 convolutional layers, max-pooling layers, dropout layers, and 3 fully connected layers. However,

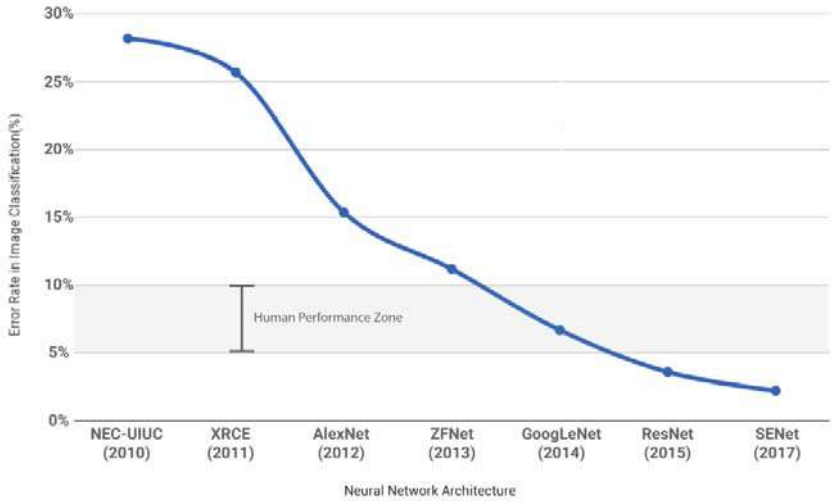
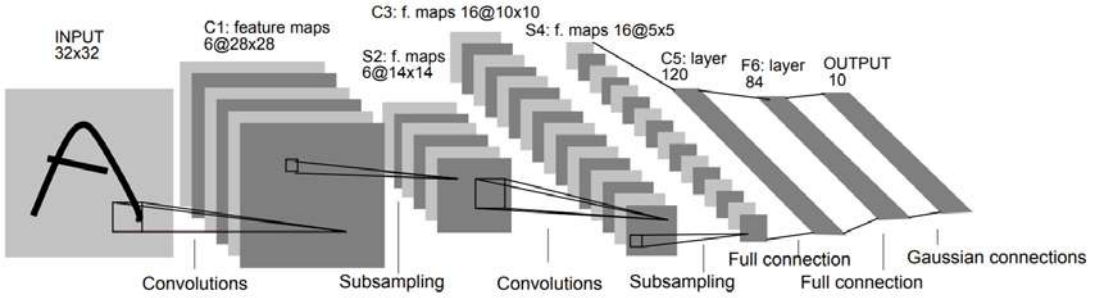
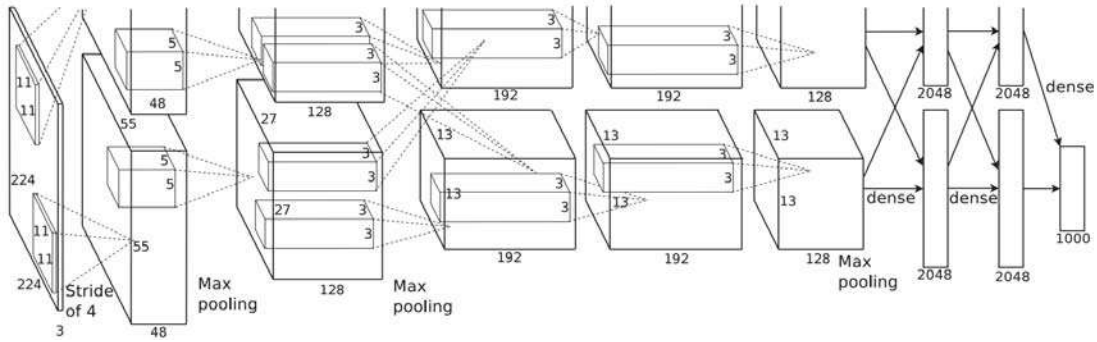


Figure 2.15 Performance of the recognition challenge (ILSVRC). The performance saw an exponential decline in top 5 error rate for neural network architecture for image classification over the past few years. Data source ImageNet [Russakovsky et al., 2015].

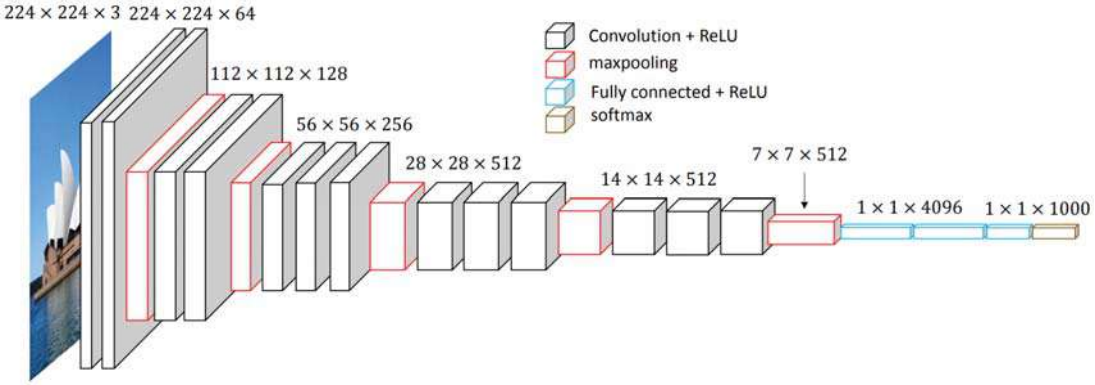
there was no standard about filter sizes to be used, how many convolutions before a max-pooling, etc. Without the insight understanding on inner mechanisms on CNNs, the development of better models is reduced to trial and error [Zeiler and Fergus, 2014]. Zeiler and Fergus [Zeiler and Fergus, 2014] provide details of a slightly modified AlexNet model and a fascinating way of visualising features maps using a new architecture named ZFNet. With the introduction of VGGNet (Figure 2.16(c)), Simonyan and Zisserman [Simonyan and Zisserman, 2014] suggested some standards such as the use of filters size of 3×3 (different from AlexNet 11×11 and ZFNet 7×7), max pooling should be placed after every 2 convolutions and the number of filters should be doubled after each max-pooling. The authors' reasoning is that the combination of two 3×3 convolution layers has an effective receptive field of 5×5. This, in turn, simulates a larger filter while keeping the benefits of smaller filter sizes. One of the benefits is a decrease in the number of parameters. VGGNet is one of the most influential proposals because it reinforced the notion that CNNs have to have deep networks of layers in order for the representation of visual data to work. The VGGNet with 16 weight layers produced the best results documented. GoogLeNet (Figure 2.16(d)), has an entirely different architecture than previous proposals. This network does not follow the idea of simplicity by proposing combinations of inception modules, each including some pooling, convolutions at different scales and concatenation operations. It also uses 1×1 feature convolutions that work like feature selectors. Overall, these crucial developments in convolutional neural networks have been the base of innumerable proposals to assess new problems in the field of computer vision.



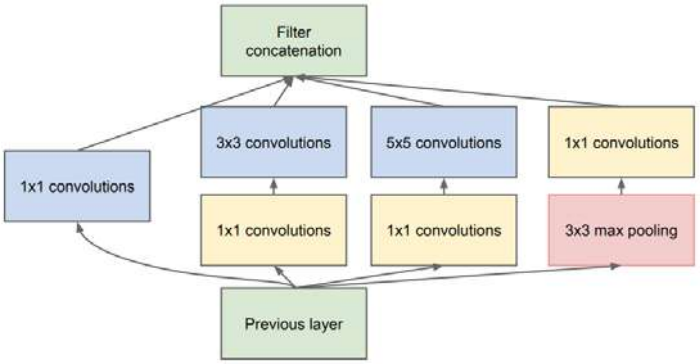
(a) LeNet-5. Image adapted from [LeCun et al., 1998].



(b) AlexNet. Image adapted from [Krizhevsky et al., 2012].



(c) VGGNet (VGG16). Image adapted from [Simonyan and Zisserman, 2014, Wang et al., 2017b].



(d) GoogLeNet. Image adapted from [Szegedy et al., 2015].

Figure 2.16 The most famous convolutional neural network architectures in the literature.

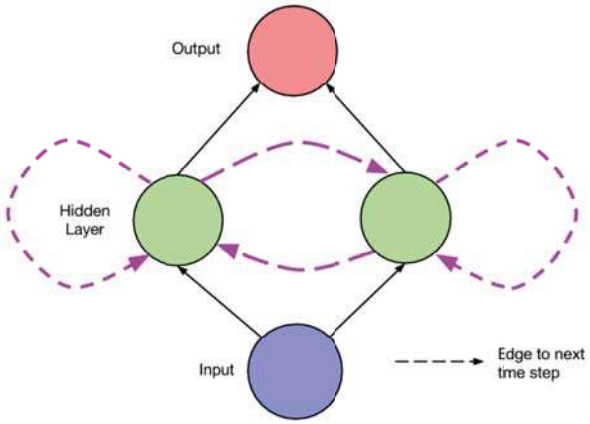
Recurrent neural networks

Recurrent Neural Networks (RNNs) introduce the notion of time into the deep model via including recurrent edges that span adjacent time steps [Lipton et al., 2015]. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Another way to think about RNNs is that they have a memory which captures information about what has been calculated so far [Lipton et al., 2015]. The edges that connect adjacent time steps are called recurrent edges. Figure 2.17(a) shows a basic RNN and Figure 2.17(b) depicts the usual way of interpreting an RNN, as a deep network with one layer per time step and shared weights across time steps. An RNN is obtained from the feedforward network by connecting the neurons' output to their inputs (Figure 2.17(c)). The short-term time-dependency is modelled by the hidden-to-hidden connections without using any time delay-taps. The RNN networks can be trained using a procedure known as backpropagation-through-time (BPTT) [Werbos, 1990].

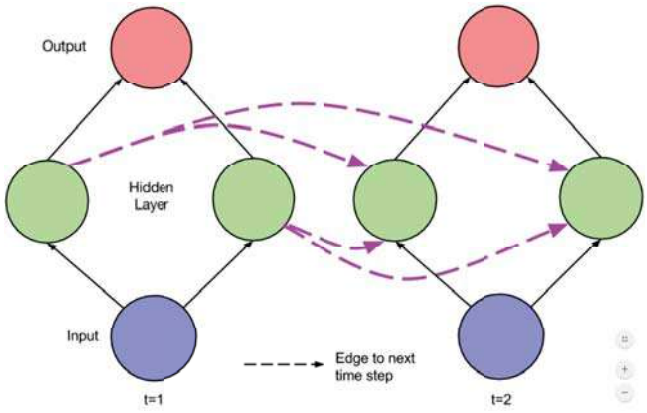
RNNs perform poorly when dealing with long sequences due to its frequently encountered drawback in gradient vanishing and exploding [Greff et al., 2017]. Long-Short-Term-Memory (LSTMs), which has evolved from the recurrent neural networks, are much better at capturing long-term dependencies. LSTMs seek to address this issue by using a gated mechanism. Three gates, *i.e.*, forget, input and output gates, are used to control the flow of information. Figure 2.18 illustrates a representation of the gates in an LSTM network. While the traditional RNN cell has a single "internal layer" acting on the current state h_{t-1} and input x_t, the LSTM cell has three. The amount of information that is let through each gate is controlled by a point-wise multiplication and sigmoid function [Greff et al., 2017].

2.4.2 Deep learning in human motion analysis

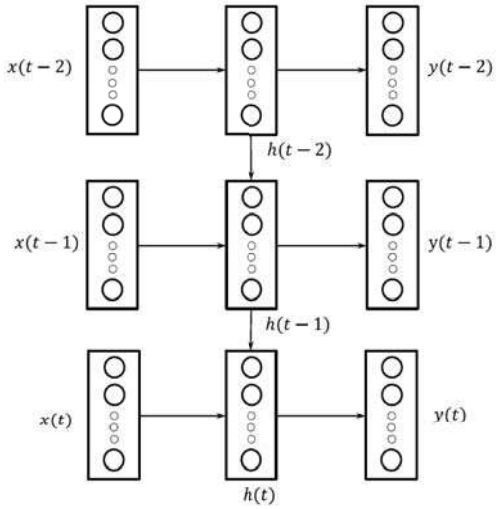
Quantifying a person's posture and limb articulation is a challenging process because of the uncoordinated movement in a person's body and the marked variation that exists among patients performing the same or similar actions. An accurate pose estimation system must be robust to heavy occlusion, severe deformation, and imaging condition changes. Deep learning techniques excel in modelling long-range dependencies between variables in structured prediction tasks, which is particularly suitable for articulated pose estimation. Since Toshev and Szegedy [Toshev and Szegedy, 2014] introduced the concept of DeepPose, CNNs have shown remarkably robust performance and high part localisation accuracy. They have outperformed classical approaches, and have widely replaced hand-crafted features from sequential prediction framework, pictorial structure, hierarchical, and graphical models. Proposed methods for articulated human pose estimation using convolutional architectures can be classified as detection-based or regression-based. Detection-based methods rely on CNN-based part detectors, which are then combined using a graphical model or refined using regression. Regression-based methods try to learn a mapping from an image and CNN features to part locations [Bulat and Tzimiropoulos, 2016a]. There are several examples of methods making successful predictions for pose estimation. Wei et al. [Wei et al., 2016]



(a) A simple recurrent network. Dashed edges connect a source node at each time t to a target node at each following time $t + 1$. Image adapted from [Lipton et al., 2015].



(b) The recurrent network unfolded across time steps. Image adapted from [Lipton et al., 2015].



(c) The input x is transformed to the output representation y via the hidden units h. The hidden units have connections from the input values of the current time frame and the hidden units from the previous time frame. Image adapted from [Längkvist et al., 2014].

Figure 2.17 Representation of recurrent neural networks.

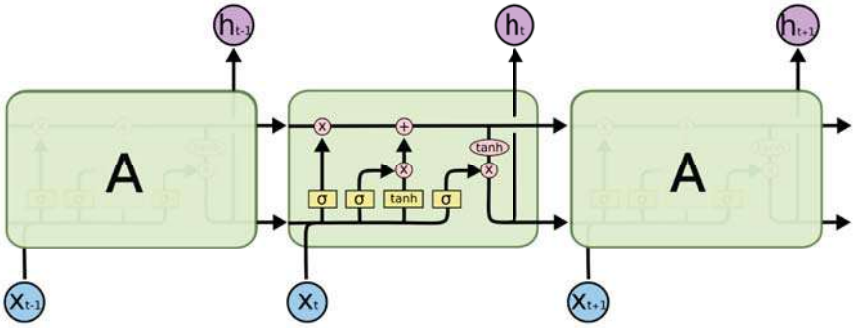


Figure 2.18 Representation of a Long-Short-Term-Memory (LSTM) cell structure. The repeating module in an LSTM contains four interacting layers.

developed a methodology with the ability to learn feature representations for both image and spatial contexts directly from the data. This approach generated heat-maps to simultaneously capture features at a variety of scales and detected body parts such as the head, neck, elbow, wrist, knee and ankle. Newell et al. [Newell et al., 2016] introduced a novel "stacked hourglass" network for predicting human pose with a more symmetric topology. Features were processed across all scales and consolidated to best capture the various spatial relationships associated with the body.

The wealth of information present in facial motions has stimulated researchers in applied deep learning architectures to create highly sophisticated models that effectively capture nonlinear mappings of intrinsic features of facial muscle motions [Unzueta et al., 2014]. Handcrafted features perform inaccurately for the facial expression recognition task under uncontrolled conditions due to a wide range of variations in pose, scale, illumination, and occlusion, and natural variations of individuals in facial shape, texture, and behaviour. Deep learning techniques have proved that they can deal with these challenges effectively for different tasks [Burkert et al., 2015, Liu et al., 2014b], including recognition of facial semantic features [Taigman et al., 2014, Liu et al., 2013], facial motions [Jaiswal and Valstar, 2016, Ghasemi et al., 2016]. Deep learning architectures have demonstrated to be the best methods to estimate reliable features to analyse facial expressions implementing landmark-based or region-based methodologies. Landmark-based approaches quantify facial movements using key facial points or landmarks [Zhang et al., 2016c]. Region-based methods process the raw sequence of images through an end-to-end deep learning model to learn the spatio-temporal features [Zhang et al., 2017a].

The aforementioned proposals that detect and quantify motions from limbs and face have confirmed that the most important advantage of deep learning is that the learned dynamic appearance or discriminative features could be generalised and applied to different datasets and applications.

2.4.3 Deep learning in healthcare

Deep learning methods have recently had a high success in medical imaging analysis, including the fields of genomic sequence, ophthalmology, pathology, cancer detection, radiology, gastrointestinal

diseases detection, tumour detection and prediction and personalised medicine [Razzak et al., 2018, Lee et al., 2017, Shen et al., 2017, Litjens et al., 2017, Esteva et al., 2017]. For example, it was reported that the automated staging and detection of Alzheimer's disease, breast and lung cancer has shown optimistic diagnostic performances. The notion of applying deep learning to medical imaging data is a fascinating and growing research area; however, scarcity of training data, privacy and legal issues, and dedicated medical experts are still the most significant barriers that slow down its progress.

Deep learning has revolutionised computer vision through end-to-end learning, and applying it to time-series data is also gaining increasing attention [Längkvist et al., 2014]. For instance, an overview of the capabilities of deep neural architectures for classifying brain signals is given in [Craik et al., 2019, Bozhkov, 2016]. These methodologies address the challenges of traditional machine learning techniques based on temporal and frequency domains, wavelet transforms, or energy analysis and are robust to analyse high-dimensional data with a poor signal-to-noise ratio and considerable variability between individual subjects and recording sessions [Bashivan et al., 2015].

Recent machine learning advancements in sensor-based mobility, computer vision and deep learning are supporting human action recognition [Wang et al., 2018b] and the motion assessment in patients with neurological diseases such as Parkinson's disease [Li et al., 2017, Li et al., 2018a, Camps et al., 2018, Kuhner et al., 2017] and Alzheimer's disease [Wang et al., 2018c]. Despite the clear benefits of deep learning to analyse motions from patients, it is still common to find recent publications that still use traditional computer-based techniques to assess movements that only exploits the benefits of sensors developments [Orlandi et al., 2018, Naghavi and Wade, 2018, Yamamoto et al., 2018]. This opens an opportunity to make significant progress in the healthcare field that suggests the use of deep features.

2.4.4 Deep learning in semiology analysis

Deep learning has excelled in many biological and medical applications but has advanced insufficiently in epilepsy analysis, specifically, in the assessment of semiology. Deep learning architectures within cognitive neuroscience and for processing electrophysiological recordings from epileptic patients have progressed considerably [Schirrmeister et al., 2017, Thodoroff et al., 2016, Antoniades et al., 2016]; however, the analysis of clinical manifestations have been very limited so far. There are only a few proposals in the context of head and limbs motions and there is no evidence of applications to assess facial expressions and hand motions.

Achilles et al. [Achilles et al., 2016c] developed a novel seizure detection method using video frames from a combined depth and infrared sensor and CNNs to extract features, as it is shown in Figure 2.19. The frames were used to train the network in a supervised fashion and detect seizure-related static and slow patient motions in various types of epileptic seizures. However, the image sequences used for training were determined by medical professionals, which could affect the generalisation of the trained system by the dependency on the experts' evaluation and the accuracy in differentiating between the electrical onset and the spreading of activity using scalp EEG recordings.

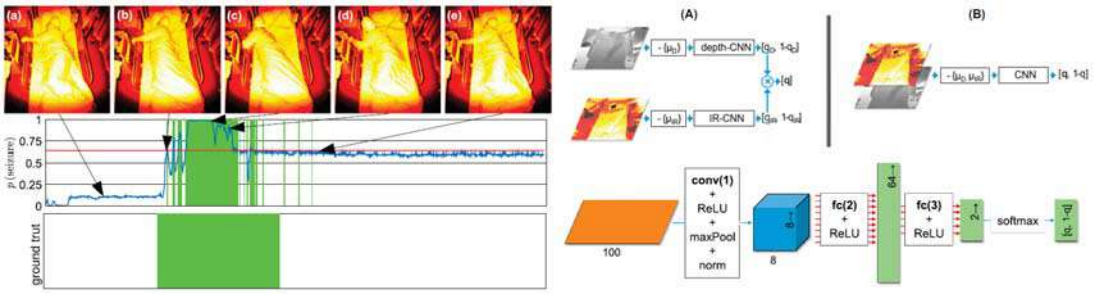


Figure 2.19 Seizure detection based on deep learning: frames from depth and IR cameras and the network architecture proposed. Image adapted from [Achilles et al., 2016c].

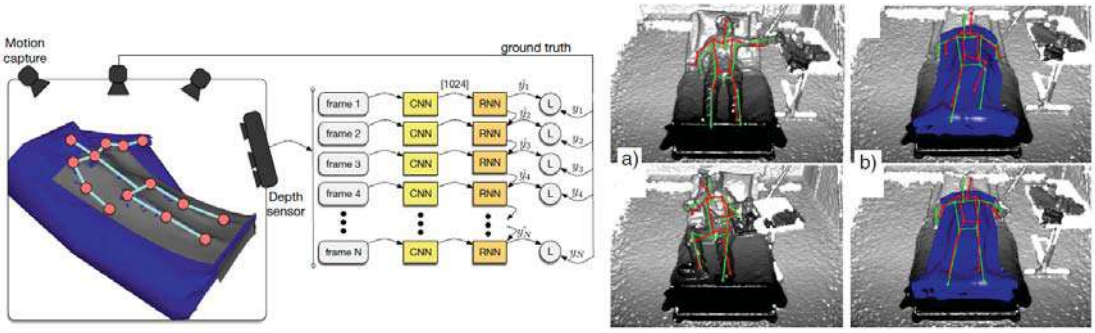


Figure 2.20 Data generation and motion capture of pose estimation based on deep learning. Pose estimation a) without and b) underneath the blanket. Image adapted from [Achilles et al., 2016b].

Although the detection accuracy outperformed state-of-the-art vision-based techniques, it is still unknown whether a higher performance was limited because of the insufficient information from the depth and infrared data or the network architecture designed.

Similarly, Fang et al. [Fang et al., 2018] proposed a spatial temporal gated recurrent unit convolutional neural network to detect seizures. However, the approach is unclear in explaining the clinical purpose of seizure detection in public places and the difference between electrical and clinical onset. The system also seems to detect the full expression of semiology instead of the seizure onset which is stated in the paper "During the start phase, seizure action does not differ from normal behaviour clearly, consequently the model cannot recognise this period."

Achilles et al. [Achilles et al., 2016b], based on their results in previous research in semiology analysis in hidden patient motion [Achilles et al., 2016a] and using a 3D Video-EEG dataset [Cunha et al., 2016a], produced an automatic human pose estimation using deep learning methods. This approach used a combination of a CNN and an RNN to train the data to increase the accuracy of a 3D joint prediction with blanket occlusions. Figure 2.20 illustrates the motion capture framework and the pose estimation results. Although these novels approaches were significant in applying deep learning architectures to detect and track motions from epileptic patients, further research to analyse and verify specific semiology were not explored. Additionally, this approach is

Table 2.3 Summary of advantages of deep learning architectures to assess epilepsy.

Methodology	Advantages
Deep learning	
General deep learning analysis	Automatically learn optimal filters that generate high-level features from raw data (reduces the need for feature engineering, one of the most time-consuming phases of machine learning practice). Benefits of cross-dataset learning (is an architecture that can be adapted to new problems using techniques such as convolutional neural networks, recurrent neural networks, and long short-term memory. Convolution leads to translation invariance. Pooling and subsampling contribute to slight translation and rotation invariance. Larger networks mean higher capability. Has best-in-class performance on problems that significantly outperforms other solutions in multiple domains; this includes speech, language, and vision.
Epilepsy / Semiology	Automated patient detection by region-based convolutional neural networks. Human pose estimation outperformed traditional techniques from sequential prediction frameworks, pictorial structures, and hierarchical and graphical models. Analysis of facial expressions with unconstrained conditions (in-the-wild) such as real patient monitoring. Face detection can be further improved, overcoming limitations of cascade-based and deformable part models. Do not need any facial landmark initialisation. Improved detection robustness of key facial attributes through multi-task learning with heterogeneous but subtly correlated tasks, *e.g.*, head pose estimation and facial attribute inference.

still an expensive alternative in real clinical practice by its dependency on motion capture and depth sensors. The contribution of our work to the knowledge-base of semiology analysis discussed in this thesis has inspired the recent work of Maria et al. [Maia et al., 2019], who proposed a CNN and multilayer perceptron to quantify and classify temporal and extra-temporal seizures. However, it was not clear if the low results of cross-validation were due to the use of infrared data or the training phase of complex deep networks.

Inspired by the breakthroughs in addressing challenges and limitation of conventional counterparts in video analysis, we discuss that the implementation of supervised deep learning can be beneficial in the epilepsy context. Table 2.3 introduces the most important benefits of implementing deep learning for semiology analysis.

2.5 Concluding remarks

According to the literature reviewed, research on vision-based systems in epilepsy is not still widespread. The main aspects considered for modelling clinical manifestations from epileptic patients can be summarised in the following items.

- Many forms of epilepsy have characteristic movements during a seizure, allowing an understanding of the underlying brain networks. Analysis of movements during seizures provides clues as to where the focus of the epilepsy may be, which in turn allows for a successful surgery. Despite the establishment of several automatic methods that help to analyse

brain electrical activity and neuroimaging, seizure semiology is still widely interpreted by visual inspection. Seizure semiology is difficult to characterise and is prone to subjective interpretation and misdiagnosis. The incorporation of quantitative methods or objective information can support the assessment criteria. Years of training and experience is required and having objective quantitative information would assist in developing and formulating a diagnosis in situations where this expertise is unavailable.

- Several groups have developed automatic and semi-automatic methods to quantify and classify epilepsy types such as TLE, FLE, and ETLE. However, these proposals measure a few seizures that involve only limbs and head movements. Researchers have proposed marker-based or marker-free systems. Marker-based systems use non-camera approaches such as accelerometers attached to key body parts to capture the trajectory of each movement, but these sensors need maintenance such as data synchrony and calibration. There are also video systems which use reference markers such as colour-based or reflective materials; however, these markers are affected by the camera position and illumination conditions. Marker-free systems, on the other hand, eliminate the markers by exploiting computer vision techniques and new sensors developments such as depth cameras. But, they have limitations in motion discontinuities and in the detection and tracking of the region of interest in videos. Although methodologies based on 3D vision-based have shown potential to solve some limitations, the tracking of body joints is still under improvement. There is a need to assess limb-free movements such as facial modifications and hand automatisms which requires the detection of fine motions.

- The recent success of deep learning in automatic feature engineering and hierarchical multiple-layer representations is making these techniques useful for biological and medical research. Deep learning is a promising option for assistive medical diagnosis of semiology, where currently automated approaches are adversely affected by challenging, unconstrained conditions encountered in the clinical environment. Examples of these challenges are high data volume analysis, body occlusion, vulnerability to image noise, sensitivity to motion discontinuities, camera position and categorisation of miscellaneous data. Additionally, it is predicted that this automatic quantification could extract information that is imperceptible to the expert inspection. Current proposals that have investigated deep learning in the epilepsy scenario, were significant in targeting the problem of seizure detection. However, they are still an expensive alternative in clinical practice by their dependency on motion capture and depth sensors. Furthermore, seizure detection is not totally relevant for the analysis of semiology and pre-surgical evaluation, but rather the analysis of the patient's behaviour during a seizure.

- The literature review has exposed that there has been considerably less work on using automated analysis in facial expression recognition from epileptic patients. The results of current approaches are documented with data under controlled environments or with data that does not belong to epileptic seizures. Quantification of facial semiology is still limited because of

the immense complexity to detect and track key facial regions. Deep learning could be an alternative to learn facial motions and the temporal relation between video frames in different monitoring scenarios (day and night monitoring) and be robust to severe variations in head pose and occlusion. However, there is no evidence to assess facial and hand semiology with these techniques.

- Video analytic systems based on deep learning are suitable for motion analysis, but their potential is yet to be fully exploited to analyse standard monitoring videos stored in the Epilepsy Monitoring Units. Deep learning proposals for distinguishing epilepsy types are unknown and there is not sufficient research documented about multi-modal systems that quantify and distinguish clinical manifestations from different body locations.

- The implementation of recent advances in deep learning techniques could increase the automatic diagnosis of epilepsy by providing standardised assessments. This opens up new opportunities development of a fully automated system that can differentiate key features from the seizure onset and the propagation using both semiological data and underlying electrographic patterns.

The state-of-the-art in the automatic quantification and evaluation of semiology using traditional and deep learning techniques has been reviewed in this chapter. This chapter also shows that is greatly needed the study of semiology considering multiple clinical manifestations from different body locations. Recent advances in deep learning techniques will be investigated in the contexts of epilepsy to address the challenges exhibited by traditional machine learning techniques. A critical analysis of state-of-the-art methodologies proposed to address each research aim will be documented in each chapter. We have proposed assessing semiology from diverse perspectives using a research dataset of video recordings during the pre-surgical evaluation of patients with drug-resistant epilepsy, as a part of the routine long-term Video-EEG and Stereo-EEG monitoring at the Mater Advanced Epilepsy Unit. Details of the research dataset considered in this thesis are included in the next chapter.

Chapter 3

Data collection and specifications

3.1 Overview

This chapter explains the process to build the research epilepsy database and the public benchmark databases of human motion analysis that support the detection and quantification of semiology. A description of the recruitment process, the video recordings definition, ethical clearance, data access, the specifications of each patient, and codification of the data is included in this chapter. A video dataset annotation of selected seizure recordings is conducted to evaluate the performance of methodologies that address each research aim. Specifications about each public benchmark dataset are also included and organised according to the specific body part or semiology type under analysis including the face, body pose and hands. The research dataset and its connection with each research aim are illustrated in Figure 3.1.

3.2 Epilepsy dataset definition and ethical clearance

To carry out experiments and evaluate the proposed research on semiology, a new dataset of epileptic seizures was developed. It is important to note that there does not exist any publicly available dataset containing semiology recorded during epileptic seizures. In the department of Mater Advanced Epilepsy Unit, Brisbane, Australia (a tertiary referral public epilepsy surgery centre), Dr. Sasha Dionisio and his team receive patients in pre-surgical exploration. These patients present a pharmacoresistant epilepsy: medical therapy failed to sufficiently control seizures, hence the possibility of surgical treatment is considered and patients stay under observation from several days to a few weeks.

Video recordings were captured as a part of the routine long-term Video-EEG and Video-SEEG monitoring protocol in Epilepsy Monitoring Units (EMUs). Participants were undergoing Phase I workup for their drug-resistant epilepsy and were monitored over a time period ranging from 5 to 7 days (epilepsy surgery cases with high seizure frequency at baseline). Each video recordings used in this research represents the seizure event via the captured semiology. Each seizure video was

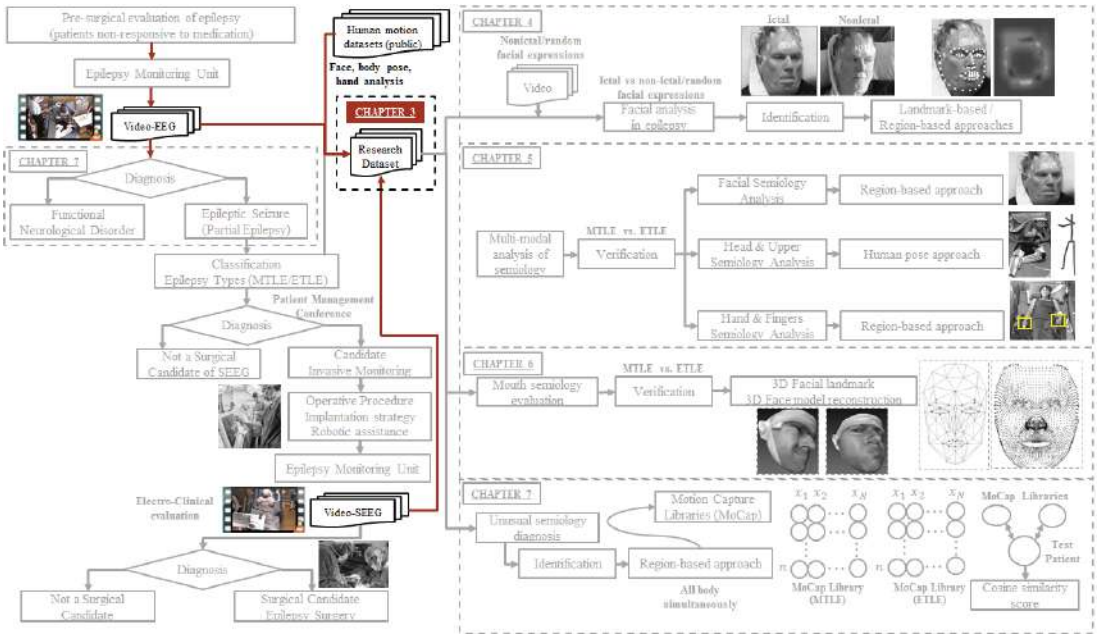


Figure 3.1 Representation of the chapters and their interconnections in the clinical evaluation of epileptic patients. Overview of the research dataset in Chapter 3.

segmented such that it showed from the first epileptic discharge until the full expression of semiology prior to version and convulsion, if it was experienced. This results in video clips of roughly of 1-2 minutes length.

As discussed in Chapter 2, this research focuses on the analysis of mesial temporal lobe epilepsy (MTLE) because it is one of the most common types of drug-resistant epilepsy and its clinical features are well described in the majority of patients. While MTLE can also present with complicated semiology, it generally has been shown to have a more limited repertoire of signs and thus is more easily distinguished from seizures arising from other regions These more homogeneous features assist with the first in the literature research of machine learning and computer vision in epilepsy and allow easier comparison. We define two categories, MTLE and ETLE (extra-temporal lobe epilepsy) in order to categorise the complexity of semiology into two defined regions and allow easier comparison. The term “Extra-temporal” was based on SEEG localization of seizures not arising from the mesial temporal structures. Thus seizures arising from the insula or opercular regions were deemed extra-temporal in origin, even if they eventually involved the temporal lobe later as part of the ictal network. In the MTLE group, patients underwent a SEEG as they were lesion negative on MRI. All patients in the ETLE group underwent SEEG. Seizure freedom was seen in all cases of no less than 2 years to add further weight to localization or categorization. Patients diagnosed with psychogenic non-epileptic seizures (PNES), also known more recently as functional neurological disorder (FND), which mainly are manifestations of psychological distress were also considered for experiments.

Figure 3.2 Video monitoring of epileptic patients. **1.** Epilepsy monitoring units at Mater Centre for Neurosciences. **2.** Selected samples of video recordings during EEG and SEEG monitoring.

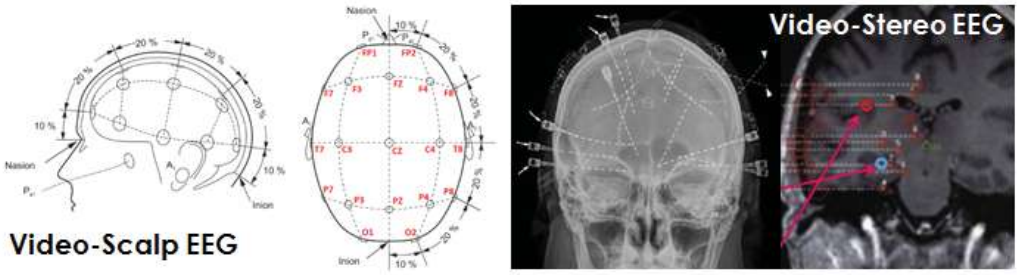


Figure 3.3 Representation of the electrodes used for each type of monitoring: scalp EEG (left) and SEEG (right).

The patient participation in this research did not involve new activities different from the regular monitoring of the patient. All clinical data and patient's records obtained were collected as a part of their existing treatment plan. The neurologist identified patients who could be candidates for the analysis of specific clinical manifestations. We include as many natural clinical settings as possible to ensure a challenging database that reflects real conditions, which includes existing collections of retrospective clinical data and patients that were under evaluation during, approximately, the past two years. This results in a challenging database where some behaviours are very rare, making them difficult to detect. The access to video recordings was distributed along with the duration of this research and the experiments related to each research aim used the available videos at the moment of conducting each experimental procedure. Existing studies, as documented in Chapter 2, have used minimal numbers of patients and seizures, which can be problematic as seizures vary dramatically for even the same type of epilepsy or between seizures of the same patient. Figure 3.2 illustrates the EMUs and samples of video recordings during EEG and SEEG monitoring. The electrodes used for each type of monitoring are depicted in Figure 3.3.

Video recordings of patient monitoring were provided after successful approval of the human ethics clearance. The documentation for this process includes the patient's consent form study and publication (Appendix A.1), the Mater Health Service Human Research Ethics Committee approval (MHS-HREC) (Appendix B.1), the Site Specific Assessment approval (SSA) from the Research

Figure 3.4 Selected samples of video images under different challenging clinical conditions in the epilepsy dataset.

Table 3.1 Selected sample of the codification used to process video recordings with seizures.

Participant Name	Participant Code	Seizure Code	Seizure Date and Hour	Duration
David	1	1	06-01-17 8:34am	72 sec
John	2	1	30-11-16 2:24am	45 sec
John	3	2	01-12-16 3:45pm	54 sec

Governance Office (Appendix B.2), the National Ethics Application form (Appendix B.3) and the executed research agreement signed by Mater Hospital and QUT. The MHS HREC approval was awarded on **January** 3^{rd} **2017** and the SSA approval on **April** 3^{th} **2017**, the date in which the study was notified to start. The information collected for, used in, or generated by, this research cannot be used for any other purpose (See Appendix refappendix:NEAF, Confidentiality/Privacy, Item 3, 21).

3.3 Research datasets

3.3.1 The epilepsy dataset

The epilepsy dataset supports and evaluates all research aims proposed in this thesis (Chapter 4-8). The use of video recording during seizures of epileptic and non-epileptic patients are required to evaluate the robustness and generalisation capability of the methods proposed under real natural clinical environments. We preprocessed each video in order to develop the data for research.

All digitised recorded high-definition videos at a frame rate of 25 frames per second have the format MPEG-2 transport video stream file ("video.M2T"), and resolutions of 1280×720 and 720×480. Figure 3.4 provides examples of video monitoring, the location of the camera and the patient position in different clinical scenarios including day and night monitoring. In some videos, it is evident that the visualisation of the patient is affected by changes in illumination, interaction with clinical staff and hospital equipment which makes a challenging dataset.

The data was collected in an identifiable and re-identifiable format. All videos were re-identified with the idea to protect the personal information of the patient, so in the experiments, the participants are not identified by name. The patients and their seizure events were assigned a consecutive number and labelled with the following structure: *Video Patient "A"."B"* ("A" indicates the code of the patient and "B" the code of the seizure). The codification of the patients is organised in alphabetical order according to the identified information of the name. Table 3.1 demonstrates the video codification for each patient and seizure event. Each video clip is processed into images (25 images per second).

Figure 3.5 Timeline of the data access to seizure recordings.

Images were created in a graphic file format that supports lossless data compression such as the portable network graphics (.PNG). The data represented by each image has the following nomenclature: *Image "A"."B"."Z"* ("A" patient code,"B" seizure code, and "Z" sequence frame of the video).

All seizures from each patient were analysed, assessed and categorised according to semiological features. The observation of semiology was the essence of the first step of this study, where it was crucial to choose well-defined terms to describe different signs. The description of each semiology was established according to the absence or presence of the sign, *i.e.* we identified which patient experienced facial, body and hand semiology.

Figure 3.5 shows the specific dates that we accessed each video set from Mater Hospital and Table 3.2 provides the specification of the patients in each subset of data including the number of seizures used for the experiments documented in each chapter. During this process, we received seizures recorded from new patients or from patients that were already codified in the dataset. In this research, we were able to collect the following total number of seizures for experiments:

Patients with MTLE: 90 seizures from 17 patients.

Patients with ETLE: 71 seizures from 17 patients.

Patients with FND: 10 seizures from 5 patients.

All these seizures were used in the experiments of the hierarchical multi-modal approach, which is the most important evaluation in the analysis of semiology in this thesis (see Section 5.3).

Elementary motor manifestations from the face, head, upper limbs, hand and fingers motions were identified. Examples of this symptomatology or specific isolated signs include unilateral blinking, chewing automatisms, unilateral mouth deviation, postictal nose wiping, ictal pouting, smacking and grimacing, arm flexion, dystonic limb posturing, tonic limb posturing, unilateral immobile limb, fencing posturing, shuddering, ictal head turning, waving, snapping finger, tapping or grabbing and claw position. Figure 3.6 displays selected samples of semiology related to facial, upper limbs and hand motions. Specific details of the analysis of each type of semiology are discussed in the following chapters.

Video dataset annotation

In order to evaluate clinical features, a ground truth of the movement detection that constitutes reliable information to assess the performance of each detector was developed. A process of manual and semi-automatic annotation of body parts of the patient was performed in selected video clips under

Table 3.2 Description of the epilepsy research dataset for experiments. Each video recording represents the seizure event via the captured semiology.

Epilepsy Type	Date of data access	Patient code	Number of seizures used in each chapter						
			CH4	CH5(F)	**CH5(H)**	CH6	CH7(A)	CH7(S)	CH8
	4Apr17,9Mar18	1	2	3	3	-	4	2	-
	4Apr17	2	3	3	3	-	3	1	-
	18Aug17,25Oct17	3	-	3	5	-	3	1	-
	4Apr17,25Oct17	4	2	2	10	-	8	1	-
	25Oct17	5	-	-	5	-	5	1	-
	4Apr17,9Mar18	6	2	2	2	-	3	-	-
	4Apr17,25Oct17	7	4	4	9	-	-	2	-
	18Aug17	8	-	3	3	-	2	1	-
MTLE	25Oct17	9	-	-	7	-	6	1	-
	4Apr17,25Oct17	10	4	4	10	5	7	2	-
	25Oct17	11	-	-	1	-	1	1	1
	4Apr17,25Oct17	12	6	6	10	5	10	2	1
	4Apr17	13	1	1	1	-	1	1	-
	18Aug17,25Oct17	14	-	5	8	-	4	2	-
	18Aug17	15	-	4	4	-	-	1	-
	25Oct17	16	-	-	4	-	-	-	-
	25Oct17	17	-	-	5	-	5	1	-
Total MTLE		**17**	**24**	**40**	**90**	**10**	**62**	**20**	**2**
	18Aug17,25Oct17	1	-	2	5	-	3	2	-
	18Aug17,25Oct17	2	-	3	6	-	7	1	-
	18Aug17,25Oct17	3	-	3	9	-	7	2	-
	18Aug17,25Oct17	4	-	2	6	-	-	1	-
	18Aug17,25Oct17	5	-	1	2	-	2	1	-
	18Aug17,25Oct17	6	-	1	6	-	5	2	-
	25Oct17,9Mar18	7	-	-	1	-	3	-	-
	25Oct17	8	-	-	1	-	1	-	-
ETLE	25Oct17	9	-	-	3	-	-	1	-
	25Oct17	10	-	-	1	-	-	1	-
	25Oct17	11	-	-	4	-	3	1	-
	25Oct17	12	-	-	4	2	4	1	-
	25Oct17,9Mar18	13	-	-	4	4	6	2	-
	25Oct17	14	-	-	9	-	6	1	-
	25Oct17	15	-	-	4	4	4	1	1
	25Oct17	16	-	-	3	-	3	1	-
	25Oct17	17	-	-	3	-	3	1	-
Total ETLE		**17**	**-**	**12**	**71**	**10**	**57**	**20**	**1**
	9Mar18	1	-	-	-	-	1	-	-
	9Mar18	2	-	-	-	-	1	-	-
Aberrant	9Mar18	3	-	-	-	-	1	-	-
	9Mar18	4	-	-	-	-	1	-	-
	9Mar18	5	-	-	-	-	1	-	-
Total Aberrant		**-**	**-**	**-**	**-**	**-**	**5**	**-**	
	16Jun18	1	-	-	-	-	-	1	-
	16Jun18	2	-	-	-	-	-	1	-
FND	16Jun18	3	-	-	-	-	-	5	-
	16Jun18	4	-	-	-	-	-	2	-
	16Jun18	5	-	-	-	-	-	1	-
Total FND		**-**	**-**	**-**	**-**	**-**	**-**	**10**	

CH4: Preliminary analysis of facial semiology; **CH5(F)**: Multi-modal fusion approach (Face, head and pose); **CH5(H)**: Multi-modal hierarchical approach (Face, head, pose, hands) (*All research dataset available*); **CH6**: Preliminary analysis of mouth semiology from a 3D perspective; **CH7(A)**: Aberrant seizure identification (motion libraries captured from all body simultaneously); **CH7(S)**: Identification of seizure disorders (epilepsy vs. FND); **CH8**: Motion signatures (face and hands).

Figure 3.6 Selected samples of semiology recorded in the epilepsy dataset. Face, upper limbs and hand motions.

different conditions of illumination (day and night monitoring) to validate the proposed methodology for each research aim.

We detect the location of the face (see Section 4.4.3), facial landmarks (see Section 4.4.4), the head, neck, shoulders, elbows, wrists (see Section 5.2.4) and hands (see Section 5.3.3). The manual annotation of all these regions of interest was performed in a selected data of 1,500 frames (for patient detection), 7,250 frames (for face detection), 1,086 frames (for facial landmarks), 2,450 frames (for 2D pose estimation), and 1,500 frames (for hand detection).

The location of the face and hands are represented with the detection of the bounding box that contains the body part following the specifications of the FDDB [Jain and Learned-Miller, 2010] and the VIVA hand challenge [Das et al., 2015] datasets. We used the architecture and trained models proposed in [Jiang and Learned-Miller, 2017] and in [Le et al., 2016] to detect the face and hands in the selected videos of our dataset, respectively. Then, a process of manual correction of the bounding box is performed. Figure 3.9 and Figure 3.23 illustrate examples of the face and hands annotations, respectively. The structure of landmarks annotation follows the distribution of the Multi-PIE 68 points mark-up [Gross et al., 2010] (See Figure 3.12 for the annotation of the 300-W dataset). A semi-automatic annotation methodology for annotating massive face datasets was implemented [Sagonas et al., 2016, Sagonas et al., 2013b]. We adopt the architecture proposed in [Baltrušaitis et al., 2016] to estimate the facial landmarks. Then, manual correction of each landmark is performed using a developed interface as it is displayed in Figure 3.7. Similarly, the detection of key body parts as shown in Figure 3.8 is implemented using the architecture in [Wei et al., 2016]. The interface developed for facial landmarks is also used to correct manually the location of the body parts.

3.3.2 The human motion analysis datasets

Public datasets used for human motion analysis listed in this section, support the analysis of epileptic patients by demonstrating the benefits of our proposed methodologies based on deep learning of cross-

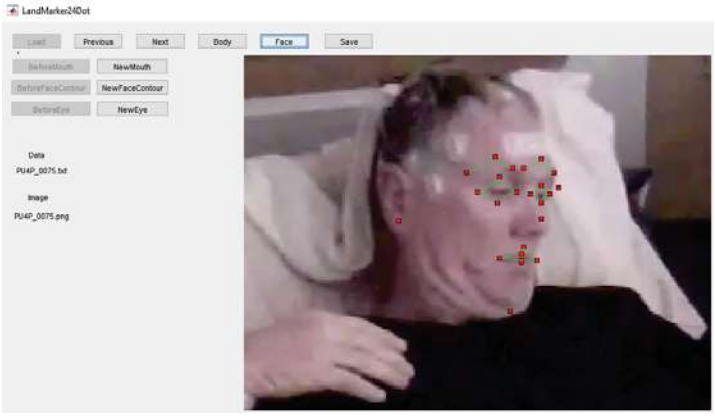


Figure 3.7 Software implemented for the semi-automatic annotations of facial landmarks.

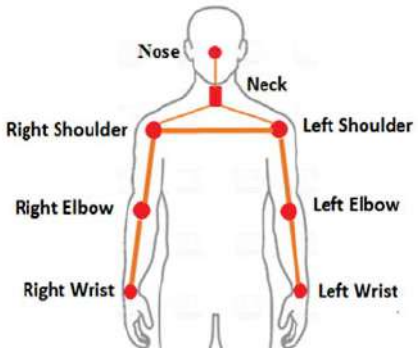


Figure 3.8 Key body points selected for the manual annotation of body pose.

dataset learning. This aims to adapt models from one domain to another, *i.e.*, to adapt well-established models pre-trained with public datasets that have been used with success in human motion analysis to the epilepsy domain. This reduces the need for large image training datasets to train a new model from scratch [Goodfellow et al., 2016].

Baseline architectures for human motion understanding were trained and tested in a wide variety of public benchmark datasets with realistic and challenging images. Datasets used by state-of-the-art methodologies proposed to support each research aim are summarised in the following tables with their specifications and we provide examples of images contained in selected datasets.

Face analysis datasets

Datasets used for face detection and facial landmark estimation under unconstrained conditions are described in Table 3.3, Table 3.4 and Table 3.5. Datasets for the experiments conducted of 3D face reconstruction are summarised in Table 3.6.

Human pose analysis datasets

The most representative datasets used in the implementation of 2D human pose estimation, pose tracking and 3D human pose estimation are listed in Table 3.7, Table 3.8 and Table 3.9, respectively.

Hand analysis datasets

Datasets adopted for experiments for hand detection and pose estimation are mentioned in Table 3.10 and Table 3.11.

Table 3.3 Summary of benchmarking datasets for face detection and tracking.

Author	Name	#Images	#Faces	Characteristics
[Yang et al., 2016]	WIDER Face	32,203	393,703	Colour (Figure 3.10)
[Klare et al., 2015]	IJB-A	24,327	49,759	Colour
[Yan et al., 2014]	PASCAL Face	851	1,341	Colour
[Zhu and Ramanan, 2012]	AFW	205	473	Colour
[Jain and Learned-Miller, 2010]	FDDB	2,846	5171	Colour (Figure 3.9)

Table 3.4 Summary of benchmarking datasets for 2D facial landmarks.

Author	Name	#Images	#Landmarks	Characteristics
[Zhang et al., 2016c]	MAFL	19,000	5	CelebA dataset
[Sagonas et al., 2016, Sagonas et al., 2013a]	300-W	4,000	51-68	Colour(Outdoor/Indoor) (Figure 3.12)
[Shen et al., 2015]	300-VW	218,595	68	114 Videos
[Zhang et al., 2014c]	MTFL	10,000	5	Colour
[Burgos-Artizzu et al., 2013]	COFW	1,007	29	Colour, Occlusions
[Belhumeur et al., 2013]	LFPW	1,432	29-35	Colour, Gray-scale
[Zhu and Ramanan, 2012]	AFW	205	6	Colour
[Le et al., 2012]	HELEN	2,330	194	Colour
[Köstinger et al., 2011]	AFLW	25,993	21	Colour, from FlickR (Figure 3.11)

Table 3.5 Summary of benchmarking datasets for 3D facial landmarks.

Author	Name	#Images	Characteristics
[Bulat and Tzimiropoulos, 2017b]	LS3D-W	230,000	
[Liu et al., 2017]	300W-LP-3D	61,225	Extension of 300W [Sagonas et al., 2013a]
[Zhu et al., 2016]	AFLW2000-3D	2,000	Extension of AFLW [Köstinger et al., 2011] (Figure 3.13)
[Jourabloo and Liu, 2015]	AFLW-LFPA	1,299	Extension of AFLW [Köstinger et al., 2011] (Figure 3.14)

Table 3.6 Summary of benchmarking datasets for 3D face reconstruction.

Author	Name	#Images	Characteristics
[Zhu et al., 2016]	AFLW2000-3D	2,000	Extension of AFLW [Köstinger et al., 2011]
[Jourabloo and Liu, 2015]	AFLW-LFPA	1,299	Extension of AFLW [Köstinger et al., 2011]
[Zhang et al., 2014b]	BP4D-Spontaneous		41 subjects (Figure3.15)
[Bagdanov et al., 2011]	Florence	212	53 subjects (Figure 3.16)
[Vijayan et al., 2011]	3D-TEC	428	214 subjects
[Savran et al., 2008]	Bosphorus	4,666	105 subjects
[Yin et al., 2008]	BU-4DFE		101 subjects
[Yin et al., 2006]	BU-3DFE	2,500	100 subjects
[Blanz and Vetter, 1999]	3DMM		200 heads of young adults

Table 3.7 Summary of benchmarking datasets for 2D pose estimation.

Author	Name	# Images	Characteristics
[Lin et al., 2014]	MS COCO Keypoints	105,698	Diverse
[Andriluka et al., 2014]	MPII Single person	26,429	491 human activities (Figure 3.18)
[Andriluka et al., 2014]	MPII Multi-person	14,993	491 human activities
[Dantone et al., 2013]	FashionPose	7,305	Fashion blogs
[Sapp and Taskar, 2013]	FLIC	5,003	Feature movies
[Johnson and Everingham, 2011]	LSP	2,000	Sports (8 act.) (Figure 3.17)

Table 3.8 Summary of benchmarking datasets for 2D pose tracking.

Author	Name	# Images	Characteristics
[Andriluka et al., 2017]	PoseTrack	153,615	Diverse
[Iqbal et al., 2017b]	Multi-person PoseTrack	16,219	Diverse
[Charles et al., 2016]	YouTube Pose	5,000	Diverse
[Zhang et al., 2013]	Penn Action	159,633	sports (15 act.) (Figure 3.19)
[Jhuang et al., 2013]	JHMDB	31,838	Diverse (21 act.)
[Sapp et al., 2011]	Video Pose 2.0	1,286	TV series

Table 3.9 Summary of benchmarking datasets for 3D pose estimation.

Author	Name	#Videos	#Subjects	Characteristics
[CMU, 2017]	CMU Lab	2,605	109	23 actions
[Mehta et al., 2017]	MPI-INF-3DHP	N/A	8	8 actions
[Ionescu et al., 2014]	Human3.6m	1,376	11	15 actions, 3.6×10^6poses (Figure 3.20)
[Rohrbach et al., 2012]	MPII	44	12	65 actions, Cooking Activities
[Sigal et al., 2010]	HumanEva-I&II	56	4	6 actions (Figure 3.21)

Table 3.10 Summary of benchmarking datasets for hand detection.

Author	Name	# Hand instances	Characteristics
[Das et al., 2015]	VIVA Hand Challenge	11,000	Driving settings (Figure 3.23)
[Mittal et al., 2011]	Oxford	13,050	Frames from films (Figure 3.22)

Table 3.11 Summary of benchmarking datasets for 2D-3D hand pose estimation.

Author	Name	# Hand instances	Characteristics
[McKee et al., 2018]	NZSL	1,500	Sign language to tell stories
[Zimmermann and Brox, 2017]	Rendered Hand Pose	2,728	Generated synthetically
[Mueller et al., 2017]	EgoDexter	3,190	4 test video (Figure 3.24)
[Sridhar et al., 2016]	Dexter+Object	3,145	6 test video (Figure 3.25)
[Zhang et al., 2016a]	Stereo Hand Pose	18,000	6 pairs of stereo sequences
[Andriluka et al., 2014]	MPII (hand cropped)	1,300	491 human activities

Figure 3.9 Selected images of the FDDB dataset. Image adapted from [Jain and Learned-Miller, 2010].

Figure 3.10 Selected images of the WIDER Face dataset. Image adapted from [Yang et al., 2016].

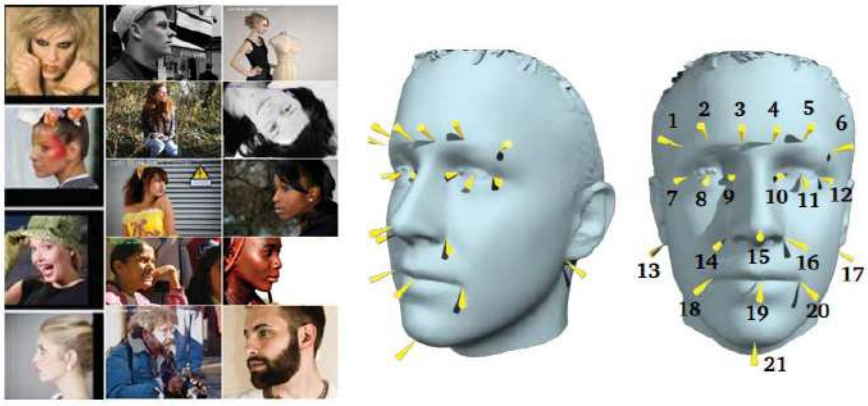


Figure 3.11 Landmark mark-up in the AFLW dataset. Image adapted from [Köstinger et al., 2011].

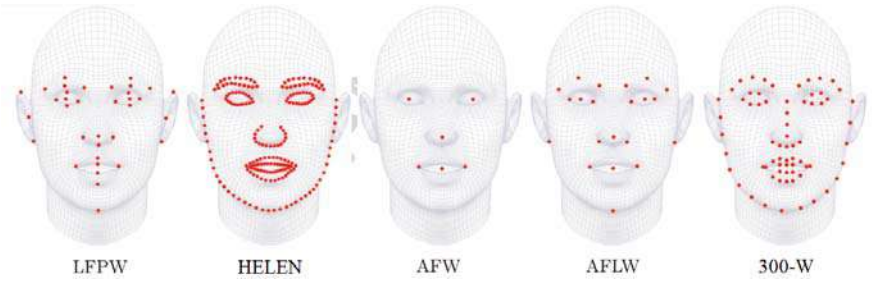


Figure 3.12 Facial landmarks configuration of existing datasets: LFPW, HELEN, AFW, AFLW and 300-W (68-landmarks). Image adapted from [Sagonas et al., 2016].

Figure 3.13 Selected images of 3D fitting and landmarks on the AFLW2000-3D dataset. Image adapted from [Zhu et al., 2016].

Figure 3.14 Selected images of 3D dense shapes and landmarks of the AFLW-LFPA dataset. Image adapted from [Jourabloo and Liu, 2016].

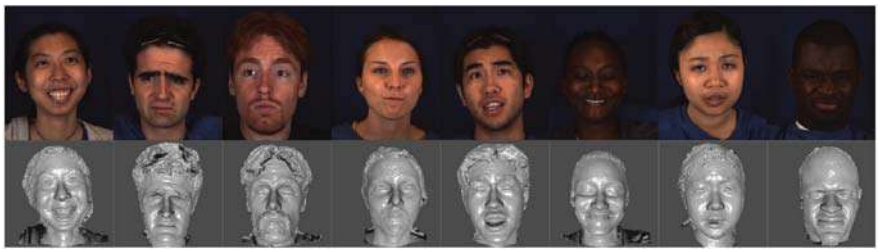


Figure 3.15 Selected images of 3D model face of the BP4D-Spontaneous dataset. Image adapted from [Zhang et al., 2014b].

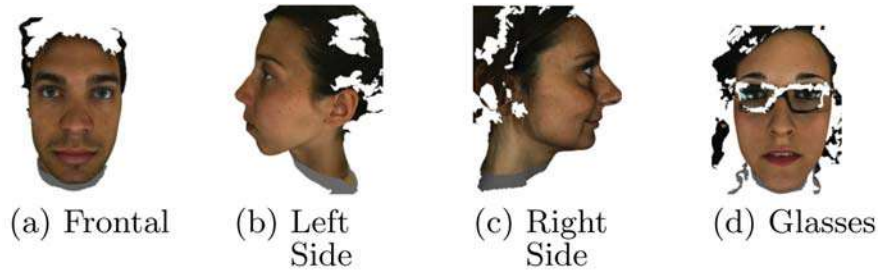


Figure 3.16 Selected images of 3D model face of the Florence dataset. Image adapted from [Bagdanov et al., 2011].

Figure 3.17 Selected images of the LSP dataset. Image adapted from [Johnson and Everingham, 2011].

Figure 3.18 Selected images of the MPII dataset. Image adapted from [Andriluka et al., 2014].

Figure 3.19 Selected images of the Penn Action dataset. An example of seed volume and its three nearest patches in the training set. Image adapted from [Zhang et al., 2013].

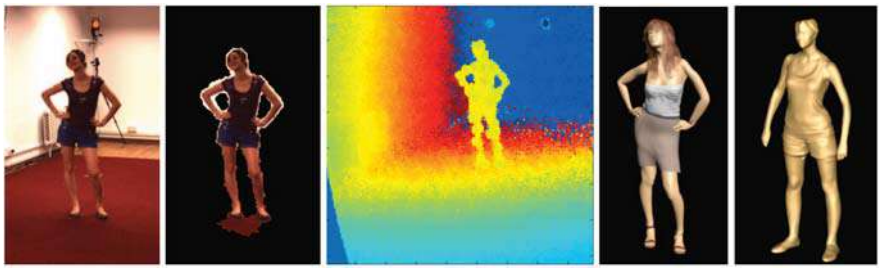


Figure 3.20 Selected images of the Human3.6 dataset. From left to right: RGB image, person silhouette, depth data, 3D pose data, accurate body surface obtained using a 3D laser scanner. Image adapted from [Ionescu et al., 2014].

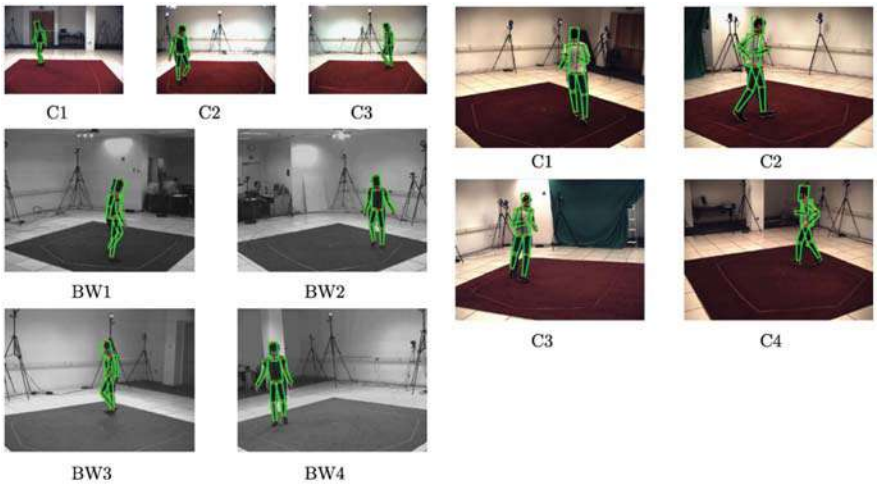


Figure 3.21 Selected images of the HumanEva dataset. HumanEva I on the left side and HumanEva II on the right side. Image adapted from [Sigal et al., 2010].

Figure 3.22 Selected images of the Oxford dataset. Hand dataset with the bounding box annotations overlaid. Image adapted from [Mittal et al., 2011].

Figure 3.23 Selected images of the VIVA hand challenge dataset. Realistic driving scenarios. Image adapted from [Das et al., 2015].

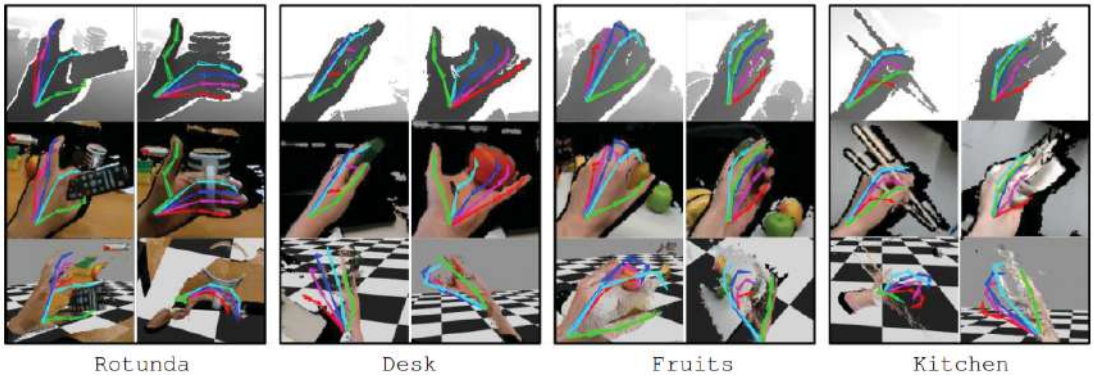


Figure 3.24 Selected images of the EgoDexter dataset. Image adapted from [Mueller et al., 2017].

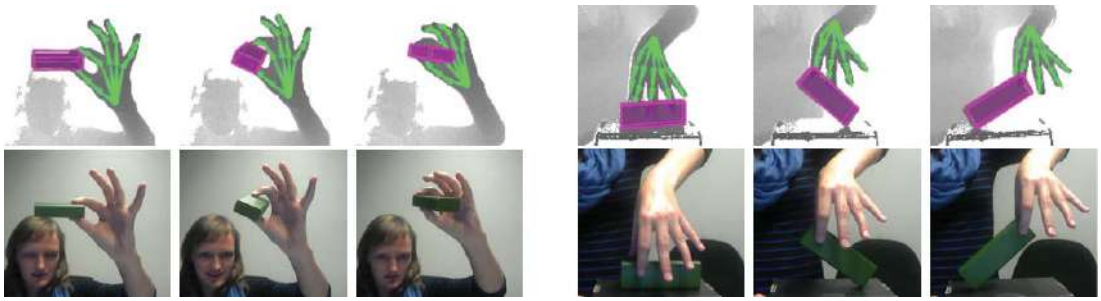


Figure 3.25 Selected images of the Dexter + Object dataset. Image adapted from [Sridhar et al., 2016].

3.4 Concluding remarks

To conduct the analysis of semiology, we created a research dataset of video recordings during the pre-surgical evaluation of patients, as a part of the routine long-term Video-EEG and Stereo-EEG monitoring. We focus on the analysis of MTLE because it is one of the most common types of drug-resistant epilepsy and its clinical features are well described in the majority of patients. Two epilepsy types, MTLE and ETLE, were chosen in order to categorise the complexity of semiology into two defined regions which allow easier comparison. Extra-temporal means seizures not arising from the mesial temporal structures. Assessment of all seizures recorded was conducted by clinical experts to guarantee the quality of the supervised models and for performance comparison.

Each video recording was segmented from the first epileptic discharge until the full expression of semiology prior to version and convulsion, if it was experienced. This resulted in videos clips of roughly 1–2 min length, captured at 25 frames per second. A total of 161 seizures from 34 patients with epilepsy were analysed, which includes existing collections of retrospective clinical data and patients that were under diagnosis during this research. Table 3.2 provides the specification of the patients in each subset of data including the number of seizures used for the experiments documented in each chapter.

Benchmark datasets for human motion evaluation are also integrated into our research dataset, to support the quantification of semiology by exploiting the ability of deep learning known as cross-dataset learning. This aims to adapt models from one domain like biometric facial recognition into the epilepsy domain. The process executed in each research aim is based on a critical analysis of broad benchmarking architectures for human motion analysis and how we could adjust and improve them in the epilepsy scenario.

In the following chapter, we endeavour to develop quantitative methods that characterise semiology from the facial expression based on deep learning. While some attempts at automating the semiology of facial expressions have been made, the field is still largely unexplored. We are motivated by the high number of semiology that came from the face in patients with MTLE and the lack of the literature to analyse these features. We have concentrated our techniques to distinguish between ictal and nonictal/random facial expressions during the patient's monitoring.

Chapter 4

Facial analysis in epilepsy

4.1 Overview

The literature review has revealed that there has been considerably limited work on developing automated techniques to assess facial semiology for epilepsy. This chapter proposes a new approach to detect clinical manifestation from facial modifications aiming to solve current challenges in video analytics under challenging real-world conditions of a hospital setting. This chapter introduces the first evaluation of facial semiology for epilepsy using methodologies that have solid evidence of improvement in computer vision such as deep learning architectures. This chapter's research aim and its relationship with the thesis is illustrated in Figure 4.1.

In epilepsy, certain facial modifications are more commonly exhibited (although not exclusive), including unilateral blinking, eye deviation, chewing automatisms, fear expression, disgust, unilateral mouth deviation, and postictal nose wiping [Ataoğlu et al., 2015, Noachtar and Peters, 2009, Fogarasi et al., 2007]. Figure 4.2 illustrates selected samples of sequences with facial expressions recorded during epileptic seizures from the epilepsy dataset.

Although some attempts at quantifying semiology of facial expressions have been described in Chapter 2, the field is largely unexplored due to the immense complexity in detecting and tracking key facial regions, where the face may often be obscured by electrodes, bedding, inadequate camera capture and positioning, poor illumination, and movements during seizures.

Researchers have demonstrated successful performance in analysing videos using deep learning, outperforming traditional techniques in emotion recognition and facial expression analysis [Rodriguez et al., 2017, Ghasemi et al., 2016, Lopes et al., 2017]. However, despite its advantages, there has not been an exploration and improvement of this technology for monitoring and detecting facial changes during epileptic seizures. While automated detection of EEG signals based on deep learning exists to help identify seizures [Thodoroff et al., 2016], there are no frameworks which detect ictal changes in semiology or clinical manifestations during seizures as discussed in Chapter 2.

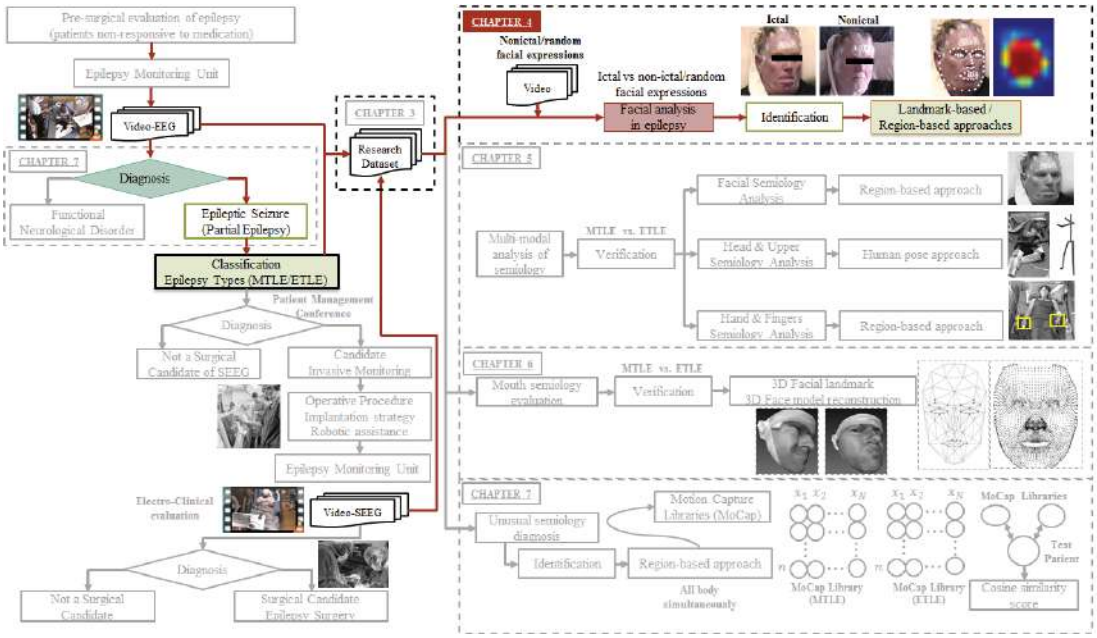


Figure 4.1 Representation of the chapters and their interconnections in the clinical evaluation of epileptic patients. Overview of the research aim in Chapter 4.

In this chapter, we propose deep learning techniques to distinguish between ictal and non-ictal/random facial expressions in patients with mesial temporal lobe epilepsy (MTLE), which is the most common type of semiology for this epilepsy group. Deep learning architectures quantify the various semiological patterns from a selected subset of patients with MTLE in order to detect ictal patterns.

The chapter is distributed as follows. Section 4.2 describes two approaches to quantify facial semiology; Section 4.3 explains the computer-vision and deep learning techniques to detect and model facial motions related to each approach; Section 4.4 describes the dataset used for this particular experiment and the results to distinguish clinical manifestation from natural facial expressions; Finally, Section 4.5 includes the discussion of the facial analysis and limitations.

This chapter is supported by the following published manuscript:

- **D. Ahmedt-Aristizabal**, C. Fookes, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, Deep facial analysis: A new phase I epilepsy evaluation using computer vision, *Epilepsy & Behavior*, 82 (2018) 17-24.

4.2 Strategies to identify facial semiology in epilepsy

The proposed framework estimates whether a facial expression sequence has an ictal pattern of MTLE. A facial expression can be observed as a dynamic variation of key parts, which are fused to form the variation of the whole face. The framework aims to capture such dynamic variations of facial physical

Figure 4.2 Selected samples of facial semiology from the epilepsy dataset. From up to down, patients experience high blinking frequency, fear expression and blinking, mouth movement, and unilateral eye blinking.

structure from consecutive frames. In order to validate the research clinical hypothesis which states that similar semiological patterns are sufficient to categorise patients with MTLE, an experimental design displayed in Figure 4.3 is proposed to assess semiology from facial movements. To analyse facial expressions from patients, two methods are considered: landmark-based and region-based methods. The landmark-based method (geometry information) is based on a detector of anatomical points of reference in the face for measurement and quantification of facial motions over time. In the region-based method, spatio-temporal features are extracted to model the variability in morphological and contextual factors of the whole face by employing a combination of CNNs and LSTMs. Spatial features present information in the facial expressions of a single video frame. On the other hand, temporal features exhibit the relationship between facial expressions revealed in consecutive video frames.

These approaches are selected to evaluate and verify which method excels at detecting facial semiology from patients with MTLE in real conditions of clinical monitoring.

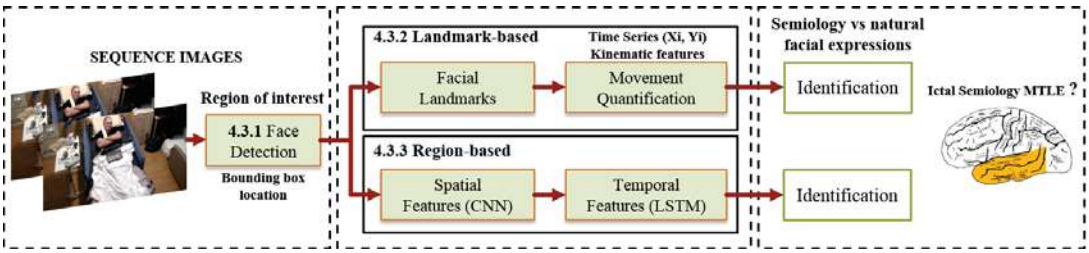


Figure 4.3 Approaches proposed for the automatic analysis of facial semiology: a landmark-based and a region-based approach.

The *landmark-based method*, which is an important representation of facial expressions, is chosen to capture the kinematic information of specific landmarks located in the mouth and eyes to analyse the frequency and amplitude of the movements. This method models the face changes over time, which is useful to capture the dynamic variation of the facial physical structure. The movement detection is represented by the two-dimensional movements, x-axis and y-axis, of each landmark during the facial expression. The resulting representation $[x, y]$ of the time series of each marker are used to perform the movement quantification using metrics based on the temporal domain to detect facial changes in acceleration or displacement of each landmark among frames.

The *region-based method*, on the other hand, extracts spatial features from the whole face. This method aims to capture the dynamic change of facial physical structure from consecutive frames by exploiting temporal information as the facial expression changes. The region-based method extracts spatio-temporal features from the raw frames using an end-to-end deep learning model by implementing a CNN-LSTM architecture. While a CNN excels at learning spatial features, an LSTM is ideal for learning the temporal features and the long-term dependencies present within sequential data.

The landmark-based approach is more intuitive to humans, but it is more related to handcrafted features since we predefine some landmark points based on humans' assumption about facial movement. In contrast, the region-based approach is less intuitive because we may not know what happens inside a CNN and an LSTM, but it eliminates the need for feature engineering.

Both landmark-based and region-based approaches are trained in a supervised manner. This methodology is divided into two main phases: feature extraction and classification [Karpathy et al., 2014]. The feature extraction process can be viewed as finding a set of measured data which adequately represent the information content of an observation. The classification phase concerned to which of a set of categories or class a new observation belongs, by a training and validation set whose class membership is known.

Inadequate training data hinder the validation of algorithms that could quantify semiology. For this reason and as discussed in Chapter 3, to exploit the ability of cross-dataset learning of deep learning architectures in the epilepsy scenario, well-established models pre-trained with public datasets that have been used with success in facial research were considered (see Section 3.3.2). In this situation, we aim to adapt models use for face verification to analyse facial semiology.

We analyse a broad range of available techniques for face detection and facial landmark estimation with the purpose of training our framework to recognise the patient face and its regions of interest (ROI). Subsequently, we implement and improve upon the selected model for the epilepsy evaluation task by comparing their performance on the epilepsy dataset and conducting a process of fine-tuning. Once the features are extracted from the video clips recorded, training and validation are performed with the aim to classify facial semiology from natural facial expressions.

4.3 Quantification of facial semiology

4.3.1 Face detection

The initial step to detect and quantify facial features for both proposed approaches is to detect the area of interest, *i.e.*, the face. This process should be robust to real-world conditions that occur during clinical monitoring including scale and pose changes, occlusions, and illumination variations. Traditional techniques used for face detection are represented by cascade-based, deformable part models, local binary patterns and histograms of oriented gradients. However, they have limitations with facial datasets under in-the-wild conditions [Kazemi and Sullivan, 2014]. The widely used Viola and Jones algorithm [Viola and Jones, 2001], provides real-time face detection, but only performs well on frontal and well-lit face images [Pediaditis et al., 2011, Pediaditis et al., 2012a]. This method coupled with a traditional tracking algorithm [Tomasi and Kanade, 1991], have been used in numerous libraries and toolboxes for imaging analysis. Matlab and its computer vision system toolbox [Guide, 1998] provides functions to perform the face detection and tracking based on the Viola-Jones, Kanade-Lucas-Tomas, continuously adaptive mean shift (CAMShift) and Kalman filtering algorithms. Similarly, OpenCV [Bradski et al., 2000], an open source computer vision library, has a practical implementation of the Viola-Jones face detector using cascade classifiers. Figure 4.4 illustrates qualitative results on how these traditional methods fail in detecting faces under natural clinical scenarios of epilepsy monitoring.

With the advent of deep learning and the development of large annotated datasets, face detection has been significantly improved [Yang et al., 2018a]. Recent deep learning models with high modelling capacity can capture non-linear mappings between intrinsic facial features and facial muscle motions [Unzueta et al., 2014], showing impressive results and overcoming limitations of traditional techniques. We evaluated representative state-of-the-art face detectors based on deep learning models and compared their performance on the epilepsy dataset. These architectures were trained and assessed with publicly datasets of faces such as FDDB and WIDER Face datasets (see Table 3.3). Different metrics are assessed such as the average precision (AP), computational cost, documentation, and training. Table 4.1 summarised these architectures.

Since the remarkable success of deep convolutional neural networks in image classification, numerous efforts have been made to port CNNs to produce state-of-the-art performances on face detection [Jiang and Learned-Miller, 2017]. Methodologies for detecting objects are extensive from using object proposals to deep networks, and region-based CNNs drive recent advances in object

Figure 4.4 Selected samples of face detection with traditional detectors. Upper: Performance using the Matlab Toolbox (yellow boxes). Lower: Performance with the OpenCV Toolbox (red boxes).

detection. Although these methodologies are computationally expensive, their cost has been decimated thanks to sharing convolutions across proposals. Pioneer models of object detection based on region-based CNN features are R-CNN [Girshick et al., 2014], SP Pnet [He et al., 2014], Fast R-CNN [Girshick, 2015] and Faster R-CNN [Ren et al., 2015, Ren et al., 2017]. Another famous model is the YOLO architecture [Redmon et al., 2016] which treats the object recognition task as a unified regression problem, different from the models in the R-CNN family which learn to solve a classification task. Among these approaches, the Faster R-CNN architecture [Ren et al., 2017] has achieved state-of-the-art object detection by replacing external object proposals with a region proposal network (RPN). The face detector proposed by Jiang and Learned-Miller [Jiang and Learned-Miller, 2017], which is based on the Faster R-CNN architecture, has been found to be a suitable option to conduct the detection of the patient's face because of its precision, reduced running time and documentation to ensure full reproducibility. The precision of the method is considered as the number of items correctly labelled as belonging to the positive class. Figure 4.5 depicts the qualitative results in a benchmark dataset.

Faster Region-based Convolutional Network (R-CNN) comprises two modules: a Region Proposal Network (RPN) which generates a set of object proposals, *i.e.*, a set of rectangles; and an Object Detection Network, based on the Fast R-CNN detector which refines the proposal location [Girshick, 2015]. In the RPN, the CNN architecture considered is the VGG-16 model [Simonyan and Zisserman, 2014]. The RPN can be trained in an end-to-end manner using back-propagation and stochastic gradient descent (SGD) [LeCun et al., 1989]. This architecture is illustrated in Figure 4.6. In the region proposal network, the position of the sliding window provides localisation information regarding the image and the box regression provides finer localisation

Table 4.1 Selected benchmarking techniques for face detection and their performance on the FDDB dataset.

Author	Title	AP/FDDB
[Jiang and Learned-Miller, 2017]	Face detection with the Faster R-CNN	0.960
[Zhang et al., 2016b]	Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks (MTCNN)	0.950
[King, 2012]	DLib C++ (v19.2, Release Oct 10, 2016)	0.920
[Chen et al., 2016]	Supervised transformer network for efficient face detection	0.920
[Ranjan et al., 2015]	A deep pyramid deformable part model for face detection (DP2MFD)	0.917
[Shuo Yang and Tang, 2015]	From Facial Parts Responses to Face Detection: A Deep Learning Approach (Faceness)	0.909
[Mathias et al., 2014]	Face detection without bells and whistles (Head Hunter-DPM)	0.880
[Yang et al., 2014]	Aggregate channel features for multi-view face detection (ACF-Multiscale)	0.860
[Li et al., 2015]	A Convolutional Neural Network Cascade for Face Detection (Cascade CNN)	0.859
[Farfade et al., 2015]	Multi-view Face Detection Using Deep Convolutional Neural Networks (DDFD)	0.840

information regarding this sliding window as it is displayed in Figure 4.7. The algorithm uses N anchor boxes at each location, where regression gives offsets from anchor boxes and classification gives the probability that each regressed box shows an object. An anchor is labelled as positive if the anchor is the one with highest intersection-over-union (IoU) overlap with a ground-truth box.

The loss function for learning region proposals is defined as,

$$L(p_i,t_i) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i,p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i,t_i^*), \tag{4.1}$$

where p_i (predicted probability of being an object for anchor i), t_i (coordinates of the predicted bounding box for anchor i), L_{cls} (Log loss), p_i^* (ground truth object label), λ (In practice $\lambda = 10$, so that both terms are roughly equally balanced), L_{reg} (smooth L1 loss), and t_i^* (true box coordinates). Additionally, the object detection network loss can be written as,

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1]L_{loc}(t^u,v), \tag{4.2}$$

where p (predicted class scores), u (true class scores), t^u (true Box coordinates), v (predicted box coordinates), L_{cls} (Log loss), and L_{reg} (smooth L1 loss). The positive samples are defined as those whose IoU overlap with a ground-truth bounding box is > 0.5.

The face detector [Jiang and Learned-Miller, 2017] was trained on the large-scale WIDER Face dataset, and used the pre-trained ImageNet model VGG-16 [Simonyan and Zisserman, 2014] to generate high-quality object proposals. The authors adopted the approximate joint learning strategy. This method trains the RPN module jointly with the Fast R-CNN network, rather than alternating between training the two. The performance was evaluated on the widely-used FDDB and IJB-A datasets (see Table 3.3).

Figure 4.5 Selected samples of qualitative results of the face detector in the FDDB dataset. Ground-truth annotations (Green boxes) and automated detection (Red boxes). Image adapted from [Jiang and Learned-Miller, 2017].

Figure 4.6 The Faster R-CNN unified network for object detection and selected samples of object detection. Image adapted from [Ren et al., 2017].

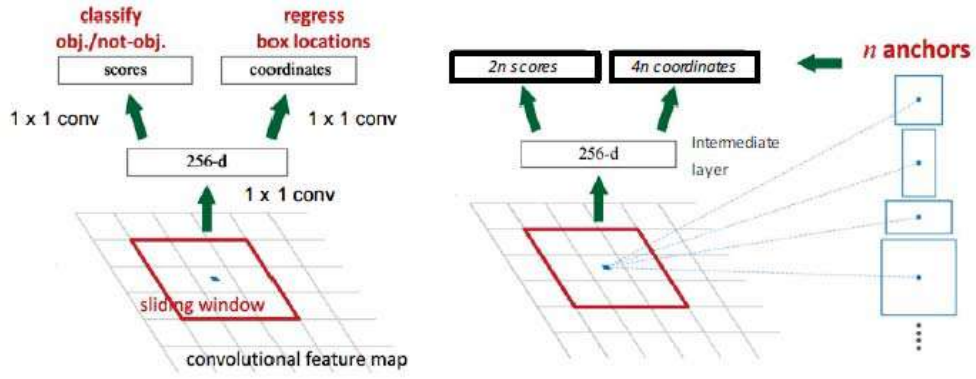


Figure 4.7 The region proposal network (RPN) in the Faster R-CNN network. Image adapted from [Ren et al., 2017].

Figure 4.8 Selected samples of objects and persons captured in the video monitoring that affect the facial semiology analysis.

Figure 4.9 Selected sample of patient detection. All images of a sequence are cropped with the area of the patient detected.

We fine-tuned the face detector to improve the performance of images recorded during night monitoring (infrared camera) which is present in the epilepsy dataset. The fine-tuning was performed only on the last fully connected layer of the VGG-16 architecture to preserve the earlier learned filters. This process was performed following the instructions in [Ren et al., 2017, Jia et al., 2014]. The framework is implemented with Python, Caffe [Jia et al., 2014] and OpenCV libraries [Bradski et al., 2000].

Patient detection

Objects and other people inside a patients' room including monitoring equipment, furniture, clinical staff and family members may affect the performance of the facial analysis by providing false face detections of the patient's face, which is shown in Figure 4.8. Therefore, a preceding phase of patient detection is considered. This ensures that the facial features used in the analysis come from the patient and not from any family members or physicians also visible in the videos. The region of interest is defined as the location of the bed that contains the patient, and it is expected that the patient will remain inside this boundary during a seizure. Detection is performed only in the first frame, to ensure that the extracted features for each video segment are extracted from the same region and reduce jitter.

We perform patient boundary detection using the Faster R-CNN approach [Ren et al., 2017] fine-tuned on the COCO dataset [Lin et al., 2014], which includes training samples of bed objects. The bounding box detected that contains the patient is expanded with an offset of 20% on each side to avoid any extremities of the patient being located outside of this boundary because of movements during a seizure. With the location of the final bounding box, we crop and resize the images to roughly make the samples of each patient the same scale to create the dataset that was used in the facial semiology quantification as it is shown in Figure 4.9.

4.3.2 Landmark-based approach

Once the face detection is completed, facial landmark estimation can be performed. Facial landmark estimation and alignment are traditional pre-processing steps for facial analysis, which aims at aligning a face image and locating fiducial points, such as the eye, nose, mouth, and contours.

During the last decades, numerous methodologies have been proposed for the problem of facial landmark points localisation. Facial landmark detection has been extensively studied since Dr Cootes' Active Shape Model (ASM) in the 1990s [Cootes et al., 1994]. These methods can be roughly divided into two categories: generative and discriminative. Generative techniques aim to find parameters that maximise the probability of the test image being generated by the model (active appearance models (AAM) and pictorial structures). Discriminative techniques or part-based approaches use discriminative response map functions (active shape models (ASMs), constrained local models (CLMs), deformable part models (DPMs), cascade of regression functions and random forests) [Jin and Tan, 2017]. These proposals mainly refine the prediction of the location of the landmarks iteratively from an initial estimation, which is highly relevant to the initialisation. Furthermore, facial landmark detection has been impeded by the problems of pose variation, occlusion, illumination, resolution and background clutter [Jourabloo and Liu, 2015]. Therefore, we aim to evaluate state-of-the-art methodologies that address these challenges based on deep learning techniques and two main perspectives, 2D and 3D methods, as proposed in thorough reviews of face alignment in-the-wild [Wang et al., 2018a, Jin and Tan, 2017].

2D methods for facial landmark estimation

Since the pioneering method of Sun et al. [Sun et al., 2013], deep convolutional neural networks have been successfully used in facial landmark localisation, overcoming limitations of traditional techniques based on generative and discriminative methods. Deep learning architectures are accurate because the geometric constraints among facial points are implicitly utilised, a huge amount of training data can be leveraged and they do not need facial landmark initialisation [Burkert et al., 2015].

Recent approaches investigated the possibility of improving the detection robustness through multi-task learning with heterogeneous but subtly correlated tasks, *e.g.*, facial landmark distribution and head pose estimation including the spatial rotations yaw, pitch, and roll. [Yang et al., 2015, Zhang et al., 2016c]. We assessed a number of benchmark methods for facial keypoint detection based on 2D deep learning approaches, analysing metrics such as the mean error, failure rate and qualitative performance to handle naturalistic, unconstrained face images with different expressions. Table 4.2 summarised the most significant approaches and their performance in well-known datasets (see Table 3.4).

For the purpose of facial modifications analysis, we adopt the framework known as Tasks-Constrained Deep Convolutional Network (TCDCN) [Zhang et al., 2016c], which is a 2D state-of-the-art facial landmark estimator system on faces with varying pose angles. Figure 4.10 depicts qualitative results in benchmark datasets. TCDCN incorporates auxiliary information into the fitting

Table 4.2 Selected benchmarking techniques for 2D facial landmark estimation and the normalised mean error in selected datasets.

Author	**Title**	Helen (68)	Helen (194)	300-W	COFW (29)
[Zhang et al., 2016c, Zhang et al., 2014c]	TCDCN	4.60	4.63	5.54	8.05
[Xiao et al., 2016]	RAR	3.99	-	4.99	6.03
[Zhu et al., 2016]	3DDFA	-	-	7.01	-
[Shao et al., 2016]	CFT	4.75	4.86	5.85	6.33
[Deng et al., 2017, Deng et al., 2016]	SDN	-	-	-	7.76
[Zhu et al., 2015]	CFSS	4.63	4.74	5.76	-
[De la Torre et al., 2015, Xiong and De la Torre, 2013]	IntraFace (SDM)	5.50	5.85	7.50	11.4
[Zhang et al., 2014a]	CFAN	5.53	-	7.69	-
[Kazemi and Sullivan, 2014, King, 2012]	ERT-DLib(V19.2)	-	4.90	6.40	-
[Tzimiropoulos and Pantic, 2014]	GN-DPM	5.69	-	-	-
Others deep approaches	Openface-CLNF [Baltrušaitis et al., 2016, Baltrušaitis et al., 2014, Baltrusaitis et al., 2013] C-DPM [Uřičář et al., 2016]				

Figure 4.10 Selected samples of qualitative results of the 2D landmark estimator in the Helen, IBUG and LPFW datasets (68 landmarks). Red rectangles indicate wrong cases. Image adapted from [Zhang et al., 2016c].

process such as head pose estimation or facial attribute inference. This architecture represents a method of transferring the representation from a pre-trained network with images annotated with sparse landmarks and attributes, to a network for dense landmark learning. The DCNN is pre-trained by five landmarks and then fine-tuned to predict the dense landmarks of 68 facial points required. The feature extraction stage contains four convolutional layers, three pooling layers and one fully connected layer. The model used filter size 5×5 conv stride 1 and filter size 2×2 max pool stride 1 [Zhang et al., 2016c], as it is illustrated in Figure 4.11. The TCDCN model was trained and tested with the MAFL, 300-W (IBUG), Helen, COFW and AFLW datasets (see Table 3.4).

In contrast to conventional multi-task learning that maximises the performance of all tasks, TCDCN aims to optimise the main task r, which is facial landmark detection, with the assistance of arbitrary number of auxiliary tasks $a \in A$, such as pose estimation and attribute inference. Thus, the problem can be formulated as,

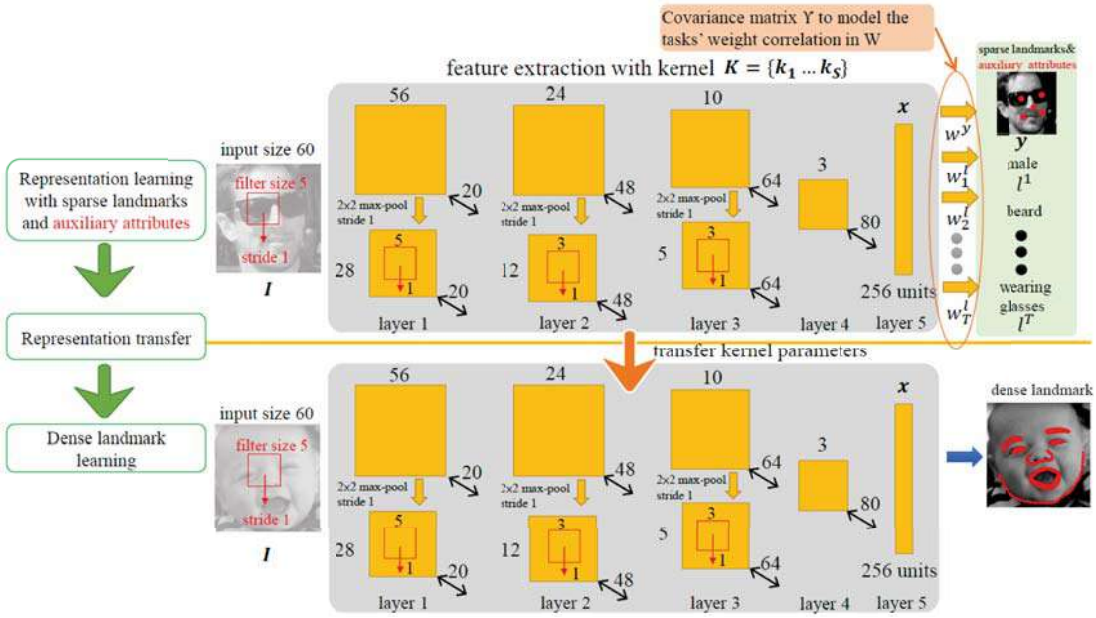


Figure 4.11 Structure specification for the TCDCN architecture. Image adapted from [Zhang et al., 2016c].

$$\left(\begin{array}{c} argmin \\ W^r, \{W^a\}_{a \in A} \end{array} \right) = \sum_{i=1}^{N} L^r(y_i^r, f(x_i; W^r)) + \sum_{i=1}^{N} \sum_{a \in A} \lambda^a L^a(y_i^a, f(x_i; W^a)), \tag{4.3}$$

where λ^a denotes the importance coefficient of $a-th$ task's error and the loss function is denoted by $L(\cdot)$. For the facial landmark detection, it is employed the least square and cross-entropy as the loss function for the main task (regression) and the auxiliary tasks (classification), respectively. So, the objective function is rewritten as,

$$\left(\begin{array}{c} argmin \\ W^r, \{W^a\} \end{array} \right) = \frac{1}{2} \sum_{i=1}^{N} || \, y_i^r - f(x_i; W^r)) \, ||^2 - \sum_{i=1}^{N} \sum_{a \in A} \lambda^a y_i^a \log(p(y_i^a \mid x_i; W^a)) + \sum_{i=1}^{T} || \, W \, ||_2^2, \tag{4.4}$$

where y_i^r is the target of landmark detection, $f(x_i; W^r) = (W^r)^T x_i$ in the first term is a linear function. The second term is a softmax function $p(y_i = m \mid x_i) = \frac{\exp\{(W_m^a)^T x_i\}}{\sum_j exp\{(W_j^a)^T x_i\}}$, that models the class posterior probability (W_j^a denotes the jth column of the matrix), and the third term penalises large weights ($W = \{W^r, \{W^a\}\}$).

3D methods for facial landmark estimation

2D facial landmark detection and alignment via cascaded regression and CNN-based methods have considerable achievements. However, these methods only regress visible points and have difficulties in dealing with large-pose changes ($\pm 90°$) and occlusions [Zhu et al., 2016]. 3D based methods contain a wide range of views using a 3D model which is robust to illumination and pose [Wang et al., 2018a]. Therefore, pioneer 3D facial landmark architectures have been

Table 4.3 Selected benchmarking techniques for 3D facial landmark estimation.

Author	**Title / Name**
[Bulat and Tzimiropoulos, 2017b]	3D Face alignment network (3D-FAN)
[Bhagavatula et al., 2017]	3D spatial transformer network (3DSTN)
[Liu et al., 2017]	Dense face alignment (DeFA)
[Zadeh et al., 2017]	Convolutional experts network (CEN)
[Jourabloo and Liu, 2017]	Pose-invariant face alignment via CNN-based dense 3D model fitting
[Liu et al., 2016]	Joint face alignment and 3d face reconstruction
[Jourabloo and Liu, 2016]	CNN regressor and 3DMM (PAWF - D3PF)
[Zhu et al., 2016]	Face alignment across large posess: a 3D solution (3DDFA)
[Jourabloo and Liu, 2015]	Pose-invariant 3D face alignment
[Jeni et al., 2015]	Dense 3D face alignment

proposed to handle face images with large pose variations [Jeni et al., 2015, Jourabloo and Liu, 2015, Jeni et al., 2016, Bulat and Tzimiropoulos, 2016b]. Table 4.3 summarises the most representative approaches to estimate facial landmarks and face alignment from single RGB images based on a 3D perspective. Some of these approaches are also used to provide a 3D face reconstruction because the estimation of landmarks are based on fitting a dense 3D shape, however, the use of a 3D face modelling strategy will be discussed in Chapter 6.

Early works to extract 3D face landmarks fit a 3D Morphable Model (3DMM - a 3D triangulated mesh built from a set of scanned faces) [Blanz and Vetter, 1999, Romdhani and Vetter, 2005] or register a 3D facial template with a 2D facial image [Zhu et al., 2016, Gou et al., 2016], *i.e.*, methods that achieve 3D face alignment by means of 3DMM fitting. 3D landmark detection is possible by selecting x, y coordinates of landmarks vertices in a reconstructed geometry. Some works have improved the points detection by using cascaded CNNs with 3DMM to proposed the face fitting [Jourabloo and Liu, 2016, Bulat and Tzimiropoulos, 2016b, Jourabloo and Liu, 2017] or by using multi-constraints to train a CNN to estimate the 3DMM parameters and provides very dense 3D alignment [Liu et al., 2017, Zadeh et al., 2017]. These frameworks seek valuable information for additional supervision and integrate them into the learning architecture. While such methods based on deep learning and 3DMM fitting achieve impressive results on challenging dataset, they are limited due to the nature of the 3D space. Some others methods directly learn the correspondence between the input image and 3D template via deep learning [Yu et al., 2017, Güler et al., 2017]; however, only visible face-region is considered and need a complex network to regress the 3D face.

Recently, an end-to-end approach [Bulat and Tzimiropoulos, 2017b] known as Face Alignment Network (FAN), has overcome this limitation of model space, by combining a state-of-the-art architecture for human landmark localisation [Newell et al., 2016] (to estimate 68 facial landmarks with 2D coordinates) with a state-of-the-art parallel and multi-scale block [Bulat and Tzimiropoulos, 2017a] to estimate depth values (see also Section 5.2.2). Figure 4.12 illustrates the network that uses four stack HourGlass (HG) networks [Newell et al., 2016], where the bottleneck block is replaced by the multi-scale block. The authors added to the common three (3) channel input (RGB) 68 additional channels, each representing a 2D landmark with a 2D Gaussian

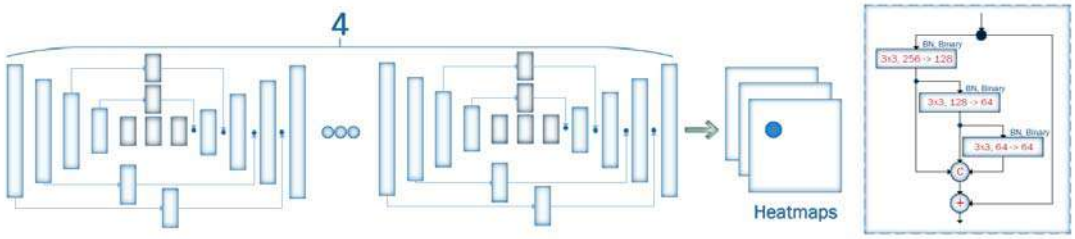


Figure 4.12 Schematic of the 3D landmark network (3D-FAN). Image adapted from [Bulat and Tzimiropoulos, 2017b].

Figure 4.13 Selected samples of qualitative results of the 3D facial landmark estimator in the 300-VW and 300-W datasets. Image adapted from [Bulat and Tzimiropoulos, 2017b].

distribution with a standard unit distribution around the landmark's location. The depth of 2D projections of the 3D landmark position is estimated using a Res-Net-152 network [He et al., 2016] adapted to accept the image and the 2D projections as input and output the depth of each landmark.

We adopt the framework proposed by Bulat and Tzimiropoulos [Bulat and Tzimiropoulos, 2017b] to estimate 3D landmarks in the epilepsy dataset. This methodology is a state-of-the-art facial 3D landmark estimation system which allows the authors to create the biggest 3D facial landmark dataset to date. The model 3D-FAN trained on the 300-W-LP-3D and the LS3D-W datasets and tested on the AFLW2000-3D dataset (see Table 3.5), outperformed benchmark architectures showing remarkable resilience to large poses, initialisation, resolution and even to the size of the network. The last one indicates that there is only a moderate performance drop related to the number of parameters used, *e.g.*, when they are reduced from 24M to 12M parameters. These parameters are varied by reducing the number of HG networks and the number of channels inside the block. Figure 4.13 illustrates qualitative results of the FAN network on public datasets.

Movement quantification

The epilepsy dataset is used to fine-tune the trained model from [Zhang et al., 2016c] and [Bulat and Tzimiropoulos, 2017b] to conduct 2D and 3D facial landmark estimation, respectively.

In order to access the locations of facial landmarks with respect to the face centre, we perform pose normalisation (frontalization) of the landmarks using the information of the head pose estimation and

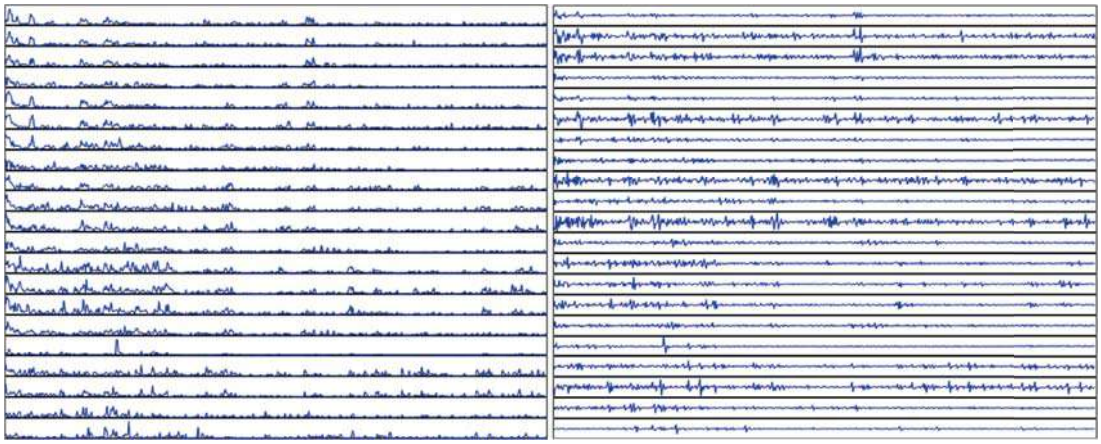


Figure 4.14 Selected samples of the trajectories of velocity (left) and acceleration (right) extracted from the landmarks detected during the full length of a seizure. The rows correspond to the location of 20 landmarks from the mouth and one nose point according to the 68-landmarks distribution.

3D shape fitting as proposed by [Huber et al., 2016] to make a fair comparison between sequences of images. Given the frontal localisation of the facial landmarks, we capture the dynamic variations of the face from consecutive frames based on spatial-temporal networks [Zhang et al., 2017a].

We generate the x-axis and y-axis movement trajectories for each detected joint and temporal features are obtained by computing 10 metrics for each landmark. The landmarks movement is analysed using a temporal window of 25 consecutive video frames to study their significance level in discriminating MTLE patients. For each landmark trajectory, the velocity and acceleration over time are calculated, and for each of these signals the standard deviation, median, mean, maximum, and minimum are measured. A representation of the trajectories of velocity and acceleration in 21 selected landmarks are illustrated in Figure 4.14. Overall, each video sequence has a feature vector with a dimensionality of $[1,680]$, which corresponds to 10 features for each of the 68 landmarks. Finally, a support vector machine (SVM) is proposed to classify facial semiology from patients with MTLE.

4.3.3 Region-based approach

The region-based method leverages the well-known cascaded network for facial expression recognition [Li and Deng, 2018], and works by extracting spatio-temporal features from video sequences to predict classes through an end-to-end deep learning model. Compared to the method based on facial key points, the performance can be enhanced by feeding raw frames to deep learning models, allowing the model to learn optimal features from the entire facial area and eliminates the need for feature engineering. For example, the selection of appropriate spatial features from the trajectories estimated in the landmark-based approach may affect the performance of the system. Facial landmark detection can also vary in their location in small proportion across consecutive frames which includes noise in the feature extraction. Additionally, the error from facial landmark estimation is not only

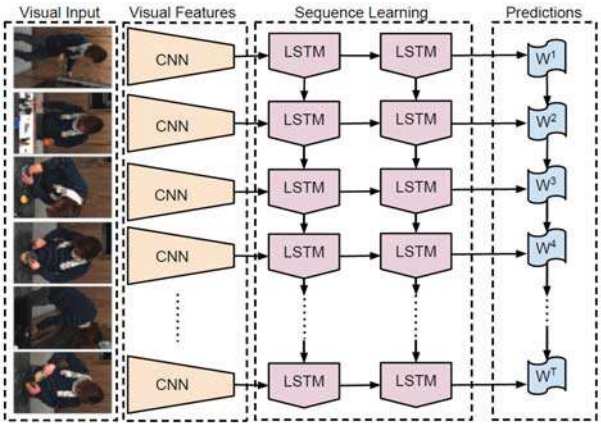


Figure 4.15 Representation of a long-term recurrent convolutional network. Image adapted from [Donahue et al., 2015].

difficult to recover in different scenarios but also propagates to the following stages affecting the motion quantification.

Cascaded networks can outperform an individual network to improve facial recognition features using dynamic-based deep architectures that exploit the crucial temporal dependency [Li and Deng, 2018]. Cascaded networks are proposed to first extract discriminative representations from static images of faces (CNNs), and then input these features to sequential networks (LSTMs). Donahue et. al [Donahue et al., 2015] proposed a novel long-term recurrent convolutional network (LRCNs) suitable for video activity recognition and video description tasks. These architectures are motivated by the development of deep convolutional architectures of 3D spatio-temporal filters (3D-CNN) [Ji et al., 2013], which also operate temporally, *i.e.*, have temporal recurrence of latent variables. Figure 4.15 illustrates the LRCNs framework. It could be argued that 3D-CNNs are superior in learning spatio-temporal features for action recognition. 3D-CNN use 3D convolutional kernels with shared weights along the time axis, and have been widely used for motion analysis such as facial expression recognition [Abbasnejad et al., 2017, Vielzeuf et al., 2017, Nguyen et al., 2017]. However, RNNs are more suitable for encoding long-term temporal information, especially from the variable-length videos [Zhang et al., 2017b], compared to short-term spatio-temporal features using a shallow 3D-CNN. Additionally, the hybrid network of CNNs and LSTMs has shown to be able to be trained with a small dataset for facial expression analysis [Kim et al., 2017, Zhang et al., 2017a]. This architecture has also been successful in medical diagnostics including facial pain recognition [Zhou et al., 2016a, Rodriguez et al., 2017, Bellantonio et al., 2016] and facial analysis of traumatic brain-injured patients [Ilyas et al., 2018]. Therefore, similar methods can be applied to the task of facial semiology in patients with epilepsy. For this reason, we adopt a CNN-LSTM framework.

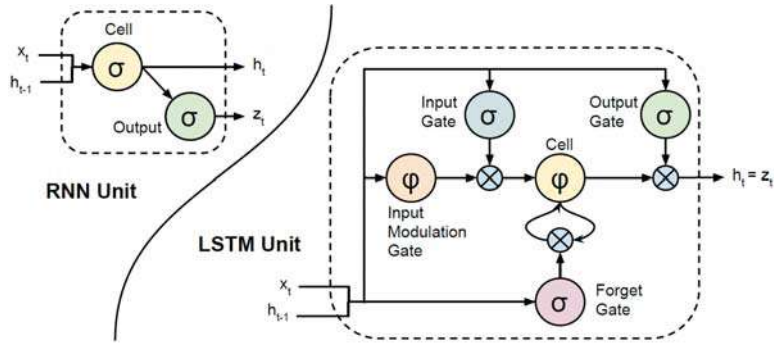


Figure 4.16 A diagram of a basic RNN cell and an LSTM memory cell. Image adapted from [Donahue et al., 2015].

As it was introduced in Chapter 2, RNNs can robustly derive information from sequences by exploiting the fact that feature vectors for successive data are connected semantically and are therefore interdependent. LSTM incorporates memory units to handle varying length sequential data with a lower computation cost, *i.e*, to allow the network to learn when to forget previous hidden states and when to update hidden states given new information. A diagram representation of an RNN and an LSTM is illustrated in Figure 4.16. For a system with input x_t, an output y_t and a hidden state h_t, a conventional RNN is constructed by defining the transition function and the output function as,

$$h_t = \phi_b(W^T h_{t-1} + U^T x_t), y_t = \phi_o(V^T h_t), \tag{4.5}$$

where W, U and V are the transition, input and output matrices respectively and ϕ_b and ϕ_o are element-wise nonlinear functions. Sigmoid or a hyperbolic tangent function are common examples of nonlinear functions. When the forget and input gates have determined how much information of the previous cell state C_{t-1} and the new cell state candidate $\hat{C}_t$ should be let through, the dynamic equations to represent the LSTM is given as,

$$\begin{aligned} \hat{C}_t &= tanh(W^T(r_t * h_{t-1}) + U^T x_t), \\ z_t &= \sigma_b(W_z^T h_{t-1} + U_z^T x_t + V_z^T C_{t-1}), \\ C_t &= f_t * C_{t-1} + i_t * \hat{C}_t, \\ h_t &= o_t * \phi_b(C_t), \end{aligned} \tag{4.6}$$

where $z = \{i, f, o, r\}$, representing the gating functions: input gate, the forget gate, the output gate and the internal gate, σ is the Sigmoid function and the trainable model parameters are $\{W, W_z, U, U_z, V_z\}$.

Deep learning architecture and training

The proposed approach uses the CNN fine-tuned in the epilepsy dataset for face detection to learn the spatial features of the face; then, these features are linked to an LSTM to exploit the temporal

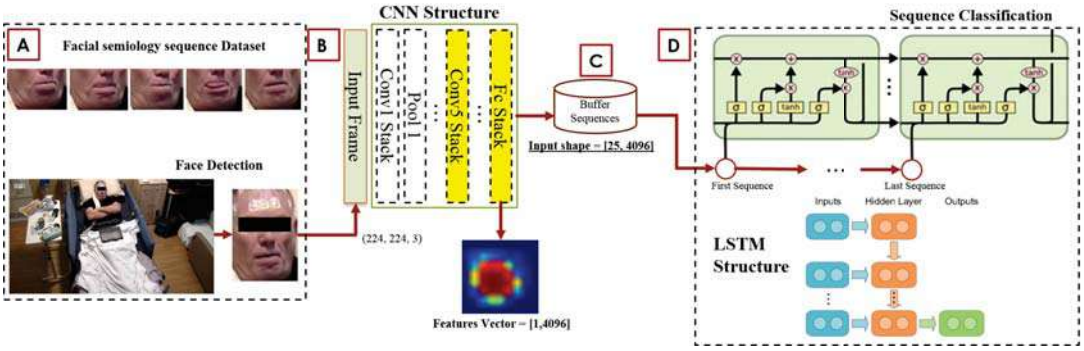


Figure 4.17 Framework proposed of the region-based methodology to quantify facial semiology. **A.** A dataset of facial semiology is created using face detection. **B.** With the CNN structure spatial features of the face image are extracted from the fully-connected layer. **C.** The temporal evolution is analysed using a temporal window of 25 consecutive video frames for each video clip. **D.** The feature sequence is fed to an LSTM to exploit the temporal relation between video frames to predict the class of each sequence.

relationship between video frames. The hybrid deep learning framework combining CNN and LSTM to exploit the spatio-temporal information of facial semiology in video sequences is displayed in Figure 4.17. A new dataset of sequences of facial semiology is created using the face detector. The image background is removed to avoid non-facial information or incorrect features by setting the input image as only the region bounded by the detected face. The sequences are analysed using a temporal window of 25 consecutive video frames. Each frame is processed through the CNN architecture to generate the visual features. The hidden layer activation is extracted from the last fully connected layer (fc7 layer) with a dimension of $[1, 4096]$, as the output of the fc7 layer in the VGG-16 network has 4,096 units.

Once the sequential features are extracted, they are fed to an LSTM network. The number of LSTM layers is one major hyper-parameter to consider in the LSTM network. We adopt a model where multiple stack LSTMs infer one output. We obtained the best performance with a network designed with 2 hidden layers of 128 and 64 hidden units, respectively. The output of the second hidden recurrent layer is fed into a densely-connected layer with a sigmoid activation function to predict the class probability for the input data sequence.

The LSTM architecture used to exploit the temporal features, is a lightweight model with approximately 800,000 trainable parameters. Training of the LSTM network is carried out by optimising the binary cross entropy loss function. The LSTM was optimised with the Adam optimiser [Kingma and Ba, 2014] with a learning factor of 10^{-3}, and decay rate of first and second moments as 0.9 and 0.999, respectively. This method for stochastic optimization has been shown to achieve competitively fast convergence rates when used for multi-layer neural networks. It was found that the SGD optimiser yielded worse performance. Dropout [Hinton et al., 2012] with a probability of 0.35 and batch size set to 4 are also used as they are considered being an effective method for reducing overfitting in deep neural networks when dealing with a huge number of parameters and a small

Table 4.4 Patients with MTLE for experiments on the detection of ictal facial expressions. Demographics of semiology during ictal activity.

Test Patient	Number of seizures	Number of frames	Main semiology
1	1	1700	Mouth and tongue movement
2	2	3750	Mouth movement and left eye blinking
3	3	1700	Fear expressions
4	2	1025	High blinking frequency
5	2	2225	Mouth and tongue movement
6	4	3750	Fear expressions and blinking
7	4	3600	Mouth movement and swallowing
8	6	6675	Mouth and tongue movement
Total	**24**	**24,425**	

training dataset. We balance the training data at the sequence level using the class weight parameters as in [Chollet, 2015]. With an imbalanced dataset, it is probable that, without class weights, a model will get biased toward the prediction of non-ictal recordings as this is the dominant case in the dataset. We performed the model training using 30 epochs and used the default initialisation parameters from Keras package [Chollet, 2015] for initialising the weights of LSTM hidden units. The LSTM is implemented in Python using Keras [Chollet, 2015] with a Theano backend [Al-Rfou et al., 2016].

4.4 Experimental results

4.4.1 Dataset specifications

A video set during seizures from 8 patients diagnosed with MTLE were considered for this preliminary analysis of facial semiology according to the data that was received until April 2017 (see Section 3.3.1 and Table 3.2). Further analysis of facial semiology with more patients and for classification purposes is discussed in Chapter 5. In this particular experiment of facial semiology identification, we included non-ictal recordings, *i.e.* video clips with non-seizure events or natural behaviour. We randomly selected 8 patients that were under monitoring, and we segmented videos such that they show facial expressions during common activities such as speaking with family members, eating and watching television. To avoid incorrect instances of natural behaviour, video recordings of interictal periods were not considered.

All seizures recorded from patients with MTLE and during ictal activity were assessed and categorised according to gestural motor behaviours including chewing, blinking, fear or wide-open eyes, eye-gaze and motions in the mouth area as they are illustrated in Figure 4.2. Table 4.4 summarises the specific symptomatology (most common ictal pattern) for these patients and the number of seizures. A total of 55 videos clips from day and night monitoring were used, representing 24 videos during seizures (Class 1) and 31 videos during natural behaviour (Class 2).

Figure 4.18 Qualitative results of face detection in the epilepsy dataset. The face detected is shown in the red bounding box.

4.4.2 Experimental setup

We adopt two approaches for cross-validation to evaluate the performance of the deep framework, a *k*-fold cross-validation and a leave-one-subject-out cross-validation. Both approaches ensure that data used for testing is completely separate from that used for training the models.

The *k*-fold cross-validation [Kohavi et al., 1995] allows us to confirm the reliability of the model by evaluating the approach for facial semiology detection of patients with MTLE on data that has not been seen during training. The sequences of all patients of the same class are randomly split into 70% for training, 20% for validation and 10% for testing *k* different folds (5-folds in this experiment). The average test accuracy of the framework is computed as the average performance of each fold.

The leave-one-subject-out cross-validation (LOSO-CV) [Xu and Huang, 2012] aims to validate the ability of the trained model to capture patient invariant features such that it can predict whether facial expressions are indicative of MTLE on patients not seen in the training set. In this scenario, we evaluate the complete video corpus for one specific patient who is not seen at any moment during the training, *i.e.*, one patient with MTLE is left out as the test patient and the remaining are used for training and validation. The prediction accuracy is computed as the average of eight models (8 of the 8 patients). This is the expected clinical scenario when analysing seizures recorded for a new patient and outputs the probability that the patient exhibit facial behaviour from MTLE.

4.4.3 Face detection

The intersection-over-union (IoU) is used to evaluate the face detection in the epilepsy dataset quantitatively. The IoU or the union of the overlapping boxes is the sum of the areas of the entire boxes minus the area of the overlap, and the intersection is divided by the union. This error returns a

value between 0 and 1, where 1 implies a perfect overlap. The fine-tuned face detector reached an average accuracy of 0.920 in the IoU in selected videos manually annotated from the epilepsy dataset (see Section 3.3.1). Figure 4.18 shows qualitative performance of the detector in different clinical scenarios of the epilepsy dataset.

4.4.4 Landmark-based analysis

The outcome of the landmark estimation is the position (x, y) for each facial keypoint and the head rotation vector represented by the yaw, pitch and roll angles in degrees. The qualitative performance under different scenarios of the facial landmark estimation using 2D and 3D approaches are depicted in Figure 4.19 and Figure 4.20, respectively.

A landmark is labelled as valid or detected if the distance between the estimated point is within a certain range (four-pixel neighbourhood) when compared to the ground truth. The 2D approach reached an average accuracy of 92% of facial point detection across all the images annotated (see Section 3.3.1). On the other hand, an average accuracy of 95.7% of detection reinforces the benefits of a 3D approach to address the difficulties in dealing with large-pose changes. These results indicate an accurate level performance with semi-frontal faces (complex head rotation across the yaw axis) and changes in illumination. However, in our data corpus consisting of scenarios for epilepsy diagnosis, more than 75% of the images with semiology were observed in cases of extreme head pose. Additionally, the landmark detection can also vary in their location in small proportion across consecutive frames which includes noise in the feature extraction. As a result, the facial landmark estimation for the 55 videos clips of our dataset was restricted by the performance of the detector. Figure 4.21 demonstrates different examples of the current challenge of the head position that affects the landmark-based analysis. Consequently, the number of clips with the available x-axis and y-axis trajectories needed to compute the metrics and extract the feature vector was very low. With a disproportionate number of features for each class, the evaluation of the classification methodology using the proposed SVM method, resulted in a very low validation accuracy of 35% (K-fold cross-validation), which infers that there are not sufficiently discriminative features to classify facial semiology from patients with MTLE. This experiment reveals that state-of-the-art landmark detection algorithms in the literature suffer when presented with these extreme pose cases and yield poor detection results in epilepsy monitoring.

4.4.5 Region-based analysis

Multi-fold cross-validation for facial semiology

The CNN-LSTM model was capable of achieving an average of 95.19% accuracy on the test set. Table 4.5 displays the comparison between the validation accuracy and test accuracy set for each fold. We compared the performance of each model from the multifold cross-validation computing the receiver operating characteristic (ROC) curve and calculating the area under the curve (AUC). The

Figure 4.19 Qualitative results of 2D facial landmark estimation in the epilepsy dataset.

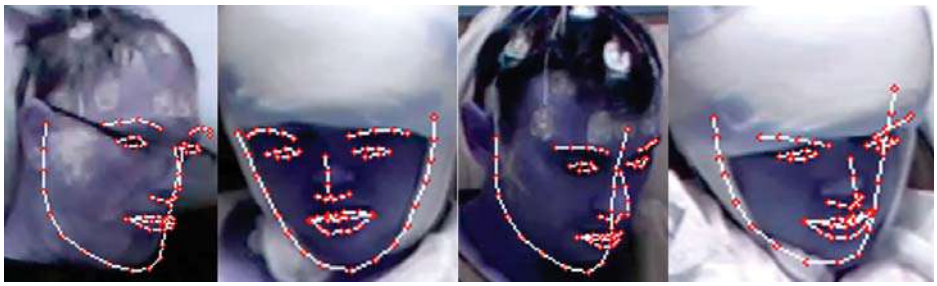


Figure 4.20 Qualitative results of 3D facial landmark estimation in the epilepsy dataset.

Figure 4.21 Selected samples that illustrate the current challenges of facial landmark estimation in the epilepsy dataset related to head pose.

Table 4.5 Multifold cross-validation performance with patients with MTLE.

Fold	Validation Accuracy (%)	Test Accuracy (%)	AUC
1	96.20	96.20	0.9984
2	93.67	94.94	0.9968
3	98.10	92.41	0.9752
4	96.84	96.20	0.9952
5	98.10	96.20	0.9976
Average	**96.58**	**95.19**	**0.9926**

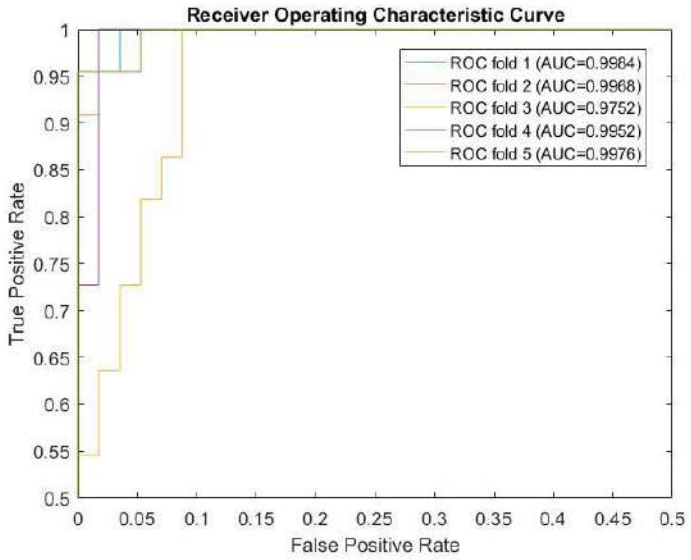


Figure 4.22 Performance of multiple ROC curves in the *K*-fold cross-validation for detection of facial expression during seizures.

area measures discrimination, *i.e.*, the ability of the test to correctly classify those with and without MTLE. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). For the five different models, the average area under the curve reached a value of 0.9926. The performance of the models is illustrated in Figure 4.22.

LOSO-CV for facial semiology

Table 4.6 illustrates the results of the performance of the validation during training and testing for each patient. The proportion of the patient in the total data indicates the number of video sequences for each patient during a seizure (class 1), where patient 8 has more sequences of seizure recorded than any other. The proposed framework reached an average validation accuracy of 97.69% and an average test accuracy of 50.85%.

Table 4.6 LOSO-CV performance of patients with MTLE during ictal activity.

Test subject	Proportion of subject in total data (%)	Validation Accuracy (%)	Test Accuracy (%)
1	6.96	97.78	87.80
2	15.35	98.10	41.11
3	6.96	97.78	9.76
4	4.19	96.97	16.67
5	9.11	98.20	81.48
6	15.35	97.63	60.00
7	14.73	97.64	50.00
8	27.32	97.42	60.85
Average		**97.69**	**50.85**

4.5 Discussion and limitations

Facial semiology has proven to be a reliable data source in epilepsy evaluation, but it is extremely difficult to characterise, and this evaluation requires standardisation among evaluators through quantitative methods. Automated analysis of facial semiology is challenging because of the immense complexity of extracting accurate features from key facial regions in the challenging conditions encountered during clinical monitoring. This study shows that quantitative facial expression analysis based on deep learning provides objective data that differentiate facial semiology from patients with MTLE during seizures from spontaneous expressions during routine monitoring. The results of the face detector, facial landmark estimator and region-based analysis illustrate the ability of cross-dataset learning to adapt models trained on other domains to new problems. This is represented by the ability to learn to evaluate neurological diseases from features learned from facial expressions. Additionally, the architectures based on LSTM proved to be a useful tool to accumulate and maintain temporal detail when processing sequence data, such as the evolution of facial expressions.

The fine-tuned face detector trained with public datasets and our dataset with epileptic patients demonstrates a successful performance to detect the patient's face under clinical monitoring scenarios. An accurate location of the face represented by the bounding box that contains the face in each frame is an important requirement to provide robust assessments with the proposed landmark-based and region-based methods. It is possible to argue that new developments for face detection may improve the quantification of facial semiology by adapting approaches such as an enhanced Faster R-CNN [Sun et al., 2018], a Face R-CNN [Wang et al., 2017a], or a face detection using R-FCN [Wang et al., 2017c] which achieved state-of-the-art results in the Wider Face dataset [Yang et al., 2016]. Most of these architectures are still based on region-based methods but with more complex deep networks, using hard example mining algorithms to boost the performance, and a different loss function such as centre loss, which leads to the improvement of training and testing efficiency. It can also be observed that by improving the face detection the spatial features extracted can enhance the discrimination between epileptic seizures. Therefore, incorporating a tracking strategy to the tracking by detection approach implemented in this experiment, may provide more consistent CNN facial features. The facial bounding box will be more consistently localised, with reduced jitter between frames (*i.e.*, less detection noise). This

enhancement will be introduced in Chapter 5. Additionally, there is evidence of deep multi-task learning frameworks that provide face detection, landmark localisation and head pose estimation simultaneously [Ranjan et al., 2019, Ranjan et al., 2017, Wu et al., 2017], which may reduce the complexity of using multiple architectures. This can make our framework more suitable for the deployment phase of a system in the hospital.

2D landmark-based methods via CNNs cascade regression and multi-task learning was introduced for the analysis of facial semiology, but the detector has difficulties in dealing with head-pose changes. A 3D landmark network has been proposed to address these difficulties achieving better results [Wang et al., 2018a]. However, we have shown that this method is also limited and is appropriate only in certain monitoring situations of the epilepsy dataset. A landmark-based approach is more intuitive because the features extracted are visually related to the amplitude and frequency of the landmark motions and will probably yield good results if the landmarks are consistently detected across challenging scenarios. Nevertheless, the performance is diminished in extreme cases of head positions which is a routine reality in the clinical monitoring environment.

The region-based method has demonstrated robust performance within the challenging, unconstrained conditions encountered in the clinical environment including changes in head pose and illumination. This method allows the extraction of a sufficient amount of features to conduct a process of categorisation of semiological behaviour. The proposed region-based approach automatically learns spatio-temporal features from raw data, which eliminates the need for feature engineering, one of the most time-consuming phases of machine learning in practice. The high performance of the multi-fold cross-validation method compared with the LOSO-CV technique (see Tables 4.5 and Table 4.6) has verified the robustness to model variations in the data but at the same time highlights the sensitivity of the system to the representativeness of data when classifying semiology. This is evident from the performance of the LOSO-CV and the differences between the average validation and test accuracies. In the particular case of Patients 3 and 4, the semiological patterns from these patients were not strongly present or similar to the semiological patterns of other patients of the dataset. This is likely because these two patients show a high blinking frequency and fear expressions which were not present in other patients and, thus, are not properly modelled by the network during the training process. However, we note that for other patients such as Patients 1 and 5, good classification results could be achieved because of their semiological patterns being exhibited by other patients in the dataset.

Although the region-based approach has been affected by the current semiology cases recorded, the results have demonstrated that the automatic feature engineering from deep models achieves promising results, and it is a novel method that should be considered for analysing patients with epilepsy as the landmark-based approach struggles in real-world conditions encountered in a hospital setting. Additionally, we have discussed that CNN-LSTM architectures are more suitable than 3D-CNN to analyse facial semiology because it is highly relevant to capture all the correlation between different types of facial expressions experienced during an epileptic seizure. However, it can be seen if more complex architectures that employed 3D-CNN for feature extraction and further choose

an LSTM to better capture the temporal dependencies for varying-length inputs may improve the analysis [Vielzeuf et al., 2017, Hasani and Mahoor, 2017].

Deep learning architectures have been useful in capturing and quantifying facial semiology. In the following chapter, we introduce the first effort on multi-modal approaches to quantify and classify seizures using clinical manifestations from the face, head, upper limbs, hands and finger movements. It remains to be seen how our framework for facial analysis and automatic feature engineering will scale when applied to a much larger dataset to evaluate facial expressions only during seizures from different epilepsy types.

Chapter 5

Multi-modal analysis of semiology in epilepsy

5.1 Overview

The analysis of semiology gives clues to the underlying cerebral networks involved. It is essential to record multiple seizures in patients with intractable seizures to establish the consistency of the semiologic features, particularly if surgery is considered. Specific patterns of facial movements, head motions, limb posturing and articulations, and hand and finger automatisms may be useful in distinguishing between mesial temporal lobe epilepsy (MTLE) and extra-temporal (ETLE) lobe epilepsy [Chauvel and McGonigal, 2014]. Inspired by the results in quantifying facial semiology introduced in Chapter 4, we are interested in exploit the benefits of deep learning to assess clinical manifestations from different body locations and classify epilepsy types based on a multi-modal perspective.

While a few single modal quantitative approaches are available to assess seizure semiology, the automated quantification of patients' behaviour across multiple modalities has seen limited advances in the literature. These systems with the purpose of classifying epilepsy types based on marker-based and marker-free approaches using 2D and 3D video sensors are limited to quantifying seizures that involve limb and head movements. Vision-based analysis of facial and hand semiology is still largely unexplored. Current methods to detect and track the movements of interest are heavily reliant on hand-crafted features that are severely affected by uncontrolled conditions. Additionally, statistical methods to evaluate differences between epilepsies have been the focus of current approaches rather than developing automated methods based on machine learning. Multi-modal systems that simultaneously evaluate different ictal phenomena are relatively rare.

Semiology encompasses the stepwise/temporal progression of signs that is reflective of the integration of connected neuronal networks. Thus, a single sign in isolation is far less informative. One single clinical manifestation cannot be used as a tool to conclude the neuronal networks affected. It is not infrequent to see conflicting lateralising or localising signs occurring in a single seizure

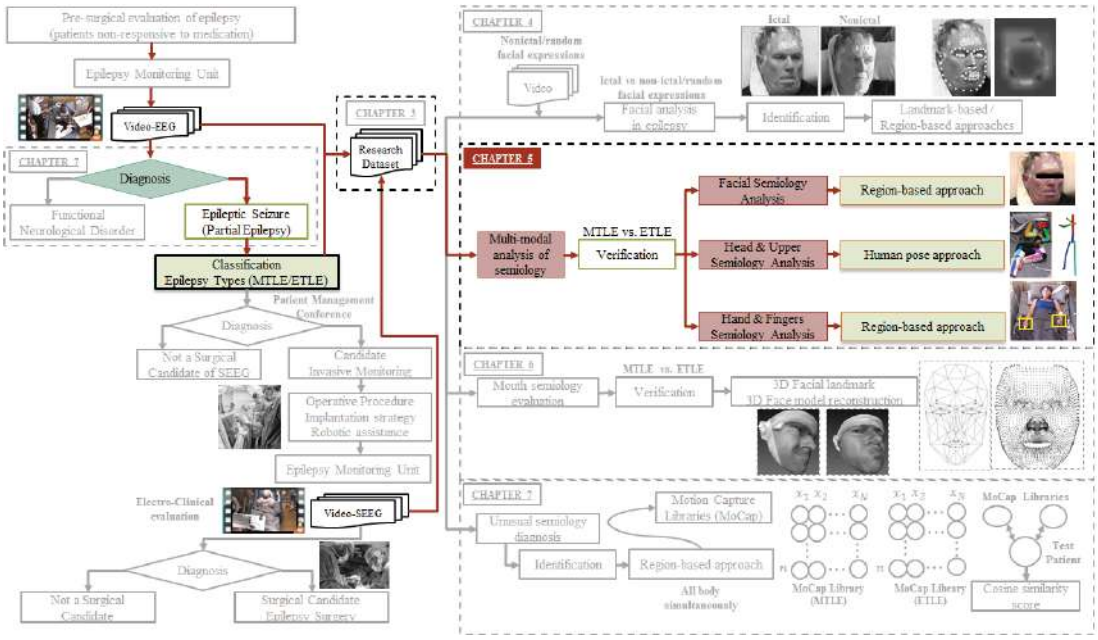


Figure 5.1 Representation of the chapters and their interconnections in the clinical evaluation of epileptic patients. Overview of the research aim in Chapter 5.

while clinical manifestations are relatively specific for MTLE and ETLE [Noachtar and Peters, 2009, Tufenkjian and Lüders, 2012]. ETLE seizures may have some similarities to temporal lobe seizures both semiologically and electrographically [Arain, 2017]. Taking this into account, in this chapter we introduce novel approaches that aim to detect and quantify semiology in order to discriminate between patients with MTLE from patients with ETLE. We proposed two different methodologies to perform multi-modal analysis: a fusion and an hierarchical approach. Our propose framework can jointly learn semiologic features from the face, body, and hand movements based on computer vision and deep learning architectures without requiring specialist (*i.e.*, depth or infrared) camera hardware. By automatically extracting spatio-temporal information and conducting different strategies of multi-modal analysis, we may enhance the understanding of semiology compared to using single modal information. The multi-modal approaches proposed in this work can assist in detecting features to differentiate between the two types of epilepsy, providing objective evidence and supporting clinicians in the pre-surgical assessment of patients with epilepsy. This chapter's research aim and its relationship with the thesis is illustrated in Figure 5.1.

The chapter is distributed as follows. Section 5.2 describes the multi-modal fusion strategy to classify epilepsy types; Section 5.3 introduces the multi-modal hierarchical approach to verify epilepsies; Finally, Section 5.4 includes the discussion of the multi-modal analysis, limitations and possible future directions.

This chapter is supported by the following published and accepted manuscripts:

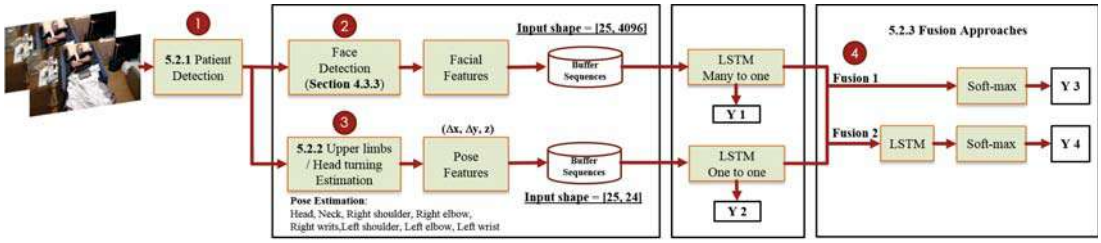


Figure 5.2 Multi-modal analysis of semiology: fusion approach of the face, head and upper limbs semiology. **1.** Detection of the patient based on an improved region-based CNN architecture. **2.** Spatio-temporal information of the facial semiology is extracted using the region-based approach proposed in Section 4.3.3. **3.** 2D pose localisation, tracking and 3D inference are also estimated to quantify the head and upper limbs movements. Auxiliary intermediate outputs ($Y1$ and $Y2$) are used to aid learning. **4.** Two approaches are considered to perform an early fusion and the final output is decided by considering the output of a soft-max layer ($Y3$ and $Y4$).

- **D. Ahmedt-Aristizabal**, C. Fookes, K. Nguyen, S. Denman, T. Fernando, S. Sridharan, S. Dionisio, A Hierarchical Multi-modal System for Motion Analysis in Epileptic Patients, *Epilepsy & Behaviour*, 87 (2018) 46-58.
- **D. Ahmedt-Aristizabal**, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, C. Fookes, Deep Motion Analysis for Epileptic Seizure Classification, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2018, 3578-3581
- A. Pemasiri, **D. Ahmedt-Aristizabal**, K. Nguyen, S. Sridharan, S. Dionisio, C. Fookes, Semantic Segmentation of Hands in Multimodal Images: A New Region-based CNN Approach, *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019.

5.2 Multi-modal strategies to classify epilepsies: fusion approach

The fusion of several modalities is one of the most critical tasks in multi-modal analyses. The fusion aims to extract useful information from a set of different input modalities and merge them in such a way that allows making more accurate and robust decisions. Data from multiple sources are semantically correlated, and sometimes it provides complementary information. A block diagram of the proposed system is displayed in Figure 5.2. For each frame, spatial information is extracted from the face, head and upper limbs movements based on convolutional neural networks and changes in the human body position. Then, temporal information for each sequence, defined as 25 consecutive frames for each modality, is extracted using a recurrent neural deep learning model via an LSTM architecture. Lastly, the two models of face and body pose are merged using two different approaches to evaluate the benefits of fusion techniques to classify epileptic seizures.

This section is distributed as follows. Section 5.2.1 provides an improved approach for patient detection; Section 5.2.2 introduces a new framework to quantify semiology from the head and upper limbs motions; To quantify facial modifications, we use the region-based approach of Section 4.3.3.

Figure 5.3 Qualitative performance of the Mask R-CNN architecture in the COCO dataset using ResNet-101-FPN. Image adapted from [He et al., 2017].

Section 5.2.3 explains the fusion approaches for this particular experiment; Finally, Section 5.2.4 summarises the dataset used and the results in classifying mesial temporal from extra-temporal lobe seizures.

5.2.1 Patient detection

As discussed in Section 4.3.1, objects and other people inside a patients' room may affect the performance of the analysis of patients with epilepsy. For this reason, to extract features related to the patients' behaviour, we first define the region of interest that contains the patient. Once all humans inside the patient room are detected in the first frame of the video clip, we categorise the patient as the person located in the bed. This helps to ensure that the extracted features of each video are consistent and independently extracted from the same region. This procedure also helps overcome changes in camera-bed viewing angles, in the inclination angle of the bed, and camera resolution, which may impact the quantification analysis. We use the bounding box coordinates of the detected patient and bed to define the region of interest locations in the x-axis of the frame and retain the original height of the video.

The strategy proposed in Chapter 4 based on the region-based Faster R-CNN architecture is improved by implementing the Mask R-CNN approach [He et al., 2017]. This proposal is a benchmark object detection method, trained on the COCO dataset [Lin et al., 2014], which includes the categories of human and bed object detection. Figure 5.3 depicts qualitative results in a benchmark dataset. Mask R-CNN is an extension of the Faster R-CNN architecture, which adds a branch for predicting segmentation masks for each region of interest in parallel with the existing branch for classification and bounding box regression, *i.e.*, includes a pixel-to-pixel alignment, which is the main missing piece of Faster R-CNN. The instance-level recognition Mask R-CNN framework is illustrated in Figure 5.4. Both stages are connected to the backbone structure. The backbone of the Mask R-CNN is based on the Feature Pyramid Network (FPN) [Lin et al., 2017] and ResNet101 (101 layers) [He et al., 2016], different from the VGG-16 model [Simonyan and Zisserman, 2014] used in the Faster R-CNN. FPN uses a top-down architecture with lateral connections to build an in-network feature pyramid from a single-scale input, *i.e.*, FPN outperforms other single CNNs mainly for the reason that it maintains strong semantically features at various resolution scales. Using a ResNet101-FPN backbone for feature extraction with Mask R-CNN gives excellent gains in both

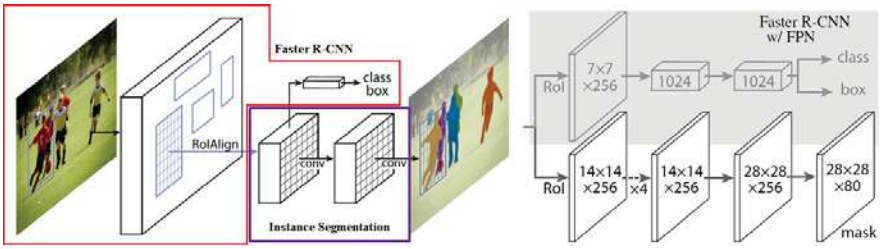


Figure 5.4 The Mask R-CNN framework for instance segmentation. Image adapted from [He et al., 2017].

Figure 5.5 Selected samples of the patient detection in the epilepsy dataset. **A.** Segmentation of the human and detected bed, and the bounding box that defines the region of interest. **B.** Selected examples of cropped images under different clinical scenarios including day and night monitoring.

accuracy and speed. Mask R-CNN improves the RoI pooling layer, named "RoIAlign layer", so that RoI can be better and more precisely mapped to the regions of the original image.

The multi-task loss function of Mask R-CNN combines the loss of classification, localisation and segmentation mask: $L = L_{cls} + L_{box} + L_{mask}$, where L_{cls} and L_{box} are the same as in Faster R-CNN. The mask branch generates a mask of dimension $m \times m$ for each RoI and each class; K classes in total. Thus, the total output is of size $K \cdot m^2$. Because the model is trying to learn a mask for each class, there is no competition among classes for generating masks. L_{mask} is defined as the average binary cross-entropy loss, only including $k-th$ mask if the region is associated with the ground truth class k,

$$L_{mask} = \frac{1}{m^2} \sum_{1 \leq i,j \leq m} y_{ij} \log \hat{y}_{ij}^k + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^k), \tag{5.1}$$

where y_{ij} is the label of a cell (i, j) in the true mask for the region of size $m \times m$; and $\hat{y}_{ij}^k$ is the predicted value of the same cell in the mask learned for the ground-truth class k.

The implementation is in Python 3, Keras, and Tensorflow [Abdulla, 2017]. The detected patient bounding box is expanded in width with an offset of 20% of the total width on each side to avoid any extremities of the patient being located outside of this boundary due to movements during a seizure.

We crop and resize all images such that each patient is extracted with a resolution of 550×720 pixels. Figure 5.5 illustrates selected examples under different scenarios. We argue that the patient detection phase does not impact generalisation. The aim of this preprocessing is to simply reduce the variability in the input space, which may, in fact, make it easier to apply our methodology elsewhere.

5.2.2 Head and upper limbs semiology

Quantifying a person's posture and limb articulation is convenient to understand semiology. The quantitative analysis of parameters such as the frequency and amplitude of the motor trajectories of the head and upper limbs reflects the clinical characteristics of a specific type of epilepsy [Noachtar and Peters, 2009, Chauvel and McGonigal, 2014]. Marker-free approaches based on 3D motion capture sensors such as depth and infrared cameras have demonstrated more accurate results compared with 2D approaches that rely on RGB cameras [Cunha et al., 2016a, Cunha et al., 2016b], but the detection of body joints is still difficult in a clinical monitoring environment, and systems still depend on semi-automatic tracking algorithms and hand-crafted features as discussed in Chapter 2. Poor predictions using depth cameras due to the lack of sufficient image cues about 3D body pose and changes in frame-rate, focus, and resolution are other limitations identified in the literature [do Carmo Vilas-Boas and Cunha, 2016]. Therefore, the need to overcome problems associated with marker-free systems is vital to make such approaches more applicable to the epilepsy monitoring environment. Like other computer vision domains, deep learning has demonstrated improved performance for estimating articulated pose and motion estimation from 2D videos [Cao et al., 2017, Sarafianos et al., 2016] such as those stored in the epilepsy monitoring units of video technology-equipped hospitals.

In order to conduct an evaluation of ictal phenomena such as arm flexion, dystonic limb posturing, tonic limb posturing, unilateral immobile limb, fencing posturing, shuddering or ictal head turning, we aim to detect and quantify head and upper limbs motions by implementing landmark-based techniques, known as human pose estimation, that have been successful in understanding human behaviour. Accurate pose estimation systems can help in analysing these actions. The main outcome of the movement detection is the location of the joints selected that will describe spatio-temporal changes of clinical manifestations.

Figure 5.6 illustrates the framework proposed to classify epileptic seizures considering 2D pose localisation, spatial-temporal inference to improve key point detection over frames (pose estimation in videos) and mapping the set of 2D detections into 3D space. The temporal information from the changes in position for each dimension $[\Delta x, \Delta y, z]$ is fed as a sequence into the LSTM models to exploit the temporal relation. The experimental design has been conducted to assess the suitability of benchmark motion methodologies to perform the kinematic analysis under large variations in illumination, pose and presence of occlusions in the epilepsy dataset. Specifications of the techniques implemented for each phase in the analysis of semiology from the head and upper limbs motion will be described in the following subsections according to the content structure displayed in Figure 5.7.

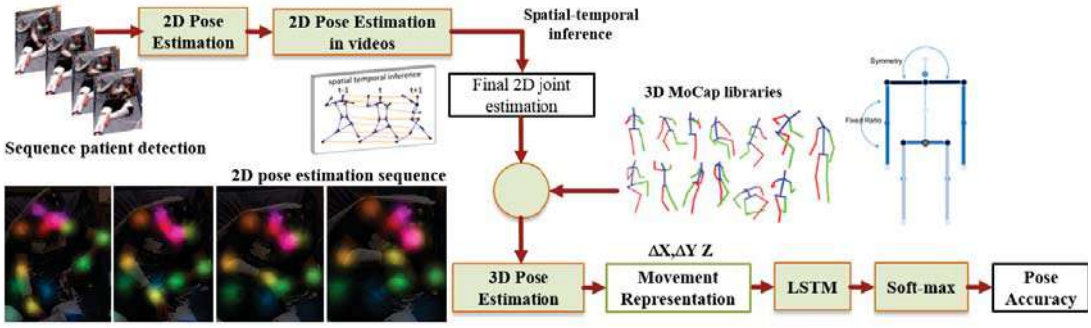


Figure 5.6 Framework proposed to quantify body semiology and classify epilepsy types. Given a number of adjacent frames, first we perform a joint detection to estimate 2D human pose, then a flow warping layer and a spatio-temporal inference layer are used to infer between body parts spatially and temporally, producing the final joint position estimated for each frame (pose estimation in videos). Depth information is estimated by matching to a library of 3D poses. Each sequence of features of the movements corresponding to three dimensions is fed to an LSTM to exploit the temporal relation and discriminate epilepsy types.

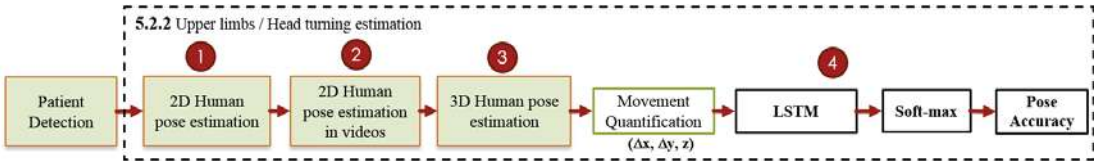


Figure 5.7 Phases to quantify semiology from head and upper limbs movements: Two-dimensional human pose estimation, tracking or human pose estimation in videos, three-dimensional inference, movement quantification and temporal analysis.

2D human pose estimation

Deep learning techniques have modelled long-range dependencies between variables in structured prediction tasks such as articulated pose estimation. Articulated body pose estimation recovers the pose of an articulated body, which consists of joints and rigid parts using image observations. Recent works that proposed deep networks to solve the challenge of movement quantification and articulated pose estimation have shown remarkably robust performance and high part localisation accuracy. These techniques outperform classical approaches, mostly replacing hand-crafted features from sequential prediction framework, pictorial structures, hierarchical and graphical models [Poppe, 2007]. For this reason, the feature extraction of the patient semiology from body posture and movement is based on these deep architectures.

Proposed methods for articulated human pose estimation using convolutional architectures can be classified as detection-based or regression-based. Detection-based methods rely on CNN-based part detectors which are then combined using a graphical model or refined using regression. Regression-based methods, on the other hand, try to learn a mapping from an image and CNN features to part locations. We have compared and assessed state-of-the-art architectures based on qualitative and

Table 5.1 Selected benchmarking techniques for 2D human pose estimation and their performance in the MPII dataset.

Author	Title	PCKh
[Cao et al., 2017]	Part Affinity Fields (PAF) and CPM	92.0
[Newell et al., 2016]	Stacked Hourglass networks for human pose estimation	90.9
[Bulat and Tzimiropoulos, 2016a]	Human Pose Estimation via Convolutional Part Heatmap Regression	89.7
[Wei et al., 2016]	Convolution Pose Machines (CPM)	88.5
[Insafutdinov et al., 2016]	DeeperCut: A deeper, stronger, and faster pose estimation model	88.5
[Belagiannis and Zisserman, 2017]	Recurrent human pose estimation	83.9
[Pishchulin et al., 2016]	Deepcut: Joint subset partition and labeling for multi person pose estimation	82.4
[Tompson et al., 2015]	Efficient object localization using convolutional networks	82.0
[Carreira et al., 2016]	Human pose estimation with iterative error feedback	81.3

Figure 5.8 Selected samples of qualitative results of the pose estimation technique CPM in public datasets. Image adapted from [Wei et al., 2016].

quantitative performance in benchmark datasets (see Table 3.7), and the epilepsy dataset. The most representative architectures in the literature are listed in Table 5.1. The evaluation of each related work is compared using metrics such as the percentage of correct keypoints (PCK) [Andriluka et al., 2014], the area under the curve (AUC), deep learning framework, computational cost, learning parameters and documentation to ensure full reproducibility for fine-tuning and testing.

The frameworks selected as the automatic technique for keypoint localisation of a single person's pose during a seizure are the part affinity fields (PAF) [Cao et al., 2017] and the convolution pose machines (CPM) [Wei et al., 2016], which are available in the open source Openpose [Hidalgo et al., 2018].

Convolution pose machines (CPM) is a sequence of convolutional networks that generate 2D belief maps for the location of each part, and the message is learned end-to-end using back-propagation. At each stage, image features and the belief maps produced by the previous stage are used as input. Figure 5.8 illustrates qualitative results of the CPM model in benchmarking datasets. The convolutional networks learn implicit image-dependent spatial models of the relationships between parts by enforcing intermediate supervision periodically through the network. This methodology combines the advantages of deep convolutional architectures with the implicit spatial modelling

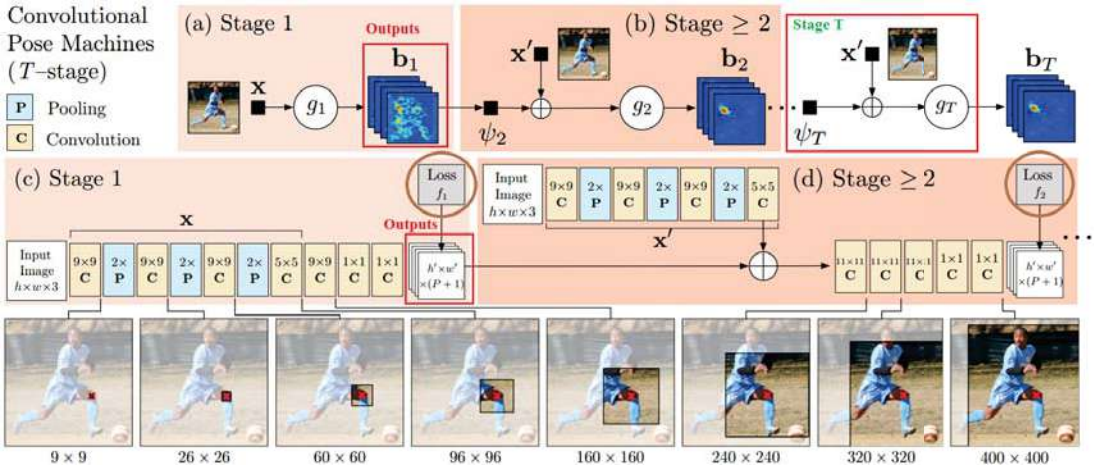


Figure 5.9 Structure specification for the convolution pose machines architecture. Image adapted from [Wei et al., 2016].

afforded by the pose machine framework [Ramakrishna et al., 2014]. The model is trained in the MPII, LSP and FLIC datasets (see Table 3.7).

The CPM architecture is shown in Figure 5.9. The first stage predicts part beliefs from only local image evidence. The network structure is composed of five convolutional layers followed by two 1×1 convolutional layers which result in a fully convolutional architecture. To achieve a certain precision, the algorithm normalises the input images to size 368×368 and the receptive field of the network shown is 160×160 pixels. The network is supervised locally after each stage using an intermediate loss layer that prevents vanishing gradients during training. A predictor in subsequent stages can use the spatial context of the noisy belief maps in a region around the image location and improve its predictions. The feature function serves to encode the landscape of the belief maps from the previous stage in a spatial region around the location of the different parts. The convolutional layers in the subsequent stage allow the classifier to combine contextual information by picking the most predictive features freely.

The loss function at the output of each stage t that minimises the distance between the predicted and ideal belief maps for each part is defined as,

$$f_t = \sum_{p=1}^{P+1} \sum_{z \in Z} \| b_t^p(z) - b_*^p(z) \|_2^2, \tag{5.2}$$

where the ideal belief map for a part p is written as $b_*^p(Y_p = z)$, which are created by putting Gaussian peaks at ground truth locations of each body part p.

Part affinity fields (PAF) is a set of 2D vector fields that encode the location and orientation of human key points over the image domain. This approach enhances the performance of CPM by learning part locations and their associations, using a novel framework trained in the large-scale COCO dataset [Lin et al., 2014] with over 100K person instances labelled with over 1 million total

Figure 5.10 Selected samples of qualitative results of the pose estimation technique PAF+CPM in a public dataset. Image adapted from [Cao et al., 2017].

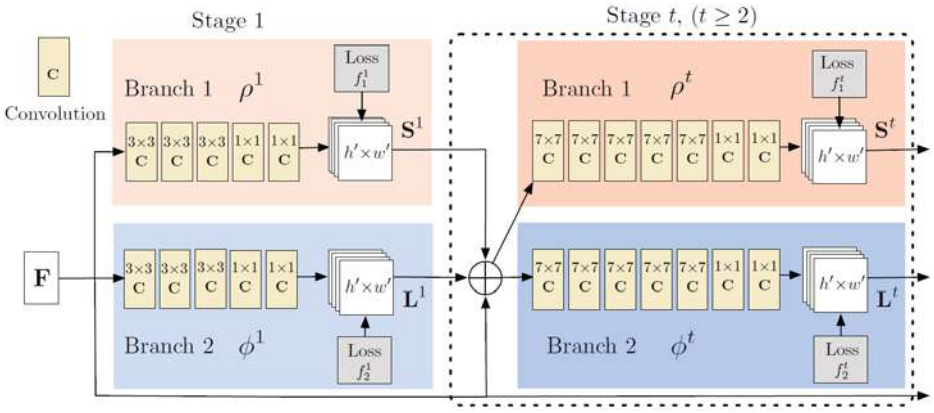


Figure 5.11 Structure specification for the part affinity fields architecture. Image adapted from [Cao et al., 2017].

keypoints (see Table 3.7). Figure 5.10 displays qualitative results of the PAF model in the COCO dataset for multi-person pose estimation.

The architecture depicted in Figure 5.11 predicts detection confidence maps and affinity fields that encode part-to-part association simultaneously. The network is divided into two branches: The bottom branch predicts the affinity fields, and the top branch predicts the confidence maps. Each branch is an iterative prediction architecture based on CPM [Wei et al., 2016], which refines the predictions over successive stages, with intermediate supervision at each stage. In the first stage, the network predicts the confidence maps and a set of part affinity fields. After the first stage, the predictions from the two branches are concatenated and used to produce refined predictions. Finally, the confidence maps and the affinity fields are parsed by greedy inference to output the 2D keypoints for the human in the image.

Table 5.2 Selected benchmarking techniques for 2D human pose estimation in videos with the PCK@0.2 metric.

Author	Title	Penn Action	JHMDB
[Song et al., 2017]	Thin-Slicing Network	96.5	92.1
[Gkioxari et al., 2016]	Chain Model	91.8	-
[Iqbal et al., 2017a]	Pose for action-action for pose	81.1	73.8

2D human pose estimation in videos

State-of-the-art approaches for key point estimation based on CNNs have shown strong performance in static images [Wei et al., 2016, Cao et al., 2017]; however, their performance is still limited in videos. Video-based tracking by detection schemes using 2D point detectors in videos does not fully exploit temporal information that can help to correct mistakes caused by joints occluded and large amounts of motion that generate motion blur. In order to alleviate this problem, a point tracking approach was considered based on the distance between key points in successive frames in [Cunha et al., 2016a]; however, this is still a semi-automatic tracking algorithm that heavily relies on a hand-crafted process. We aim to enhance the 2D pose prediction for semiology analysis by extracting consistent poses in sequences based on pose tracking approaches which integrate time information at the instance level. State-of-the-art methodologies are summarised in Table 5.2.

The baseline system for articulated pose tracking in unconstrained videos is a two-stage tracking-by-detection approach. The first stage combines a person detector with a frame-level pose estimation method. In the second stage, the single-frame pose estimations are linked temporally, where the assignment is performed at the level of body poses or individual body parts, *i.e.*, tracking based on associations between neighbouring frames [Andriluka et al., 2017]. Although these approaches have shown robust performance in a keypoint tracking challenge [Girdhar et al., 2018], they are not end-to-end networks able to directly infer articulated person tracks from videos, and are severely limited when confronted with fast camera motions and complex articulations during seizures.

Based on proposals that incorporate spatial and temporal information into a unified deep learning architecture, we adopt the framework based on the Thin-Slicing Network [Song et al., 2017]. This approach incorporates into the network an adjustment from the optical flow and a spatio-temporal message passing layer to smooth the joint predictions over space and time. Similar to [Pfister et al., 2015], this approach directly propagates joint position from previous to the current frame via optical flow. Song et al. [Song et al., 2017] show an improvement over frame-level prediction compared with tracking by detection [Wei et al., 2016] in two large-scale video pose estimation benchmarks: Penn Action [Zhang et al., 2013] and JHMDB [Jhuang et al., 2013] datasets (see Table 3.8), without the computational complexity of integer programming optimisation as used by state-of-the-art models [Iqbal et al., 2017b, Insafutdinov et al., 2017]. Additionally, although in [Insafutdinov et al., 2017] the model uses spatial-temporal models on top of a joint regressor for pose

estimation in videos, the optimisation of the graphic model is independent of the future learning process.

The framework proposed by [Song et al., 2017] is illustrated in Figure 5.12. The heatmaps produced by the pose estimator is passed through the flow warping layer to align heatmaps from one frame to the target neighbour. The warped heat-maps serve as input to the spatio-temporal inference layer. The novelty of this approach is the spatio-temporal inference layer that incorporates spatio-temporal dependencies. Regarding modelling, the single image pose estimation problem can be formulated as maximising the score $S(I,p)$ for a pose p given an image I,

$$S(I,p) = \sum_{i \in V} \phi(p_i|I) + \sum_{(i,j) \in E_s} \psi_{i,j}(p_i,p_j), \tag{5.3}$$

where $\phi(p_i|I)$ is the unary term for the body part i at the position p_i in the image I and $\psi_{i,j}(p_i,p_j)$ is the pairwise term modelling the spatial compatibility of two neighbouring parts i and j. Given a video sequence $\Gamma = (I_1, I_2, ..., I_T)$, the temporal links are introduced among neighbouring frames in order to impose temporal consistency for estimated poses $\Upsilon = (p^1, p^2, ..., p^T)$. The objective function is then the summation of the scores of single frames and the pairwise term of the temporal link. This objective score function is given by,

$$S(\Gamma,\Upsilon)_{slice} = \sum_{t=1}^{T} S(I^t,p^t) + \sum_{(i,j) \in E_f} \psi_{i,i*}(p_i,p_{i*}^t), \tag{5.4}$$

here, the pairwise term $\psi_{i,i*}(p_i,p_{i*}^t)$ regularises the temporal consistency of the part i in neighbouring frames. The elements of the temporal link are the current prediction and the flow of the prediction from neighbouring frame. Then, inference corresponds to maximising S_{slice} over p for the images sequence slice. At each iteration, a part i sends a message to its neighbours and also receives a reciprocal message along the edges given by,

$$score_i(p_i) \longleftarrow \phi(p_i|I) + \sum_{k \in child(i)} m_{ki}(p_i). \tag{5.5}$$

Therefore, for each iteration the scope part i is the summation of its unary terms and the message collected across different spatio-temporal neighbours. The message $m_{ki}(p_i)$ sent from body part k to part i are given by,

$$m_{ki}(p_i) \longleftarrow max_{p_k}(score_k(p_k) + \psi_{k,i}(p_k,p_i)). \tag{5.6}$$

This cost maximisation process is efficiently solved via the generalised distance transforms [Felzenszwalb and Huttenlocher, 2012].

In our implementation, the joint confidence maps (heatmaps) of eight key joints position (Nose, Neck, Right-Left Shoulder, Right-Left Elbow, and Right-Left Wrist) are detected using PAF and CPM [Wei et al., 2016, Cao et al., 2017]. Then, these heat-maps are passed through a flow warping layer to align heat-maps from one frame to the targeted neighbour. Finally, the part heat-maps and the

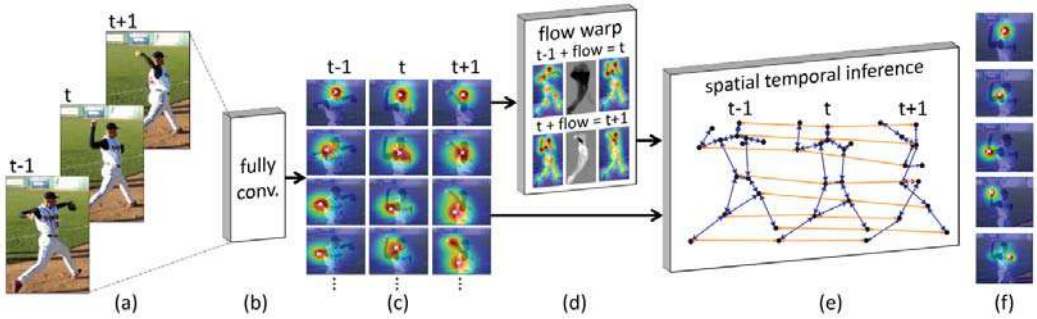


Figure 5.12 Schematic of the human pose estimation in videos. Given a sequence (a), the fully convolutional layers (b) regress initial confidence maps for each frame (c). The flow propagates information temporary. A flow-based warping layer aligns joint heat-maps to the current frame (d). A spatio-temporal inference layer performs iterative message passing along both spatial and temporal edges of the loopy pose configuration graph (e) and produces final joint prediction (f). Image adapted from [Song et al., 2017].

.

warped heat-maps are fed to a spatio-temporal inference layer to produce the final 2D joint positions estimated for all frames in the video [Song et al., 2017]. The model generates the final 2D body-part locations represented by the x and y values for each frame. The epilepsy dataset is used for fine-tuning the pose estimator [Cao et al., 2017], to improve the performance on real clinical scenarios during day and night monitoring.

3D human pose estimation

Inspired by state-of-the-art models for 3D human pose estimation from RGB images [Sarafianos et al., 2016], we employ the nearest neighbour matching of a given 2D prediction using a database of 2D to 3D correspondences to increase the number of spatial features of the patient's body movement in each frame without the use of hand-crafted features [Cunha et al., 2016a]. An overview of the problem of reconstructing 3D human motion from monocular image sequences with traditional approaches is provided in [Holte et al., 2012, Sminchisescu, 2008].

Similar to 2D human pose estimation and in order to enhance the 3D pose prediction task, new proposals have used the undeniable impact of deep learning. Current methods for 3D estimation can be classified as intermediate 2D pose (from 2D to 3D joints) and deep regression (deep-net-based). Intermediate 2D pose approaches are considered as a constrained optimisation problem, which aims to minimises the 2D reprojection error of an unknown 3D pose and camera using large databases representing human pose as sparse combinations [Chen and Ramanan, 2017, Zhou et al., 2016d, Akhter and Black, 2015, Ramakrishna et al., 2012]. Deep regression methods, on the other hand, aims to formulate the problem as a direct 2D image to a 3D pose regression task, *i.e.*, learn the mapping between 2D and 3D with deep neural networks [Zhou et al., 2017, Pavlakos et al., 2017, Moreno-Noguer, 2017, Tekin et al., 2016, Li and Chan, 2014]. Deep learning is used to train a

Table 5.3 Selected benchmarking techniques for 3D human pose estimation and their performance in the Human3.6M dataset using the mean Euclidean distance (mm) (MPJPE).

Author	Title	Protocol-I	Protocol-II
[Zhou et al., 2017]	3D pose a weakly-supervised approach	-	64.90
[Pavlakos et al., 2017]	Coarse-to-fine volumetric prediction	-	66.92
[Mehta et al., 2017]	Monocular 3D pose using CNN supervision	-	74.14
[Tome et al., 2017]	Lifting from the deep	70.7	88.4
[Moreno-Noguer, 2017]	3D pose via distance matrix regression	76.5	87.3
[Chen and Ramanan, 2017]	3D pose = 2D pose estimation+ matching	82.7	114.2
[Rogez and Schmid, 2016]	Mocap-guided data augmentation	88.1	-
[Sanzari et al., 2016]	Bayesian image based 3D pose	93.2	-
[Yasin et al., 2016]	A dual source approach for 3D pose	108.3	-
[Zhou et al., 2016d]	Sparseness meets deepness	-	113
[Tekin et al., 2016]	3D from motion compensated sequences	-	125.0
[Ionescu et al., 2014]	Human3.6m - predictive methods	-	162.1
[Ramakrishna et al., 2012]	3D Human Pose from 2D Image Landmarks	-	157.3

regression model to predict 3D directly from images, where the central challenge is to generalise novel poses outside the training set [Mehta et al., 2017, Rogez and Schmid, 2016, Yasin et al., 2016].

The most relevant 3D human pose prediction approaches are summarised in Table 5.3. We focus the selection of the method based on the performance in the Human3.6M dataset [Ionescu et al., 2014] using the Protocol-I and Protocol-II (see Table 3.9), qualitative results in the epilepsy dataset without performing fine-tuning, documentation of the framework, availability of training models and deep learning libraries. In Protocol-I six training subjects (S1, S5, S6, S7, S8, S9) and one testing subject (S11) are used. In Protocol-II five subjects (S1, S5, S6, S7, S8) are used for training and two subjects (S9, S11) for testing. The evaluation metric is the mean per joint position error (MPJPE) in *mm* after aligning the depths of the root joints.

The reconstruction of 3D points from a single monocular RGB image has two main challenges that affect its performance: (i) it is an ill-posed problem because similar image projections can be derived from different 3D poses and (ii) it is an ill-conditioned problem since minor errors in the locations of the 2D body joints can have large consequences in the 3D space. We adopt the approaches proposed in [Chen and Ramanan, 2017] and [Zhou et al., 2017], to estimate depth values of the 2D human pose location detected. Figure 5.13 illustrates the frameworks for the two approaches selected and Figure 5.14 depicts their qualitative performance in a public dataset.

The problem of inferring 3D joints from their 2D projections are traced back to the classic works in [Lee et al., 1985]. Following the idea of exploiting nearest neighbours for refining the result of pose inference, the proposal in [Chen and Ramanan, 2017] employs nearest neighbour matching of a given 2D prediction using a database of 2D to 3D correspondences, *i.e.*, predict depth values for the estimated 2D joints. The approach uses a data-driving matching with a modest number of exemplars (200,000), that when combined with a simple closed-form warping algorithm, yields a fast and accurate 3D solution outperforming more complex methods. The approach in [Chen and Ramanan, 2017],

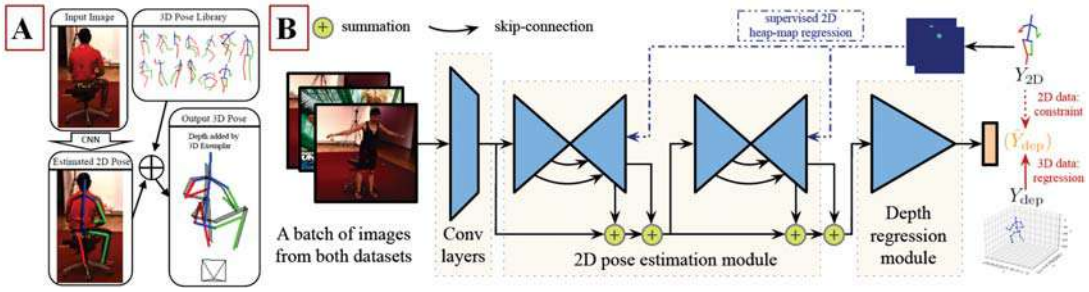


Figure 5.13 Schematic of the 3D human pose estimation. **A.** The approach proposed in [Chen and Ramanan, 2017]. The depth is estimated by matching to a library of 3D poses. **B.** The approach proposed in [Zhou et al., 2017]. The model includes a depth regression module which predicts the depth values, with the usage of a 3D geometric constraint induced loss.

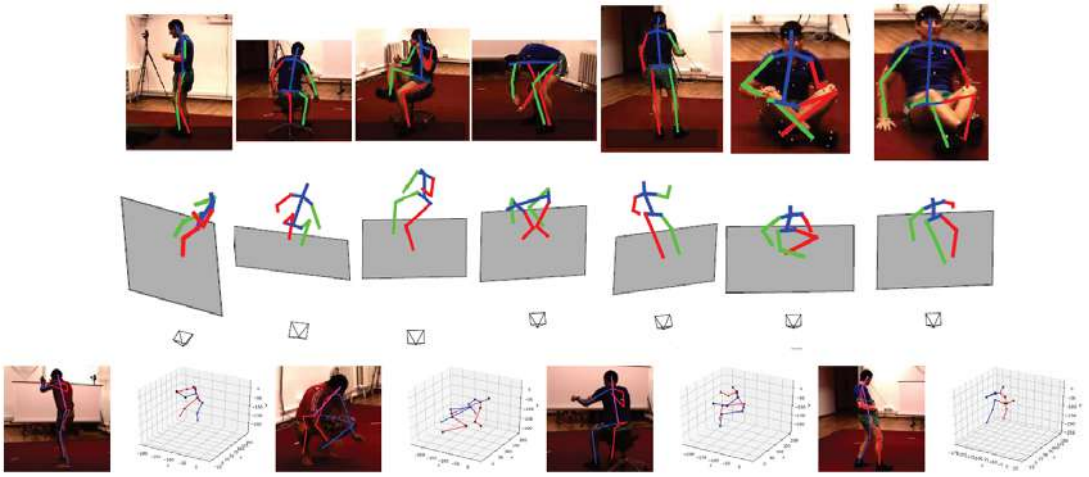


Figure 5.14 Selected samples of qualitative results of the 3D pose estimation approaches in the Human3.6M dataset. Images with 2D pose estimation and 3D pose in a selected view. Qualitative performance in [Chen and Ramanan, 2017] (Upper) and in [Zhou et al., 2017] (Lower).

is straightforward to implement with off-the-shelf 2D pose estimation systems and 3D MoCap libraries (motion capture libraries) to reliably estimated 3D poses in the wild (an intermediate 2D pose approach). Accurate 3D estimations can be obtained without the complexity of fine-tuning multi-stage deep architecture or end-to-end frameworks [Tome et al., 2017]. Typically, there are multiple 3D interpretations for a single 2D skeleton. For this reason, based on [Zhou et al., 2017], we include in the framework a depth regression module and its weakly-supervised 3D geometric constraint loss. We compare the 3D dimensional correspondence based on the nearest-neighbour model with the depth information estimated from the regression module and the 3D geometric constraint in order to improve the geometric validity of the estimated pose. Therefore, once the 2D pose estimation in videos is performed using PAF and CPM [Cao et al., 2017, Wei et al., 2016], and the tracking

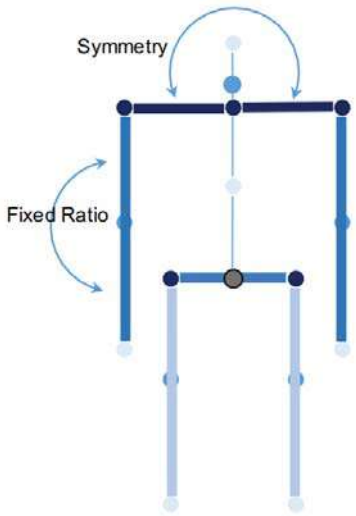


Figure 5.15 Schematic of the weakly-supervised geometric constraint. Left/right shoulder bones share the same length and upper/lower arms have a fixed length ratio. Image adapted from [Zhou et al., 2017].

scheme [Song et al., 2017], we predict the 3D skeleton based on the 2D correspondence and the 3D geometric correction following the instructions in [Chen and Ramanan, 2017, Tome et al., 2017].

Given a 2D skeleton $\mathbf{x} = [x, y]$, the prediction of its corresponding 3D skeleton $\mathbf{X} = [X, Y, Z]$ would not be affected by 2D image I measurements. This conditional independence is a reasonable approximation to write the joint probability as,

$$p(\mathbf{X}, \mathbf{x}, I) = \underbrace{p(\mathbf{X}|\mathbf{x})}_{\text{NN}} \cdot \underbrace{p(\mathbf{x}|I)}_{\text{CNN}} \cdot p(I), \tag{5.7}$$

where the first term is estimated with a non-parametric nearest-neighbour (NN) and the second term corresponds to the framework that predicts 2D keypoint heatmaps. A library of 3D poses $\{\mathbf{X}_i\}$ is paired with a particular camera projection matrix $\{M_i\}$, such that the associated 2D poses are given by $\{M_i(\mathbf{X}_i)\}$. By considering multiple cameras for a single 3D pose, the distribution over 3D poses is based on a reprojection error. In order to reduce the squared reprojection error involves solving a camera resectioning problem, where an interactive solver can be initialised with M_i,

$$M_i^* = \underset{M}{argmin} \; || M(\mathbf{X}_i - \mathbf{x}) ||^2 . \tag{5.8}$$

Warping exemplars are introduced to match better the 2D pose estimations, which can be seen as an inverse kinematics optimisation problem. First, the 3D exemplar is aligned to the camera-coordinate system used to compute the projection $\mathbf{x}$. Given this alignment, the system replaces the (X_i, Y_i) exemplary coordinates with their scaled 2D counterparts (x, y) under a weak perspective camera model given by,

$$\mathbf{X}_i^* = [sx, sy, Z_i], \tag{5.9}$$

where $s = \frac{average(Z_i)}{f}$, f is the local length of the camera (given by the intrinsic in M_i) and $average(Z_i)$ is the average depth of the 3D joints.

Finally, the proposed geometric loss, $L_{geo}(\hat{Y}_{dep}|Y_{2D})$, is based on the fact that ratios between bone lengths remain fixed in a human skeleton as it is illustrated in Figure 5.15. Given R_I a set of involved bones in a skeleton group i, l_e the length of bone e and $\bar{l}_e$ the length of bone in a canonical skeleton (set as the average of all training subject [Zhou et al., 2017]), the ratio $\frac{l_e}{\bar{l}_e}$ for each bone e in each group R_i should have a variance of 0. Therefore, the loss measures are the sum of variance among $\{\frac{l_e}{\bar{l}_e}\} e \in R_i$ of each R_i,

$$L_{geo}(\hat{Y}_{dep}|Y_{2D}) = \sum_i \frac{1}{R_i} \sum_{e \in R_i} (\frac{l_e}{\bar{l}_e} - \bar{r}_i)^2, \tag{5.10}$$

where $\bar{r}_i = \frac{1}{R_i} \sum_{e \in R_i} \frac{l_e}{\bar{l}_e}$ and L_{geo} is continuous and differentiable with respect to $\hat{Y}_{dep}$.

Movement quantification and temporal analysis

The kinematic representation of the movements from the head and upper limbs is denoted as $[\Delta x, \Delta y, z]$, where Δx and Δy are the changes in the position x and y between the current frame and the previous one, and z is the depth estimation for each frame. Each sequence of features of the movements has a dimensionality of [25,24], capturing 25 frames, each with 24 features corresponding to 8 points for the three dimensions. We consider that the variation of the key body points or landmarks is an important representation of the body location and the LSTM has the advantage of modelling the evolution of properties to capture the dynamic variation of the bodies physical structure. This strategy is different from the approach in Section 4.3.2 for the analysis of facial landmarks, which aim to compute metrics from the movement trajectories. However, this method is limited in the selection of appropriate spatial features from each landmark trajectory. In this scenario, the changes in position are directly fed to an LSTM and the output is fed into a densely-connected layer with a sigmoid activation function to estimate the epilepsy type. For each LSTM, we experimented with various numbers of layers and memory cells based on the dimensionality of the features, and we choose to use one single LSTM layer, with 8 hidden units. To train the LSTM network, we follow the same procedure to train the network to classify facial semiology in Section 4.3.3. Training is carried out by optimising the binary cross entropy loss function and using the Adam optimiser. We perform the model training using 50 epochs and use the default initialisation parameters from Keras [Chollet, 2015] to initialise the weights of the LSTM hidden units with a Theano backend [Al-Rfou et al., 2016].

5.2.3 Fusion techniques

The fusion procedure allows designing systems robust against noise and failures as well as improving reliability and accuracy with interpretable models [Dumas et al., 2009, Oviatt, 2003]. The fusion of different modalities is not a straightforward task and it can be executed in three different levels: at

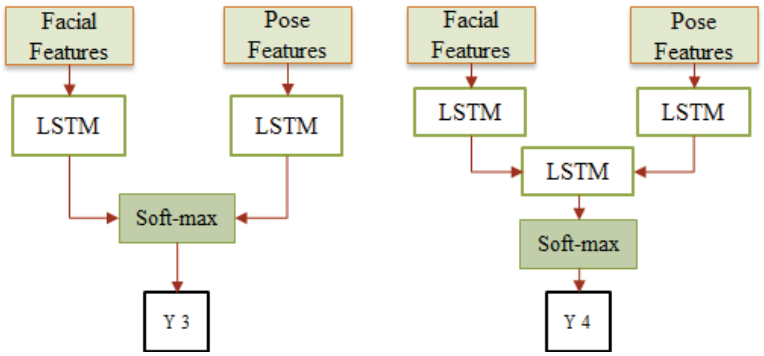


Figure 5.16 Simplified diagrams of the two fusion approaches. Sequence-to-one LSTM (left) and sequence-to-sequence LSTM (right).

the data level, at feature level (early fusion), and a decision level (late fusion). Based on Figure 5.16, which is a simplified diagram of the fusion strategy illustrated in Figure 5.2, we adopt an early fusion level. In this fusion method, also known as fusion in feature space, there is only one learning phase handling all features from different clinical manifestations, *i.e.*, facial expressions and pose estimation.

To conduct the fusion at the feature level, we merge the two networks using the LSTM outputs of each modality with two different stream approaches where joint back propagations between streams are plausible [Gammulle et al., 2017]. We take facial features extracted from the last fully connected layer with a dimension of $[25,4096]$ per sequence (see Section 4.3.3) and the representation of the head and upper limb movements for each sequence with a dimension of $[25,24]$ (see Section 5.2.2). Given the facial features X_{face} and the pose features X_{pose}, the LSTM outputs of each modality for the i^{ih} video sequence are defined as,

$$\begin{aligned} h^i_{face} &= LSTM(X_{face}), \\ h^i_{pose} &= LSTM(X_{pose}). \end{aligned} \tag{5.11}$$

The *first fusion method* merges the output of each LSTM through a soft-max layer to evaluate the final classification. This is known as sequence-to-one LSTM (Figure 5.16 left) and can be represented as,

$$y^i = softmax(W[h^i_{face}, h^i_{pose}]). \tag{5.12}$$

The *second fusion method* uses the hidden states of each LSTM as input to another LSTM, which generates a single hidden unit representing the entire video. This is called a sequence-to-sequence LSTM. The motivation behind having multiple layers of LSTMs is to capture information hierarchically. With this approach, each type of semiology is able to communicate with each other and improves the back-propagation process. For this extra LSTM, we adopt one single LSTM layer, with 4 hidden units. Given the LSTMs use in a sequence-to-sequence scheme represented as $LSTM^*$,

$$
\begin{aligned}
h^{i,t}_{face} &= LSTM^*(X^{i,t}_{face}, h^{i,t-1}_{face}),\\
h^{i}_{face} &= [h^{i,1}_{face}, h^{i,2}_{face}, ..., h^{i,T}_{face}],\\
h^{i,t}_{pose} &= LSTM^*(X^{i,t}_{pose}, h^{i,t-1}_{pose}),\\
h^{i}_{pose} &= [h^{i,1}_{pose}, h^{i,2}_{pose}, ..., h^{i,T}_{pose}].
\end{aligned}
\tag{5.13}
$$

Then, the resultant sequence of predictions are fed to a final LSTM which works in a sequence-to-one manner (Figure 5.16 right), where the outputs are concatenated into a single densely connected layer with a softmax activation function to predict the label classification for each data sequence,

$$
\begin{aligned}
h^i &= LSTM(W[h^i_{face}, h^i_{pose}]),\\
y^i &= softmax(h^i).
\end{aligned}
\tag{5.14}
$$

5.2.4 Experimental results

Dataset specification

We select video recordings that were possible to identify semiology from face, head and upper limbs motions simultaneously, including small segments of less than 5 seconds. This is important because in order to perform a fusion strategy, the face and pose features must have the same number of frames recorded. This segmentation was performed manually using the data available until August 2017 (see Section 3.2 and Table 3.2). A video set from 18 patients diagnosed with MTLE and ETLE were considered for the first experiment in the literature of a fusion approach of different types of semiology. A total of 52 video clips were processed, consisting of 40 videos from 12 patients with MTLE (Class 1) and 12 videos from 6 patients with ETLE (Class 2).

Experimental setup

Similar to the experimental setup proposed in Section 4.4.2, we adopt a *k*-fold cross-validation and LOSO-CV to evaluate the fusion approach. With the *k*-fold cross-validation, we analyse the flexibility of the system to model the variations in clinical seizure data. The sequences used for validation and testing are completely separate to that used for training the models, but it is possible to have sequences from the same patient in each set. The validation and test accuracy of the framework is computed as the average performance of each fold (10-folds in this experiment).

The LOSO-CV allow us to evaluate the subject-to-subject variation that occurs. This results in an unbiased estimation of the true generalisation error [Xu and Huang, 2012]. We test on the entire video corpus for a specific patient, who is totally excluded from the training data. Each patient will, at some point, constitute a “test set” all of its own. The prediction accuracy is computed as the average of all patients accuracies.

Figure 5.17 Qualitative results of 2D pose estimation in a sequence from the epilepsy dataset. Dystonic posturing with heat maps (Upper), right-hand automatism with limbs detection (Lower).

Head and upper limbs detection

Figure 5.17 illustrates qualitative results of the 2D pose estimation in a sequence during a seizure. Heatmaps show the estimated joint locations. The percentage of detected joints is considered to measure the performance of the 2D pose estimation in videos. The joint is considered detected if the distance between the predicted and ground truth location is within a range of 10 pixels (user-set range). The evaluation of detected joint accuracy using convolutional operations and spatio-temporal adjustment for the 2D pose predictions achieved an accuracy of 93.4% in selected manually annotated videos (see Section 3.3.1), where the wrists and elbows estimation was the most challenging. With this experimental evaluation, we confirm the advantage of integrating time information rather than simply tracking the association between key points in successive frames, which achieved 91% accuracy on the detected joints.

The 3D pose estimation in a selected patient is depicted in Figure 5.18. Although the 3D positions estimated from extracted 2D pose increases the number of features that characterise the body motion, this process also adds noise to the system. Although it is possible to further increase the number of extracted measurements by computing handcrafted features from the 2D tracking information as it was documented in [Cunha et al., 2016a], we argue that further effort would better be spent improving 3D joint location directly from colour images. For example, it is feasible to create a new library of 3D structures only from videos during seizures, without using pre-defined 3D position of a different domain (see Table 3.9).

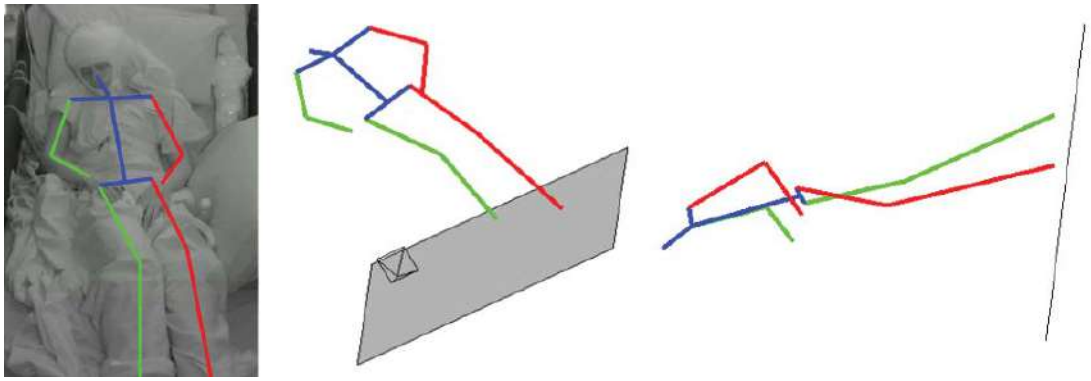


Figure 5.18 Qualitative results of 3D pose estimation in the epilepsy dataset. 3D pose estimation with different view camera angles.

Seizure classification with the fusion approach

To quantify facial semiology, we extract deep features from the detected facial bounding box using the fine-tuned face detector [Jiang and Learned-Miller, 2017]. These features are characterised by the hidden layer activation from the last fully connected layer, which has 4096 units. Then, sequences of these spatial representations of facial expressions are used as input to an LSTM model with a network configured with 2 hidden layers of 128 and 64 hidden units, respectively. To estimate the output $Y1$ as depicted in Figure 5.2, we perform classification using a soft-max layer using the trained network discussed in Section 4.3.3.

The performance of each cross-validation approach is evaluated with each type of semiology individually (face semiology output $Y1$ and head and upper limbs semiology output $Y2$) and with the two fusion methods ($Y3$ and $Y4$) as illustrated in Figure 5.2. The multi-fold cross-validation performance is illustrated in Table 5.4, while Table 5.5 describes the validation of LOSO-CV for each patient with MTLE.

Comparing the results produced by each model, the second fusion approach produces the best performance in the 10-fold cross-validation. This approach achieved an average of 85% accuracy on the test set and an average area under the curve of 0.95. This confirmed that the individual sensing modalities integrated by a multi-modal approach can improve the performance over single-modal approaches. Using facial semiology shows similar performance to that discussed in Chapter 4 with less number of video recordings (see Table 4.5); however, using only the head and upper limbs movements results in very low accuracy. This is likely due to the semiology from body movements being uncommon in our dataset and the number of spatial features being small, which infers that there are insufficient discriminative features to classify each type of epilepsy. Despite this, the fusion approach is able to identify complementary information to analyse semiology. Additionally, the performance is further improved by adding a third LSTM network for late fusion, as this provides the

Table 5.4 Multifold cross-validation performance with the fusion approach.

Model	Output	Validation Accuracy (%)	Test Accuracy (%)	Test Sensitivity (%)	Test Specificity (%)	Test Precision (%)	AUC
Facial Expression	Y1	90.10	80.90	70.00	90.00	87.5	0.89
Head & Upper Limbs movements	Y2	60.67	55.50	60.00	60.00	60.00	0.76
Fusion 1	Y2	91.30	82.50	81.00	84.00	84.00	0.91
Fusion 2	Y2	**92.10**	**85.00**	80.00	90.00	80.00	**0.95**

Table 5.5 LOSO-CV performance for patients with MTLE with the fusion approach.

Test. Acc (%)	Test Patient												Average
	1	2	3	4	5	6	7	8	9	10	11	12	
Y1	45.6	32.3	64	52.67	23.2	82	14.3	67	50	80.6	52.4	8.5	47.71
Y2	21.5	4.5	12.2	32.4	74.2	5.6	19.8	38.4	15.1	9.72	42.4	60	27.99
Y3	46.2	32.2	66.2	55.6	73.2	82	15.2	70.1	50	82.6	61.2	60.1	57.89
Y4	43.4	32.3	67.4	56.7	74.8	82	**16.3**	72.3	50	**83.1**	63.5	60.1	**58.49**

model with the additional capacity to model the relationship, and explicitly consider the two streams temporally.

In contrast, the LOSO-CV highlights the sensitivity of the system to the representativeness of data when classifying semiology. This framework reached an average test accuracy of 58.49% and the classification performance ranges from 16.3% to 83.1% with the best fusion approach. This indicates that for particular cases such as Patient 1, 2, and 7, the observed semiological patterns are not similar to the features of other patients in the dataset. This infers that while performance at present is limited, a larger-scale training dataset that better captures the wide variety of semilogical patterns that can occur will result in significantly improved performance.

One major drawback of the fusion approach is that the fusion strategy may not be applicable to every patient because not all patients experience face and body semiology simultaneously. In order to merge the LSTMs, it is required that each modality (face and pose) has the same number of frames in the sequence, which is a condition difficult to accomplish.

One single clinical manifestation cannot be used as a tool to conclude the neuronal networks affected. Each ictal semiology contributes equally to the final decision about the most effective therapeutic option. For this reason, the hierarchical approach should be more suitable to assess the patient because the decision of the system is independent of the type of semiology.

5.3 Multi-modal strategies to classify epilepsies: hierarchical approach

In this strategy, we present an approach where the classification of epilepsy types is carried out hierarchically based on the individual result of clinical manifestation in the form of facial expressions, head and upper limbs motions, and hands and fingers movements. By hierarchy, we mean that the system is composed of interrelated sub-systems. A hierarchical approach tackles complex problems by reducing them to a smaller set of interrelated problems, where they are solved separately and the results are analysed to find a description of the symptomatology of the patient. In this scenario, we quantify and classify epilepsy types based on the result of each isolated semiology to evaluate the general patient's condition. The hierarchical strategy can provide findings from different perspectives with improvements in the time and space complexity for both learning and execution.

The importance of each clinical sign is based on the development and sequence of multiple semiological features to identify the seizure initiation and propagation. Each ictal semiology contributes equally to the final decision about the most effective therapeutic option. For this reason in this strategy, we do not perform decision fusion to calculate a final quantitative score from the results of each classifier. The result of each type of semiology should be correlated with the brain electrical and neuroimaging findings, whose concordance is necessary for the ultimate evaluation of epilepsy.

To achieve the hierarchical results, we use a framework, the structure of which is presented in Figure 5.19. The proposed approach starts by detecting the patient in the video. Once the patient is detected, spatial information of the face, head, upper limbs and hands are extracted using CNNs. Then, temporal information from a sequence of frames (25 consecutive video frames for each modality), is extracted using an LSTM architecture. Lastly, for each clinical manifestation, the spatiotemporal features are evaluated for classifying MTLE and ETLE.

This section is distributed as follows. Section 5.3.1 provides an improved approach to quantify facial semiology; To quantify semiology from the head and upper limbs movements, we adopt the approach of Section 5.2.2. Section 5.3.2 introduces a new framework to quantify semiology from hands and fingers motions; Finally, Section 5.2.4 summarises the dataset used and the results in classifying epilepsy types.

5.3.1 Facial semiology

In this section, we enhance the vision-based approach discussed in Section 4.3.3 for the automatic quantification and classification of facial semiology. The extraction of facial features is based on the fine-tuned face detector [Jiang and Learned-Miller, 2017], which uses the Faster R-CNN framework [Ren et al., 2017]. In contrast to the approach proposed in Section 4.3.1, which extracts features from the face-bounding box detected by performing a process of tracking by detection, we adopt a tracking technique from [Jin et al., 2017]. This methodology aims to improve the stability of the bounding box detected regarding size and position in each frame and to recover from tracking errors where a

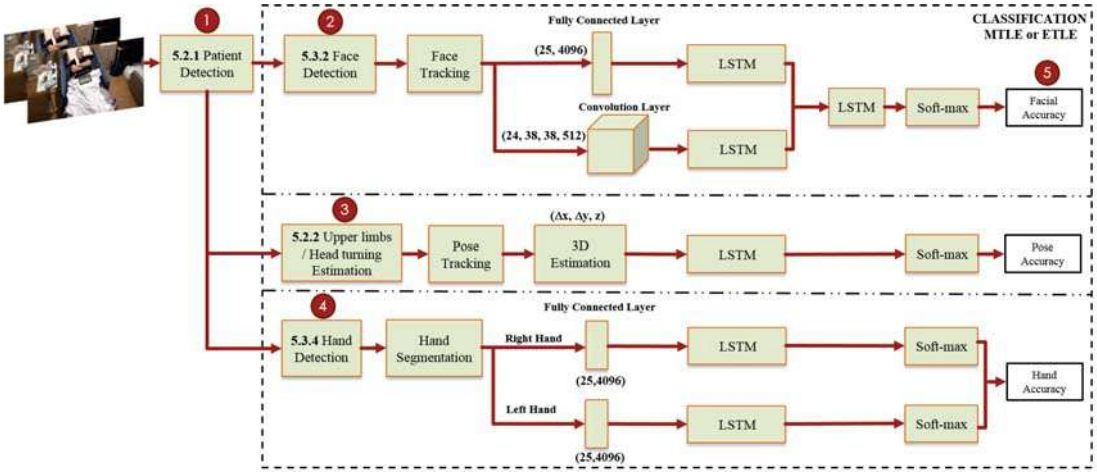


Figure 5.19 Multi-modal analysis of semiology: hierarchical approach of the face, head, upper limbs, and hands semiology. **1.** Detection of the patient (see Section 5.2.1). **2.** An improved approach to assess facial semiology which extracts spatial information from two hidden layer activations. These features are merged and the temporal information is computed to perform the classification. **3.** 2D pose localisation in videos and 3D inference are used to quantify semiology from the head and upper limbs motions (see Section 5.2.2). **4.** A novel region-based approach is proposed to evaluate semiology from hand and fingers motions. The hand accuracy is defined as the average performance between the left and right-hand classification. **5.** The output of the system is represented by the classification accuracy of each ictal symptom, which will be used by clinical experts to conduct further evaluation of each patient.

face is not detected because of conditions such as occlusions. This approach allows the combination of the high-quality face detector with a generic tracking architecture known as distribution fields for tracking [Sevilla-Lara and Learned-Miller, 2012]. The tracker's aims to find in the next frame the object most similar to the target in the current frame. Tracking helps not only to catch false negatives but also to link faces of equivalent identity in different frames. The tracker is executed forward and backwards in time from every single face detection for a fixed number of frames. In the joint detection-tracking strategy [Jin et al., 2017], the authors applied the intersection over union (IoU) and the Hungarian algorithm [Kuhn, 1955] to establish the correspondences between detection and tracking results in each frame. This detection-tracking strategy was selected as a robust option to improve the detection rate for all frames, compared to proposals that improve the tracking by detection by performing a method of bounding box aggregation where the results of different face detectors are merged in each frame [Feng et al., 2017]. One major disadvantage of the bounding box aggregation is that all face detectors may not detect the face in one specific frame, while in our proposal we use the information of the face detected before or after this event.

In this facial analysis approach, rather than extract a single type of feature from the face bounding box (see Section 4.3.3), we extract two different spatial features from the hidden layer activations: the last convolutional layer X_{conv} and the last fully connected layer X_{fc}. The network

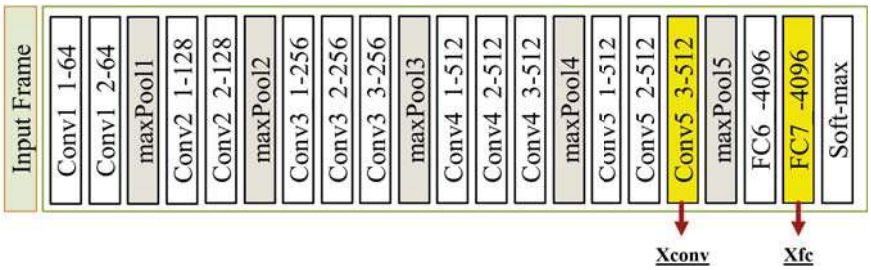


Figure 5.20 VGG-16 architecture used for the features extraction facial semiology. The features are extracted from the last convolutional layer and the last fully connected layer.

architecture of the face detector based on the VGG-16 network [Simonyan and Zisserman, 2014] is illustrated in Figure 5.20. In this figure, convolutional layers are represented as $Conv[\text{name of layer}]-[\text{number of channels}]$ and fully connected layers as $FC[\text{name of layer}]-[\text{number of channels}]$. Once the features for each frame are extracted, the representation for each sequence has a dimension of $[25,38,38,512]$ for the convolution features and $[25,4096]$ for the fully connected features. The approach in Section 4.3.1 may omit some important features, as features in convolutional layer outputs that have wider reception fields, resulting in a greater granularity [Sun et al., 2018]. We are interested in evaluating the hypothesis that the convolution layer contains more spatial information and the fully connected layer has more discriminative features or semantic information, which may enhance the facial expression detection [Gammulle et al., 2017].

A simplified block diagram of the new proposed system to analyse facial semiology used on the hierarchical approach is presented in Figure 5.21. In this approach, the spatiotemporal features are extracted from overlapping windows of facial sequences to capture the dynamic temporal aspects of the facial semiology [Nguyen et al., 2017]. To fuse the two types of features extracted from the facial bounding box, we adopt a sequence-to-sequence LSTM approach (see Section 5.2.3). The sequence of the spatial representation of each feature, convolution and fully connected layer, is used as an input to an LSTM model to exploit the temporal relation between video frames. We merge the LSTM outputs by using the hidden states of each LSTM as input to another LSTM, which generates a single hidden unit representing the entire video [Gammulle et al., 2017]. We experiment with different numbers of layers and memory cells based on the feature dimensionality, and we choose the best configuration with two stacked LSTM layers, each with 128 memory cells. More complex architectures do not show significant performance gains. The outputs of the final LSTM are concatenated into a single densely connected layer with a soft-max activation function to make a single prediction for every sequence of each seizure. To train the LSTM network, we follow the procedure to train the network to classify facial semiology in Section 4.3.3. Training is carried out by optimising the binary cross entropy loss function and using the Adam optimiser. We perform the model training using 250 epochs and use the default initialisation parameters from Keras [Chollet, 2015] to initialise the weights of the LSTM hidden units with a Theano backend [Al-Rfou et al., 2016].

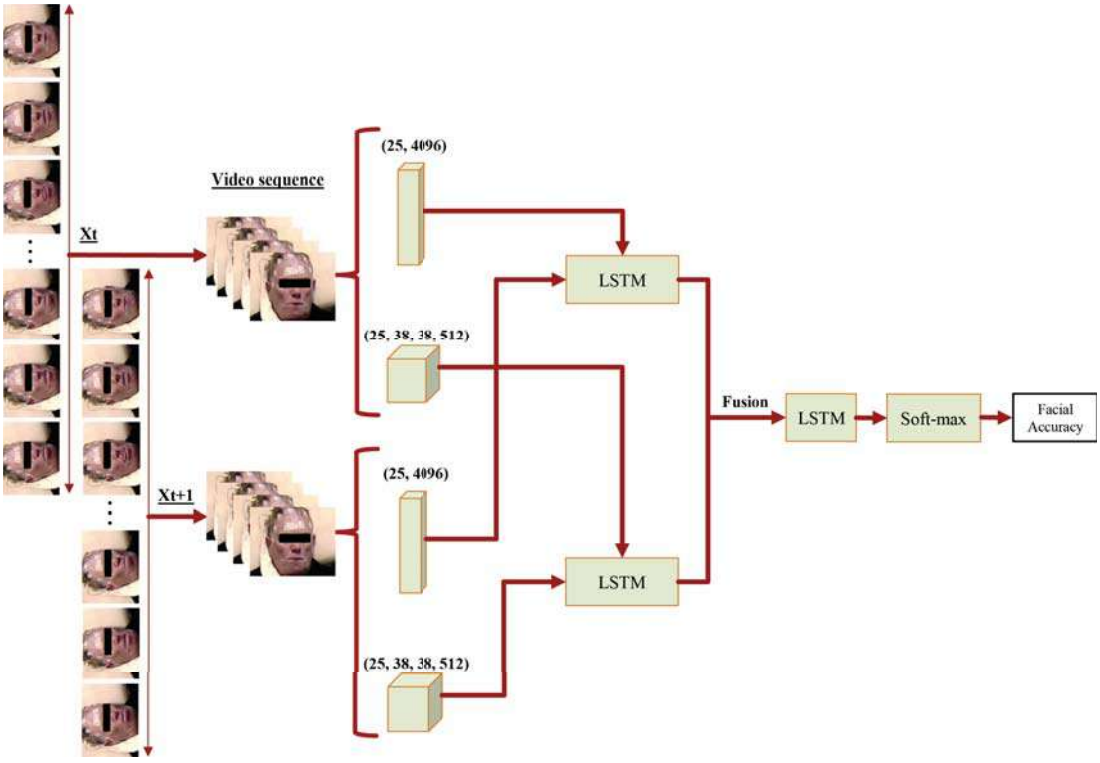


Figure 5.21 Framework proposed to quantify facial semiology and classify epilepsy types. Spatio-temporal features are extracted from overlapping windows and a fusion approach is performed to identify complementary information to analyse the facial semiology of the patient.

5.3.2 Hands and fingers semiology

In this section, we introduce a pioneering methodology to quantify semiology that is detected from hand and finger movements such as hand dystonia, finger claw position, tapping or grabbing of objects or bedclothes, snapping the fingers, fisted hand extension, thumb adduction or flumbing [Bleasel et al., 1997, Jobst et al., 2000, Noachtar and Peters, 2009]. Hand analysis is very challenging because of numerous variations in hand images in real clinical scenarios such as low-resolution, high levels of occlusion, low lighting, blurring due to hand movement and relative viewpoint changes due to pose changes.

In order to validate semiology from hands and fingers motions, two methods were considered: a landmark-based and a region-based approach. These two methods are inspired by different strategies proposed to analyse facial semiology in Chapter 4. We adopt these approaches in order to evaluate and verify which method excels at detecting the location, orientation and articulation of the hands encountered in a natural clinical monitoring setting.

The landmark-based approach or hand keypoint estimation follows similar methodologies of facial landmarks estimation and human pose estimation. We capitalise on the capability of deep networks to learn hand pose estimation from single colour images without requiring depth

Table 5.6 Selected benchmarking techniques for hand pose estimation.

Author	Title
[Baltrušaitis et al., 2018]	Using a single RGB frame for real time 3D hand pose estimation in the wild
[Simon et al., 2017]	Hand keypoint detection using multiview bootstrapping
[Mueller et al., 2017]	Real-time hand tracking from an egocentric RGB-d sensors
[Zimmermann and Brox, 2017]	Learning to estimate 3D hand pose from single RGB images
[Sridhar et al., 2016]	Real-time joint tracking of a hand manipulating an object from RGB-D input
[Tzionas et al., 2016]	Discriminative salient points and physics simulation

information [Simon et al., 2017, Zimmermann and Brox, 2017]. The method operates in sequence, the hand is detected, after which the estimation joint locations are performed in the detected hand region. Then, movement quantification and temporal analysis are performed based on the detection and tracking of the hand keypoints in videos. In contrast, the region-based method extracts spatial features from the detected hand-bounding box, which are subsequently fused to capture the temporal change of the whole hand. This methodology follows the architecture documented for the facial semiology analysis in Section 4.3.3.

Landmark-based approach

Methods that can localise hand joints in single RGB images are not as common as the image-based face and body keypoint localisations. This is surprising given the significant role of the hands. One main reason is that there are not enough large datasets of annotated keypoints for hands (see Table 3.11). The problem, however, shares several properties with human body pose estimation and many approaches proposed for the human body can be adapted for hand pose estimation. Traditional approaches have shown excellent accuracies but under highly controlled conditions. Additionally, with the introduction of low-cost depth cameras, few proposals rely on depth and synthetic data [Zhou et al., 2016c, Oberweger et al., 2015a, Oberweger et al., 2015b, Tang et al., 2014, Tompson et al., 2014], or they only perform tracking based on an initial pose rather than full pose estimation [Sharp et al., 2015]. We assess proposals that enable pose hand tracking in RGB videos based on qualitative performance in our dataset, documentation and architecture. Table 5.6 summarises the most significant approaches for 2D hand keypoint detection.

In order to quantify semiology based on hand pose estimation, we adopt the proposal of Simon et al. [Simon et al., 2017], which is a robust multi-view bootstrapping architecture to generalise and produce 2D-3D locations of hand joints. Figure 5.22 depicts qualitative results of the method selected in benchmark datasets (see Table 3.11).

The single view detector, triangulated over multiple views, enables 3D markerless hand motion capture with complex object interactions. Discriminative methods that rely on deep architectures, require large datasets [Tang et al., 2014]; however, the multiview bootstrapping technique allows the generation of large annotated datasets using a weak initial detector, and provides automated supervision based on multiview geometry and keypoint detection.

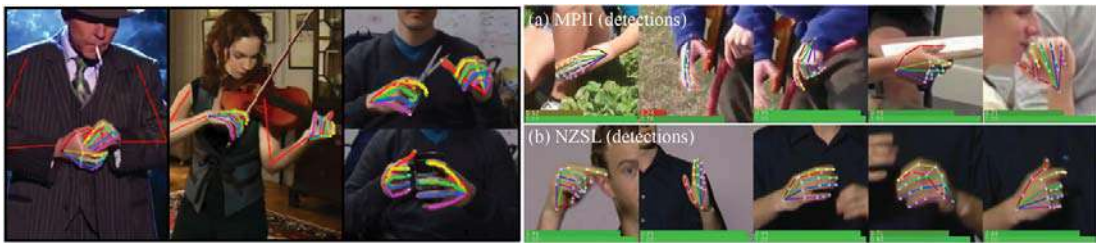


Figure 5.22 Selected samples of qualitative results of the hand keypoint detection in Youtube and Webcam videos (left side), and the MPII human pose and the New Zealand sign language datasets (right side). Image adapted from [Simon et al., 2017].

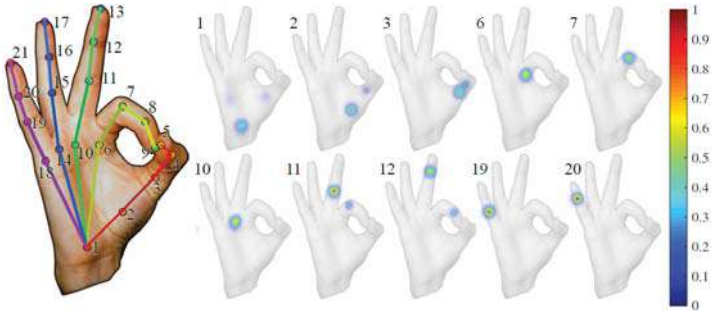


Figure 5.23 Representation of the distribution of each hand keypoint and the confidence maps of detection. Image adapted from [Simon et al., 2017].

The keypoint detector follows the architecture of [Wei et al., 2016], but uses a pre-initialized VGG-19 network [Simonyan and Zisserman, 2014]. This detector predicts a confidence map for each keypoint representing the keypoint's location, where the final position is obtained by finding the maximum peak in each confidence map. The output of the detection phase is 21 heatmaps that correspond to one wrist point and 20 hand keypoints (four per finger). Figure 5.23 illustrates the distribution of the keypoints and selected confidence maps. A hand detection phase is performed prior to the keypoint detection based on a body keypoint prediction [Cao et al., 2017], and the location of the wrists. The hand bounding box detection aims to improve the keypoint detection via confidence maps.

For every frame, the algorithm runs the detector and robustly triangulates the point detection. The detector d_i on each image I_v^f, yields a set D of 2D location candidates,

$$D \leftarrow d_i(I_v^f) \text{ for } v \in [1...V], \tag{5.15}$$

where each keypoint p have V detections (X_p^v, c_p^v). X_p^v corresponds to the location of point p in view v and $c_p^v \in [0,1]$ to the confidence measure. To triangulate each point p into a 3D location, the method use RANSAC [Fischler and Bolles, 1981] on points in D with confidence above a detection threshold λ and minimise the reprojection error [Agarwal et al., 2012] to obtain the final triangulation

Figure 5.24 Qualitative result of the hand keypoint detector in the epilepsy dataset.

position,

$$X_p^f = \underset{X}{argmin} \sum_{v \in \tau_p^f} || P_v(X) - x_p^v ||_2^2, \tag{5.16}$$

where τ_p^f is the inlier set, with $X_p^f \in \mathbb{R}^3$ the 3D triangulated keypoint p in frame f, and $P_v(X) \in \mathbb{R}^2$ stands for projection of 3D point X into view v. Figure 5.24 depicts qualitative results of the pose estimation for both hands in a selected sequence during a seizure (tapping or grabbing).

The landmark-based method is more intuitive because the features extracted are visually related to the position of the fingers and will likely yield good results, but the performance is diminished because the keypoints are not consistently detected across sequences. Furthermore, the detection can also vary in their location in small proportion across consecutive frames, which is a common limitation of the facial landmarks approach (see Section 4.3.2). Keypoints are frequently lost or miss-detected such that very large (and erroneous) movements are detected between frames. These errors are caused by hand images being small and/or frequently occluded. Overall, these errors mean that hand landmarks are not suitable for semiology analysis, as we cannot guarantee the correctness of the underlying features. Figure 5.25A displays a clear example on how the key points are not all detected during a sequence of the full length of the seizure.

In contrast, a region-based method allows the extraction of a sufficient number of features to conduct a process of categorisation of semiology based on a consistent and robust hand detection that considers an RGB-only marker-less hand tracking problem. Figure 5.25B depicts how the hand-bounding box is consistently identified for both hands. Therefore, it can be argued that a region-based approach based on the extraction of features from this bounding box detected can be more convenient for the quantitative analysis of hand semiology analysis.

Region-based approach

A block diagram of the new framework for hand analysis during seizures is displayed in Figure 5.26. The region-based method extracts spatial features from the detected hand bounding box; therefore, we need to select the appropriate methodology to detect the area of interest, which should achieve a high-quality detection rate and keep runtime overheads low. For each hand, if they are both detected, the sequential features are fed to an LSTM layer to model the spatio-temporal features of the hand. Similar to the facial analysis, the output of the hidden recurrent layer is fed into a densely connected layer with a soft-max activation function to predict the class probability. The final hand accuracy

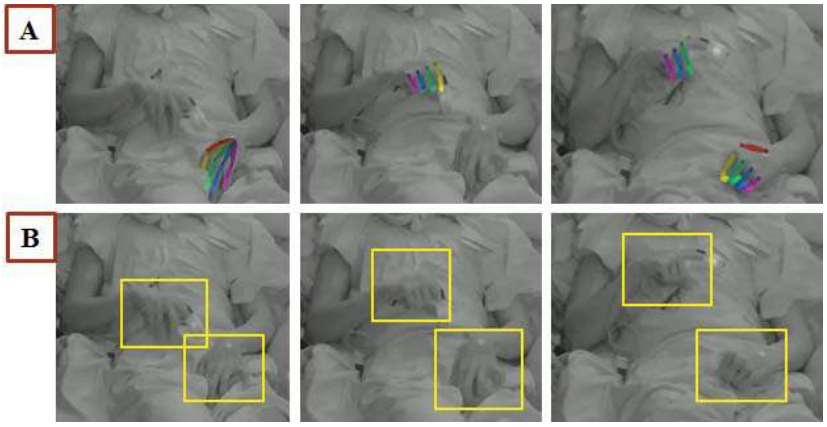


Figure 5.25 Selected sample that illustrates the disadvantage of the hand keypoint strategy. **A.** 2D hand pose estimation in a sequence. **B.** Hand bounding box detection for the region-based analysis.

is defined as the average performance between the left and right-hand classification of the ictal movement detected.

Hand detection has progressed significantly with the use of CNNs and different architectures have been proposed. Table 5.7 displays the most representative proposals in the literature that provide robust hand detection. To this end, we choose to follow the architecture from [Simon et al., 2017], where the hand-bounding box detection is performed prior the body key-point prediction, allowing the search area to be heavily constrained on the predicted pose [Wei et al., 2016, Cao et al., 2017]. Therefore, this strategy is not limited for large datasets of annotated hands to train the system because incorporates trained architectures for human pose estimation, which is a task that can be addressed with more datasets available (see Section 3.3.2). This approach keeps runtime overheads low while also achieves a high-quality detection rate with a low number of false positives. Additionally, the dimension of the hand bounding box avoids the fingers of the patient being located outside of this boundary due to fast movements during a seizure. Figure 5.27 depicts an example of the hand detected and events when a different architecture fails in capturing the totality of the hand, *i.e.*, some fingers are not included inside the bounding box. The hand detection is evaluated on two challenging hand databases: Oxford [Mittal et al., 2011] and VIVA Challenge datasets (see Table 3.10). For each hand in the input image, the detector produces a likelihood estimated and the coordinates of the hand's bounding box.

The hand detector from the hand keypoints detection strategy [Simon et al., 2017], is based on the VGG-19 network [Simonyan and Zisserman, 2014]. For each sequence of hands detected, the feature map F is extracted from the last convolutional layer which represents the feature map input of the following prediction stage that produces score maps for each keypoint. Figure 5.28 displays the VGG-19 architecture where the convolution layers are represented as *Conv*[name of layer] – [number of channels]. The hidden layer activation is extracted with a dimension of [25; 46; 46; 128] for the left and right hand, and fed to the LSTM network to infer the output for each hand.

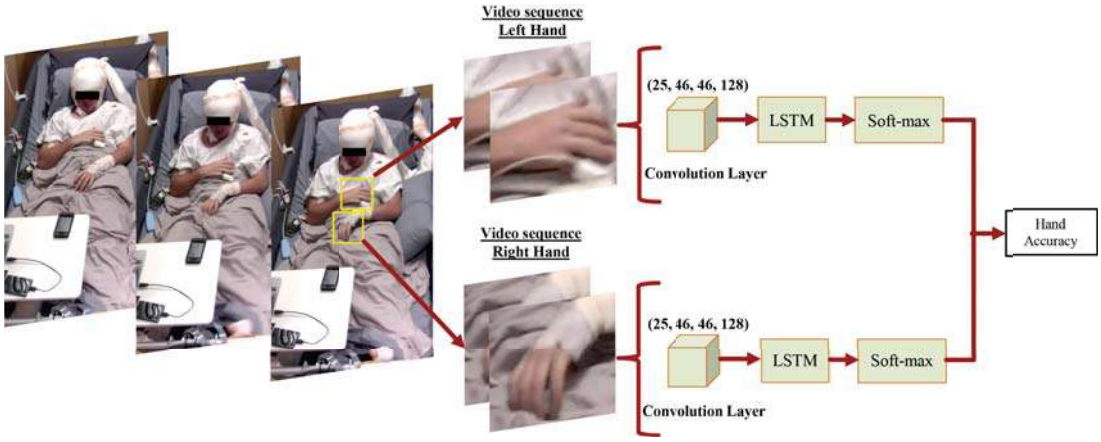


Figure 5.26 Framework proposed to quantify hand semiology and classify epilepsy types. A dataset for hand semiology is created using hand detection. Spatial features of the hand bounding box detected are extracted from the last convolutional layer. The spatio-temporal relation between frames is analysed using an LSTM to predict the class of the sequence.

Table 5.7 Selected benchmarking techniques for hand detection.

Author	Method
[Baltrušaitis et al., 2018]	YOLO v2 [Redmon and Farhadi, 2017]
[Simon et al., 2017]	Body pose in videos [Cao et al., 2017, Song et al., 2017]
[Le et al., 2017]	MS-RFCN
[Yan et al., 2017]	Multiscale Fast R-CNN
[Le et al., 2016]	MS-Faster RCNN
[Zhou et al., 2016b]	Hierarchical context-aware
[Mittal et al., 2011]	Context-based and skin-based

Semantic segmentation of hands: Although the region-based methodology can capture hand motions during a seizure, the computational cost is high and it is still unable to refine the boundaries of the hand clearly. When analysing the features from the bounding box location, there are events where the information of both hands are overlapped or information that is irrelevant to the hand semiology such as motions in bedding and monitoring equipment is captured, which can adversely affect the results. In this scenario, utilising semantic segmentation can assist in accurately locating the region of interest. Recent attention has been focused on semantic segmentation, which is an evolved version of the traditional segmentation task. Compared with the rectangular bounding boxes of the objects output from the traditional segmentation, semantic segmentation attempts to partition the image into semantically meaningful parts with more fine-grained masks rather than rectangular ones. Semantic segmentation can be very useful for medical imaging analysis of the human body due to a more accurate anatomical localisation of the region of interest [Stolojescu-CriŞan and Holban, 2013]. A block diagram of the proposed framework is displayed in Figure 5.29, where the spatial features are extracted from a semantic segmentation architecture instead of the hand bounding box detected.

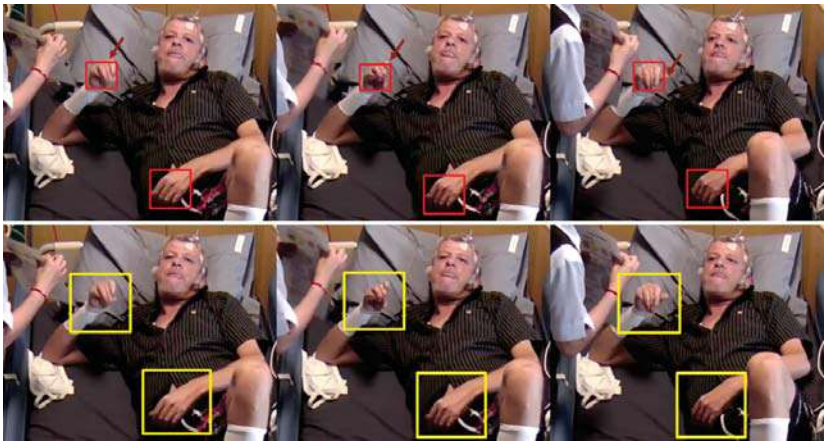


Figure 5.27 Comparative of proposals for hand detection. Hand detection based on a region-based CNN (MS-FRCN) [Le et al., 2017] (Upper). The hand-bounding box detection is constrained on the predicted human pose [Simon et al., 2017] (lower).

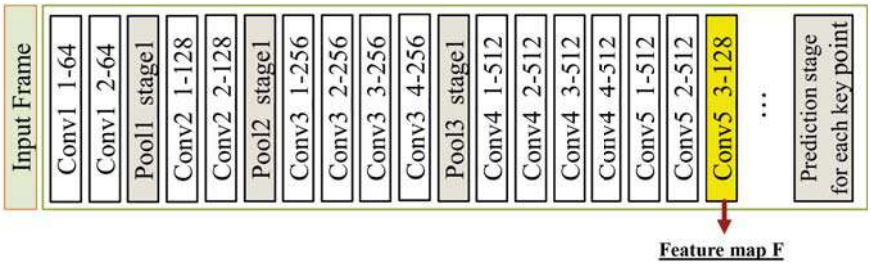


Figure 5.28 VGG-19 architecture used for the features extraction of hand semiology. The features are extracted from the last convolutional layer before starting the prediction stage layers of key points.

While Mask R-CNN [He et al., 2017] is the most widely used semantic segmentation method, its performance is diminished when considering hand segmentation under natural clinical settings. In the existing CNN approaches which are for semantic segmentation, the layers are arranged in a cascaded manner where the successive layers use the outputs from previous layers as the input. By observing the outputs of the Mask R-CNN, it was identified that not all the layers in the network generate features for different modalities. Inspired by this observation, it was defined a new architecture known as VX-Mask R-CNN [Pemasiri et al., 2019b], which is displayed in Figure 5.30. In this architecture, the layers that generate discerning features are fused as a single layer. This layer is then subjected to class predictions, the region of interest identification and mask identification tasks, using a multi-task loss function.

Once the semantic segmentation phase is executed, spatial features are extracted from the last fully connected layer with a dimension of $[1, 4096]$, as the output of the layer in the network has 4,096 units. Each sequence of features of the movements has a dimensionality of [25,4096], capturing twenty-five frames, each with 4,096 features. We further capture the dynamic variations of the hand

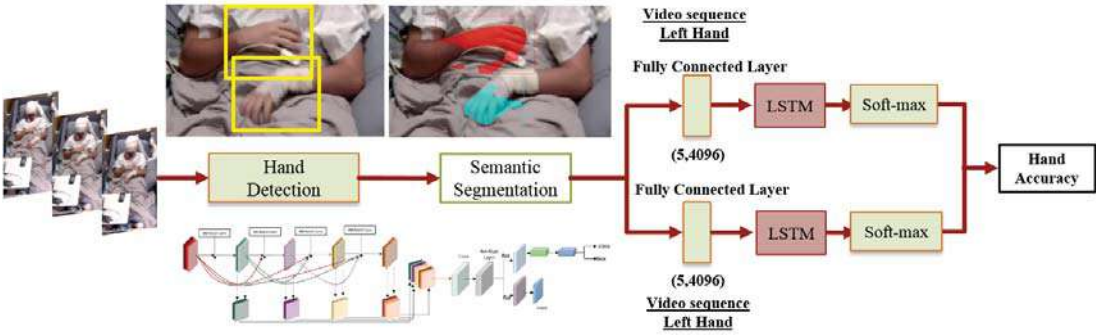


Figure 5.29 Framework enhanced to quantify hand semiology with a semantic segmentation phase. Spatial information is extracted from the last fully connected layer of the semantic segmentation results.

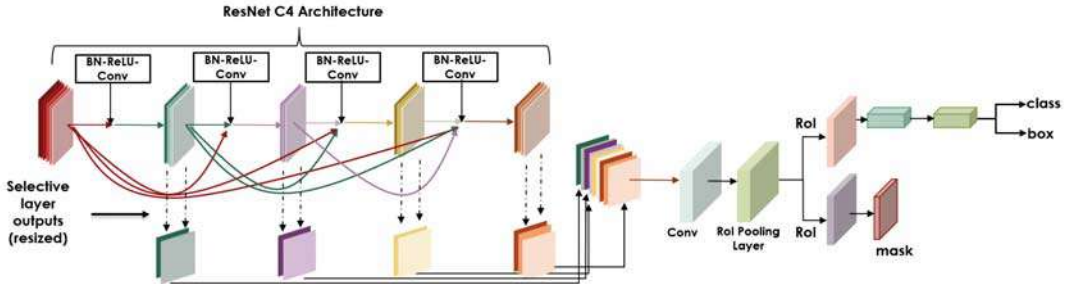


Figure 5.30 Semantic segmentation architecture for hand analysis. The selected layer outputs are resized and fused to generate the new layer, which is then subjected for RoI extraction, mask extraction and class label extraction. Image adapted from [Pemasiri et al., 2019b].

in sequences and estimate the type of seizure using an LSTM network and a single densely connected layer with a soft-max activation function. We adopt a configuration with two stacked LSTM layers, each with 128 memory cells, and we train the LSTM networks following the strategy in Section 4.3.3.

5.3.3 Experimental results

Dataset specification

Existing studies have used minimal numbers of patients and seizures, which can be problematic as seizures vary dramatically for even the same type of epilepsy or between seizures of the same patient. In the hierarchical approach, we have consolidated the best strategies to assess the face, the head, upper limbs, and hands semiology addressing the limitation of the fusion approach discussed in this chapter. For this reason, in this experiment we opt to use all the research data available until October 2017 (see Table 3.2), making this analysis the most important contribution in the evaluation of semiology in this thesis.

Table 5.8 Description of the dataset for the hierarchical approach. Not all patients exhibit all types of semiology, due to differences in the neuronal network affected and other factors such as occlusions which prevent some semiology being analysed.

Epilepsy Type	**Patient**	**Number of Seizure**	**Semiology Identified**			**Number of Video Frames**
			Face	Head/Pose	Hands	
MTLE	1	3	✓	-	-	
	2	3	✓	-	-	
	3	5	✓	-	-	
	4	10	✓	-	-	
	5	5	✓	-	✓	
	6	2	✓	✓	-	
	7	9	✓	✓	-	
	8	3	✓	-	-	
	9	7	✓	-	✓	
	10	10	✓	-	✓	
	11	1	✓	-	✓	
	12	10	✓	-	-	
	13	1	✓	-	✓	
	14	8	✓	-	-	
	15	4	✓	-	✓	
	16	4	✓	-	-	
	17	5	✓	✓	✓	
Total MTLE	**17**	**90**	**17**	**3**	**7**	**68,100**
ETLE	1	5	✓	✓	-	
	2	6	✓	✓	-	
	3	9	✓	✓	✓	
	4	6	✓	✓	-	
	5	2	✓	✓	-	
	6	6	✓	✓	✓	
	7	1	✓	-	✓	
	8	1	✓	-	-	
	9	3	✓	-	✓	
	10	1	✓	-	-	
	11	4	✓	✓	-	
	12	4	✓	✓	-	
	13	4	✓	-	-	
	14	9	✓	✓	✓	
	15	4	✓	✓	-	
	16	3	✓	✓	-	
	17	3	✓	✓	-	
Total ETLE	**17**	**71**	**17**	**11**	**6**	**35,300**
Total Dataset	**34**	**161**	**34**	**14**	**13**	**103,400**

For the hierarchical analysis, we analyse 34 patients, 17 with MTLE and 17 with ETLE. A total of 161 videos clips were recorded, consisting of 90 videos for patients with MTLE (Class 1) and 71 videos for patients with ETLE (Class 2). Table 5.8 displays the demographics of the patients included in each class, where not all patients exhibit face, body and hand semiology simultaneously. This results in a challenging database where some behaviours are very rare, making them difficult to detect.

Experimental setup

To evaluate the hierarchical approach, we also adopt a *10*-fold cross-validation to quantify and model the variations in the data, and a LOSO-CV to demonstrate the robustness of the system in an unbiased

Figure 5.31 Qualitative result of a sequence using the face detector and tracking strategy in the epilepsy dataset. Jitter illustrated in the methodology described in Section 4.3.1. The tracking strategy allows a more consistent facial bounding box localisation (Lower).

estimation of the true generalisation error (see Section 4.4.2 and Section 5.2.4). However, in this evaluation, we propose two approaches of LOSO-CV. In the first LOSO-CV scheme, we test on the entire video corpus for a specific patient, who is totally excluded from the training data. The performance of each ictal semiology is calculated as the average of all patients accuracies for each type of epilepsy. In the second approach, the training process is performed using the totality of the dataset (161 seizure recordings), where 90% of the data are used for training and 10% for validation. To conduct the test, a small number of seizures from patients previously diagnosed with MTLE or ETLE and completely separate from the dataset (Table 3.2), are selected to conduct one single classification for each clinical manifestation. The main difficulty for this evaluation is the limited data available for the training set for all types of semiology, but this is the situation of real clinical scenarios. Detailed specifications for this test dataset are displayed in Table 5.12.

Multi-fold cross-validation for facial semiology

The fine-tuned face detector coupled with the tracking approach reached an average accuracy of 0.945 for the intersection-over-union (IoU) in selected videos manually annotated from the data. This represents an improvement on the face detection performance compared with the strategy of tracking by detection discussed in Section 4.4.3, which achieved 0.920. The facial bounding box is more consistently localised, with reduced jitter between frames (*i.e.*, less detection noise) as illustrated in Figure 5.31. This helps to ensure that extracted features are more consistently extracted from the face.

Table 5.9 summarises the results of conducting a 10-fold cross-validation using the proposed facial classification with the tracking and fusion of features compared with the strategy in Section 4.3.3. The proposed approach achieves 90.00% accuracy on the test set. This result demonstrates that the combination of two feature sets, the last convolutional and the last fully connected layer output, offers higher recognition ability, from a more consistent facial bounding box localisation. The system is still challenged by extreme head rotations which result in these situations the head being poorly localised. In these circumstances, the face bounding box also contains information from the hair, electrodes or bandages (under SEEG monitoring) in addition to the face. In order to avoid incorrect spatial features

Table 5.9 Multifold cross-validation performance in the facial analysis.

Model	Validation Accuracy (%)	Test Accuracy (%)	Test Sensitivity (%)	Test Specificity (%)	Test Precision (%)	AUC
Facial Approach (Chapter 4)	90.25	82.00	82.00	82.00	84.78	0.903
Facial Approach (Fusion features)	92.50	90.00	85.00	95.00	94.37	0.965

Table 5.10 Multifold cross-validation performance in the pose analysis.

Model	Validation Accuracy (%)	Test Accuracy (%)	Test Sensitivity (%)	Test Specificity (%)	Test Precision (%)	AUC
Pose Approach (without tracking)	74.67	60.00	40.00	80.00	66.67	0.71
Pose Approach (with tracking)	77.50	65.00	70.00	60.00	63.64	0.73

being extracted, a process of semi-automatic filtering has been included; however, an automated approach can be considered through the use of a deep regression network [Liu et al., 2014a] or a dual-pathway proposal-refinement architecture [Wu et al., 2018]. This method may allow the detection of specific facial parts under structured spatial constraints and obtain a discriminative part-based representation simultaneously.

Multi-fold cross-validation for head and upper limbs semiology

Considering the strategy proposed in Section 5.2.2 and all the epilepsy dataset, the performance of the 10-fold cross-validation using the tracking strategy or not is given in Table 5.10. According to this result, the performance increased by 5% on the test set. Our approach confirms that the ability to capture temporal correlations among video frames (pose estimation in videos) helps to mitigate mistakes caused by severe occlusions and large rapid motions which blur the video (see Section 5.2.2). It is important to note that pose estimation can also vary in their location in small proportion across consecutive frames which includes noise in the feature extraction. This can also affect the 3D pose estimation because minor errors in the locations of the 2D body joints can have large consequences in the 3D space. However, we have confirmed that by conducting the cross-validation using only information from the 2D estimation, the performance accuracy is reduced by approximately 5%. This result supports the decision to still include 3D estimation because using fewer features (*i.e.*, 2D coordinates only) resulted in a lower validation accuracy.

Multi-fold cross-validation hand and fingers semiology

Given an input image, the resulting hand detector delineates the two hands as well as their bounding boxes. To allow a quantitative evaluation of the hand analysis, without an existing baseline of hand detection systems in epilepsy, we evaluate the performance of the hand detector on the epilepsy dataset with and without semantic segmentation. From selected manually annotated videos, the model

Table 5.11 Multifold cross-Validation performance in the hand analysis.

Model	Validation Accuracy (%)	Test Accuracy (%)	Test Sensitivity (%)	Test Specificity (%)	Test Precision (%)	AUC
Hand approach (without segmentation)	90.00	75.00	80.00	70.00	72.73	0.83
Hand approach (with segmentation)	93.4	82.7	88.00	75.00	88.00	0.89

reached an acceptance average accuracy of 93.2% in which the IoU for each frame was greater than a threshold defined as 0.75. Figure 5.32 shows qualitative results of each hand detected under different scenarios, day and night monitoring.

Table 5.11 shows the cross-validation in identifying semiology from hands and fingers between patients with MTLE and ETLE where the test set obtained an average accuracy of 75% without segmentation and 82.7% with segmentation. This represents an improvement on the hand analysis in extracting features from the semantic segmentation compared with the strategy of extracting spatial features directly from the bounding box of the hand detected. Qualitative results of the semantic segmentation in a selected sample is depicted in Figure 5.33.

It is important to note, that the performance of both approaches (with or without segmentation) is affected significantly in images with incomplete hands because of occlusions caused by bedding, motion discontinuities and camera angle. Furthermore, although the segmentation approach showed impressive results, the features extracted from the detected hand were inaccurate in scenarios of total overlapping of hands, which is common in clinical manifestations.

Seizure classification (LOSO-CV)

The proposed hierarchical multi-modal approach aims to verify spatio-temporal representation of the patient's behaviour in order to discriminate each epilepsy type. Although this approach considers local features through the proposed architecture, it is not possible to extract or determine which areas and movements of the face/body/hands contribute towards individual classification decisions, *i.e.*, the system does not discriminate nor determine whether particular clinical signs are better predictors than others. Local actions cannot be classified because of the lack of labels for them. For example, categorise between dystonic limb posturing and tonic limb posturing.

All quantitative features described for each semiology representation and the two different seizure pattern classes (temporal and extra-temporal) were compared in order to evaluate the performance in each patient.

The LOSO-CV performance of each patient and the present ictal sign are depicted in Figure 5.34. Patients with MTLE are represented in blue bars, while patients with ETLE are shown using cyan bars. The hierarchical multi-modal system reached an average accuracy (Ave. Acc.) of 53.39%, 56.31% and 55.10% for face, body and hands, respectively. Figure 5.35 illustrates vertical boxplots that represent the extent of the classification accuracies using LOSO-CV over each semiology. It can be observed that the identification performance ranges from 12% to 83.4% for the face, from

Figure 5.32 Qualitative results of the hand detection in the epilepsy dataset. The hands detected are shown in the yellow bounding boxes.

Figure 5.33 Qualitative results of the hand detection with semantic segmentation in the epilepsy dataset. Left: Hand bounding box detection. Right: Mask segmentation and bounding box identification.

41.2% to 80.1% for the body pose and from 32.8% to 69.3% for the hand analysis. This marked variability between patients reflects challenges related to the heterogeneity and variable frequency of semiological features within the data. For example, some patients exhibit semiology entirely separate for all others and so are poorly classified.

The obtained results using the epilepsy dataset showed that MTLE facial features were significantly more frequent and discriminative than ETLE (Average accuracy 64.11% vs 42.66%). Significant differences were also found from the head and upper limbs movements, indicating that ETLE presented more consistent characteristics of this semiology (Average accuracy 47.33% vs 65.28%). Hand motion

Figure 5.34 LOSO-CV performance for patients with MTLE and ETLE with the hierarchical approach. The *x*-axis represents the label of each patient that was described in Table 5.8. Patients with MTLE are represented with blue bars and patients with ETLE with cian bars. The average accuracy of all patients of each class is illustrated with the red line.

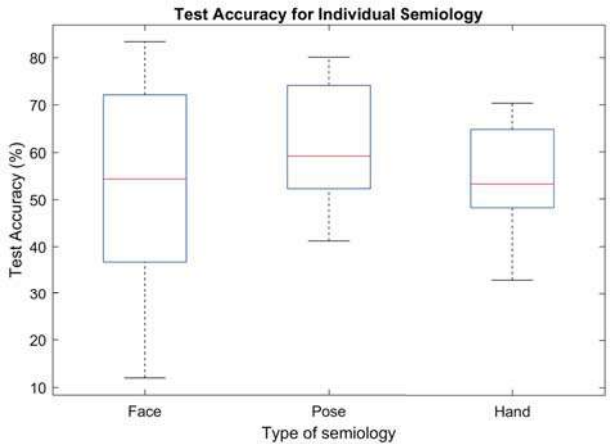


Figure 5.35 LOSO-CV performance for individual semiology with the hierarchical approach. The identification performance of patients with MTLE and ETLE are combined for each type of semiology.

also separated both classes, but there are not sufficiently discriminative features to train the system (Average accuracy 55.69% vs 54.52%).

It is important to note the difference in classification performance between patients. This demonstrates the wide variety in semiological patterns that are present in the database, and how some semiological patterns presented by some patients (*i.e.*, Patient 21, for who we obtain a Face Ave. Acc. of 12%) are dissimilar to those of all other patients present. However for other patients, such as Patient 4, 10 and 12, high classification performance is achieved (Face Ave. Acc. of 83.4%, 80.6%, and 82.1%, respectively), as they exhibit semiological patterns that are consistent with other patients in the dataset. This demonstrates the primary weakness of the proposed approach: it is sensitive to the

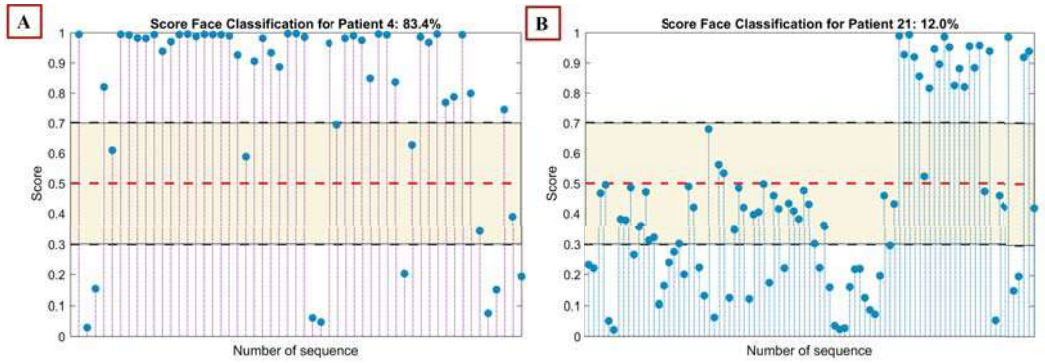


Figure 5.36 Visualisation of the classification scores in a selected window of sequences of two patients. Classification areas are defined as confident (0-0.3 and 0.7-1) and uncertain (0.3-0.7). **A.** Score distribution for facial semiology classification for Patient 4. **B.** Score distribution for facial semiology classification for Patient 21.

amount of data and unusual semiological characteristics present in the data. However, the encouraging performance on patients for who similar semiology exists demonstrates the potential of this proposed approach, as it is reasonable to expect performance to improve as more data that better captures the variety of semiological patterns is incorporated.

It is also important to discuss that the classification system has a certain level of uncertainty, as it is possible to see in the score values over an entire sequence illustrated in Figure 5.36. This allows us to better reflect the classifier's confidence and to better interpret the "correct" or "incorrect" classifier decisions. The system considers a sequence to be correctly classified when its probability is greater than 0.5, but when the system classifies the sequence incorrectly, it does not mean that for a particular sequence of a patient, the system is confident. For this reason, in order to visualise the uncertainty of classification between patients, we have defined two boundaries: the positive confidence area (score range from 0-0.3 and 0.7-1) and the uncertainty area (score range from 0.3-0.7). These boundary levels were defined based on visual inspection of the results, where the main aim was to identify when the classifiers were very uncertain. In Figure 5.36A, it is possible to note that the classification for Patient 4 (Acc. 83.4%) shows more sequences classified with positive confidence (94% of the sequences), while in Figure 5.36B for Patient 21 (Acc. 83.4%) almost 41% of the sequences were classified in the uncertainty area. Although our system is still attempting to make a hard decision based on "right" or "wrong" classification, we argue that an effort to include a quality assessment of the classifier confidence value should be considered based on a larger seizure semiology database. The primary objective of this proposal is to provide more robust information to support clinical evaluation.

Table 5.12 displays the multi-modal performance of the framework using test patients of each class that were not included in the dataset described in Table 5.8. For these patients with epilepsy, the clinical manifestation was unknown for the system, *i.e.*, the system automatically performs the facial, body and hands analysis simultaneously without knowing the type of semiology that was included in the test set. In this situation, movements from the face could be identified that differentiated better

Table 5.12 Hierarchical multimodal performance for semiology analysis on unseen patients.

Test Set	# Patients	# Seizures	Face Accuracy (%)	Body Accuracy (%)	Hand Accuracy (%)
MTLE	2	3	**62.30%**	44.30%	**55.8%**
ETLE	1	2	43.80%	**71.20%**	N/A
Average	-	-	53.05%	57.75%	55.8%

patients with MTLE (62.30%), in contrast to motions from the head and upper limbs that showed the best accuracy performance in patients with ETLE (71.20%). In order to evaluate hand motions from the video clips with ETLE, the extraction of features was not possible because of occlusions from the bed sheets.

5.4 Discussion and limitations

Seizure semiology is a very useful tool; however, it requires standardisation among evaluators. Our purpose is to demonstrate that automated analysis of seizures with similar semiological patterns from different patients could involve neuronal activities within the same specific brain networks, and could be sufficiently reliable to categorise patients with a specific type of epilepsy.

In this chapter, we show the plausibility of developing approaches using high-performance computer vision to determine if the analysis of the face, head and upper limbs, and hands changes can distinguish between patients with MTLE and ETLE.

The state-of-the-art technology has shown that it is possible to apply automated video analysis to extract features for a patient's clinical manifestations and recognise various kinematic patterns related to epileptic seizures. However, the major limitation of current approaches is that they address only a limited set of epileptic phenomena because of the difficulty in extracting robust features in the clinical environment, even with new video detection devices. Furthermore, current automated systems for seizure quantification are constrained to measuring seizures that involve limb and head movements.

The results reported here in this chapter, constitute the first multi-modal implementation of ictal signs based on deep learning architectures in a real hospital environment, in the context of the assessment of patients with epilepsy. The proposed approaches demonstrate the capability of deep learning of learning across datasets. Although the fusion approach demonstrated that it is an option to improve the classification accuracy over isolated semiology, this strategy is sometimes difficult to accomplish because not all patients experience face and body semiology during all seizures; and because for some videos, it is possible to extract only facial features due to occlusions from objects such as blankets. The hierarchical approach, on the other hand, provides individual classification accuracy according to the semiology under evaluation, which in this scenario also incorporates ictal symptoms from the hands and fingers motions.

Our study shows that a multi-modal methodology is viable for analysing 2D monitoring videos using the existing technology in the hospital with challenging imaging conditions typical of an epilepsy monitoring unit, and is capable of detecting and tracking human behaviour using supervised

deep learning architectures. Our approach has demonstrated robustness to model variations in the data (k-fold cross-validation), but also highlights the challenge posed by the high variation in the data when classifying semiology of particular patients (LOSO-CV). It is likely that the ictal patterns from patients with MTLE and ETLE were not strongly present or similar to the semiological patterns of the dataset. This is evident from the difference in performance between patients as illustrated Figure 5.34. Although we can expect that a larger seizure semiology database may be used to mitigate the accuracy problems we report in this chapter, we are also interested in developing an automated detection system for unusual or aberrant behaviours. This approach could support each identification system with an active learning phase that includes new semiological features detected by the system. Aberrant detection methods can be very useful in identifying interesting, concerning, or unknown events using past patient cases stored in health records. This strategy will be described in Section 7.2.

Undoubtedly, our approach highlights an important limitation of the system to evaluate specific semiology. For instance, the system cannot determine if the most discriminative facial features to distinguish between seizures are related to facial modifications from the eyes or mouth automatisms. For this reason, we also propose methodologies that are capable of detecting and tracking specific clinical signs, and demonstrate the discriminative power of individual motions, as it will be discussed in Chapter 6 for the 3D mouth semiology analysis.

Although deep learning approaches often require extraordinary computational resources, the most expensive phase is the training, which only needs to be performed once on high-performance computing infrastructure. The network is fast at test time when classifying new data. For example, to run one video clip of approximately 2 min or 3,000 frames takes 5 min and 2GB of RAM, which includes the preprocessing of the images and the feature extraction of each ictal sign. Once the model has been trained, it can run smoothly on low-end devices such as portable devices or embedded devices.

A direct comparison with the state-of-the-art in seizure quantification, specifically for the head and upper limbs motion, is difficult because researchers use their own datasets and distribution of this data is restricted by legal and ethical considerations. There is still the need for standardisation of the hardware and software solutions which will happen as more research groups embrace computer vision approaches for human motion analysis in epilepsy.

While the proposed multi-modal semiology system is still far from reaching the ideal performance expected by clinical experts and patients, it is, however, a critical step in that direction. In time, and with further data and resources, it has the potential to form a key tool for clinicians, providing decision making support in epilepsy centres, likely through off-line review of recorded footage when assessing patients. Our proposed system is a modular and flexible architecture that allows us to improve the performance of the semiology analysis by replacing each module of human motion quantification with new, more accurate and robust computer vision approaches. Table 5.13 summarises a number of more recent approaches that have only been introduced after our work in this research. These approaches, when plugged into our proposed modular system, are expected to improve the overall results of analysing semiology.

Table 5.13 Summary of deep learning architectures that may improve the quantification of semiology.

Clinical manifestation	**New approaches**
Facial Semiology	
Face detection	Please refers to Section 4.5.
Face tracking	A Prior-Less Method for Multi-Face Tracking in Unconstrained Videos [Lin and Hung, 2018]. Self-supervised learning of face representations for video face clustering [Sharma et al., 2019]
Head / upper limbs semiology	
2D pose estimation in videos	Efficient detection and tracking in Videos [Girdhar et al., 2018].
2D-3D pose estimation	A dual-source approach for 3D human pose estimation from single images [Iqbal et al., 2018a]. Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB [Mehta et al., 2018]. (3D pose estimation from one single image that does not require correspondences between 2D and 3D points or 3D pose libraries). Can 3D Pose be Learned from 2D Projections Alone? [Drover et al., 2018]. 3D human pose estimation in the wild by adversarial learning [Yang et al., 2018b] (Approaches that use adversarial frameworks to impose a prior on the 3D structure, learned solely from their random 2D projections or to generate plausible poses even on unannotated in-the-wild data.
Dense pose estimation	End-to-end recovery of human shape and pose [Kanazawa et al., 2018]. Densepose: Dense human pose estimation in the wild [Güler et al., 2018, Güler et al., 2017]. (Dense approaches enable more robust analysis by providing more information and not just selected keypoints in the body).
Hands / fingers semiology	
3D hand landmarks	GANerated Hands for Real-time 3D Hand Tracking from Monocular RGB [Mueller et al., 2018]. Hand Pose Estimation via Latent 2.5 D Heatmap Regression [Iqbal et al., 2018b] Weakly-supervised 3D hand pose estimation from monocular RGB images [Cai et al., 2018] End-to-end Hand Mesh Recovery from a Monocular RGB Image [Zhang et al., 2019]
Multi-modal detection	Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies, which is an Adam model obtained through controlled 3D surface acquisition [Joo et al., 2018].

In the following chapter, we introduce a novel framework to quantify and classify mouth semiology based on detailed information from 3D face reconstruction in the epilepsy scenario. We aim to address the limitation of the facial region-based approach discussed in this chapter in distinguishing which features from the bounding box are related to mouth motions. Mouth motions are heavily examined by epileptologists to distinguish seizure types, but current architectures are unable to quantified properly this isolated clinical sign.

Chapter 6

3D Mouth analysis in epilepsy

6.1 Overview

The methodology proposed in Chapter 4 and 5 demonstrated the potential of deep learning architectures to capture and quantify facial semiology during epileptic seizures. Nevertheless, these techniques are unable to accurately assess individual behaviours of interest such as mouth semiology, which is heavily examined by epileptologists to distinguish between seizure types [Aupy et al., 2018]. Clinical experts rely on certain mouth motions such as chewing automatisms, unilateral mouth deviation, smacking and grimacing [Chauvel and McGonigal, 2014], to assess where the epilepsy may be arising from and thus enable the correct procedures to be administered for the patient. Focusing on one clinical sign is a necessity in order to reduce the problem dimensionality, as more than 40 possible facial descriptions can be linked to epilepsy [Noachtar and Peters, 2009]. For example, ictal pouting which mimics expressions of fear, displeasure, or disgust, has shown to be a reliable indicator of seizures arising from the prefrontal-insula region when associated with intense emotional changes and hypermotor behaviour [Souirti et al., 2014], and the occurrence of mouth chewing (an oroalimentary automatism) is related to a theta discharge in the opercular region [Aupy et al., 2018]. Our aim is to automatically detect these facial modifications from video feeds and this chapter proposes a robust technique suitable for the challenging imaging conditions of clinical monitoring. This chapter's research aim and its relationship with the thesis is illustrated in Figure 6.1.

The automated analysis of mouth semiology is still largely unexplored due to the immense complexity in detecting and tracking key facial regions, and the challenging conditions posed by the healthcare environments and/or patient positioning [Sathyanarayana et al., 2018, Thevenot et al., 2017]. Conventional 2D region-based techniques based on the detection and quantification of the mouth as an image [Pediaditis et al., 2011, Pediaditis et al., 2012a] have limitations in handling large pose variations, and thus make a fair comparison between samples difficult due to the variety of poses present. Additionally, in these proposals, the detection of the mouth is reliant on detection methods that are only suitable for frontal and well-lit face images, such as the Viola-Jones algorithm [Viola and Jones, 2001]. Recent breakthroughs in computer vision and

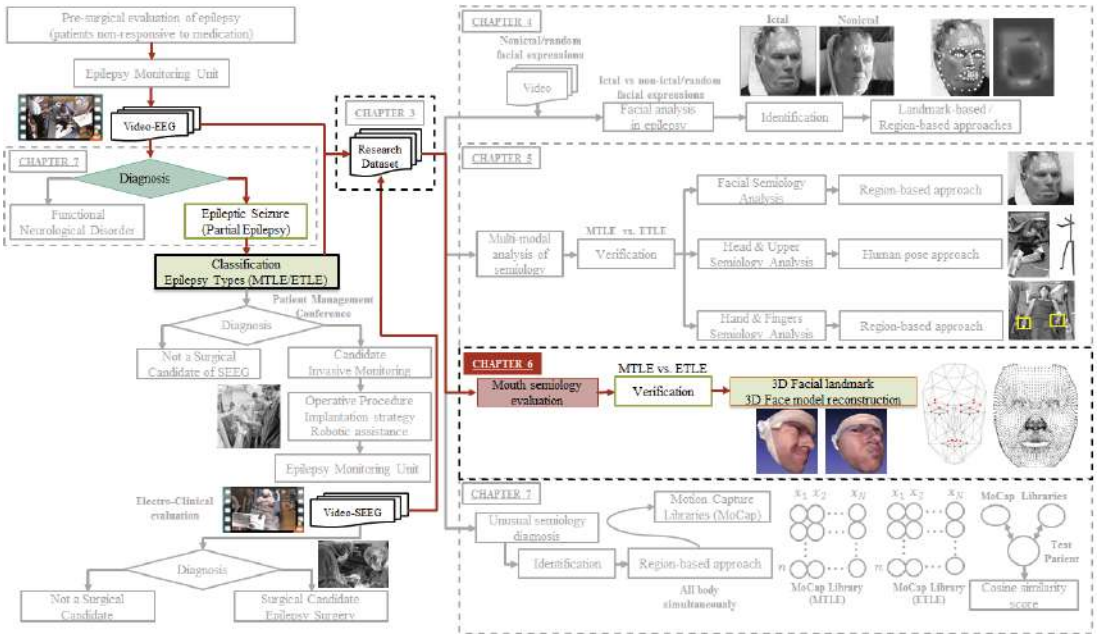
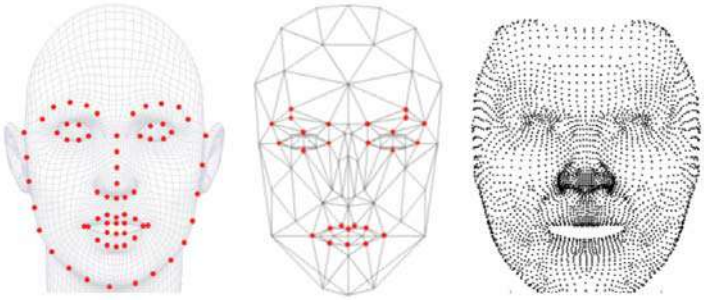


Figure 6.1 Representation of the chapters and their interconnections in the clinical evaluation of epileptic patients. Overview of the research aim in Chapter 6.

deep learning offer a new and exciting direction to overcome these limitations; however, most existing approaches to assess facial symptoms eschew deep features [Thevenot et al., 2017]. Approaches such as facial landmarks have been proposed to quantify the mouth motion (See Section 4.3.2), however, they are unable to fully represent motions in the mouth and cheeks (*e.g.*, ictal pouting) and are affected by non-visible landmarks. Despite region-based success in extracting features from the face-bounding box detected for medical diagnosis [Guo et al., 2017, Rodriguez et al., 2017] and for facial semiology assessment (see Section 5.3.1), this approach is unable to determine which motion(s) contribute towards a given decision. In the epilepsy setting for example, our approach cannot determine if facial features which distinguish between seizures are related to a motion from the eyes or mouth. 3D approaches such as face fitting, on the other hand, retain rich information about the shape and appearance of faces, simplifying alignment for comparison between image sequences. However, there is insufficient research documented using 3D face models in epilepsy to conduct a deep analysis of mouth motions. In this chapter, we propose a novel framework based on a 3D reconstruction of the face and deep learning techniques to detect and quantify mouth semiology.

The chapter is distributed as follows. Section 6.2 describes the strategy to distinguish mouth semiology between epilepsies base on 3D models; Section 6.3 describes the dataset, the experimental setup and the results on quantifying the isolated mouth semiology; Finally, Section 6.4 summarises the discussion of the proposed architecture, limitations and possible future directions.

This chapter is supported by the following accepted manuscript:



(a) From left to right: Traditional techniques (landmarks and triangulation) and dense representation.

(b) Selected samples that represent complex mouth semiology from the epilepsy dataset.

Figure 6.2 Traditional and dense representation of facial expressions. Selected sequences in the epilepsy dataset with mouth semiology which are challenging to model with traditional techniques.

- **D. Ahmedt-Aristizabal**, K. Nguyen, S. Denman, M. Saquib Sarfraz, S. Sridharan, S. Dionisio, C. Fookes, Vision-Based Mouth Motion Analysis in Epilepsy: A 3D Perspective, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2019.

6.2 Strategy to assess mouth semiology

Extracting dense facial shapes from monocular images can have substantial benefits, as it provides detailed information on facial expressions and location of facial parts. Dense representation, compared to simplified approaches such as landmarks and triangulations, retains rich information about the shape and appearance of faces, as illustrated in Figure 6.2(a). This approach has only recently become possible due to breakthroughs in deep learning which have enabled the use of very complex models. 3D dense facial reconstruction allows the reproduction of a variety of facial expressions through textures that are challenging to capture with traditional 2D landmark-based methods [Zhang et al., 2016c] (as discussed in Section 4.3.2), triangulations (candide models) [Maurel et al., 2008], or with global and local features from 3D cameras [Dittmar et al., 2017]. Several samples of these complex facial expressions are depicted in Figure 2.7(a) and Figure 6.2(b). 2D keypoint-based approaches are unable to accurately quantify motions in the face such as lip smacking, grimacing and downward mouth displacement (chapeau de gendarme) during seizures [Souirti et al., 2014], and the analysis is affected by non-visible landmarks. Triangulation is simplistic, is unable to fully represent the facial expressions, and does not handle large changes in head orientations [Salam and Séguier, 2018].

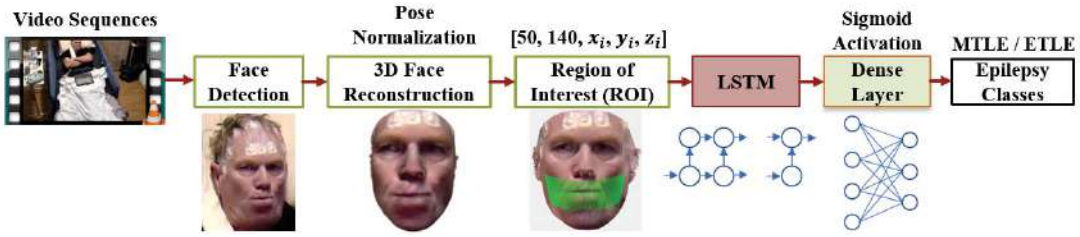


Figure 6.3 Framework proposed to capture and quantify mouth motions for seizure classification. Firstly, the face is detected from the video sequences. Secondly, the 3D face is reconstructed to capture the motions of the facial muscles located in the mouth and cheeks (pose-invariant representation). From the resultant 3D mesh of the face, we defined a ROI of size $[50, 140]$ (illustrated in green in the image). The x_i, y_i, z_i representation corresponds to the 3D location of each dense point. These spatial features are fed to an LSTM to extract temporal relations between sequences and for classification of each type of seizure.

Ideally, face models should embed a set of facial muscles that produce expressions by deforming the model in the physical realm.

We conduct investigations to ascertain if the recent advances in 3D face modelling could be a solution to capture and analyse the diverse types of mouth motions exhibited during epileptic seizures compared to purely 2D image-based approaches. In particular, this research aims to compare the patterns of semiology from complex mouth motions in patients with mesial temporal and extra-temporal lobe epilepsy. To achieve this, we propose a framework, depicted in Figure 6.3, which receives the location of the patient's face from a video sequence. Then, a 3D face reconstruction model is adopted to extract a 3D face shape from a 2D image to capture the range of shapes and deformations of the moving cheeks and muscles in the face including the anatomical points around the mouth. Later, we adopt a 3D dense point array from a selected region of interest (ROI) to track changes in the mouth area. Finally, the spatial representation of the mouth is captured via an LSTM network to encode the temporal information presented in the sequence, and to estimate the seizure type of each sequence.

Figure 6.4 Selected samples of qualitative results of the face detector in the WIDER Face dataset. We visualise one example for each attribute and scale. Image adapted from [Hu and Ramanan, 2017].

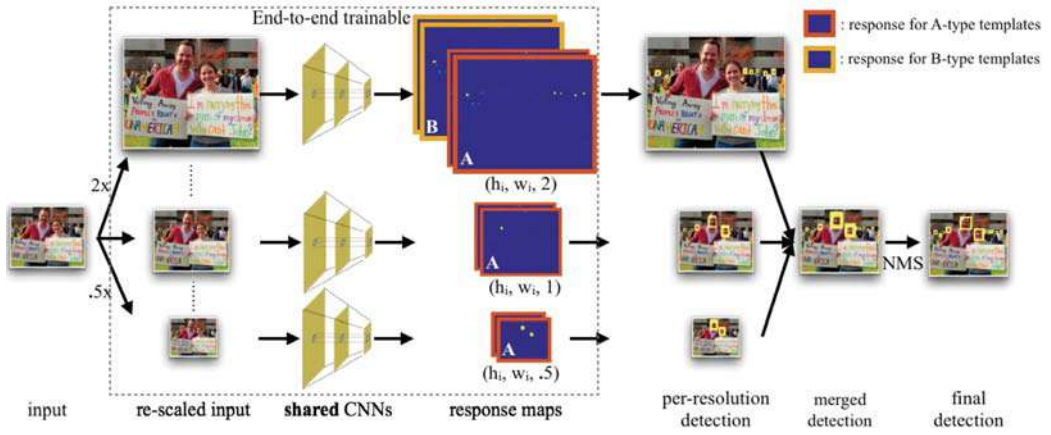


Figure 6.5 Overview of the scale aware face detection pipeline. Starting with an input image, a coarse image pyramid (including 2X interpolation) is created. Then the scaled input is fed into a CNN to predict template responses (for both detection and regression) at every resolution. In the end, a non-maximum suppression (NMS) at the original resolution to get the final detection results is applied. The dotted box represents the end-to-end trainable part. Image adapted from [Hu and Ramanan, 2017].

6.2.1 Face detection and tracking

Due to the power of deep learning, the accuracy of face detection has been improved by the use of CNNs as discussed in Section 5.3.1. However, the fine-tuned face detector based on [Jiang and Learned-Miller, 2017] still struggles in dealing with scale variations, resolution and complex head rotations, *i.e.*, the model is not rotationally invariant and spatially aware. To address this challenge, we adopt the implementation of [Hu and Ramanan, 2017] available in [Hu, 2018] to detect the patient's face during epileptic seizures. This approach can perform face detection in challenging scenarios, and better deals with scale variation in the benchmark face datasets FDDB and WIDER Face, outperforming the previous results documented in Chapter 5. Figure 6.4 depicts qualitative results in a benchmark dataset.

Figure 6.5 depicts the end-to-end trainable model known as tiny faces algorithm. This model which uses a backbone architecture based on ResNetXt with a depth of 101 layers [Xie et al., 2017], first creates a coarse image pyramid with the input image and $2X$ interpolation. Then, shared CNNs predict template responses (for both detection and regression) at every resolution. Finally, the model uses non-maximum suppression (NMS) at the original resolution to get the final detection results. To improve the conditions where faces appear in a wide range of poses, the model was fine-tuned on the public dataset FERA 2017 [Valstar et al., 2017], which includes facial expressions in a wide range of head orientations and is derived from the BP4D-Spontaneous database (see Table 3.6).

To consistently localise the face bounding box in terms of size and position with minimal jitter between frames, we fuse the face detector with a tracking algorithm for videos sequences based on the open source SORT tracker [Bewley et al., 2016]. This tracker achieves a high level of performance efficiently with a combination of traditional techniques such as the Kalman Filter and

Figure 6.6 Selected sequences of facial semiology captured with the face detection and tracking architecture.

Table 6.1 Selected benchmarking techniques for 3D face model reconstruction, where some of them are also used to detect 3D facial landmark simultaneously.

Author	Title / Name	3D Landmarks	3D Shape
[Feng et al., 2018a]	Position map regression network (PRN)	✓	✓
[Tran et al., 2018]	Extreme 3D face reconstruction		✓
[Jackson et al., 2017]	Volumetric regression network (VRN)		✓
[Sela et al., 2017]	Unrestricted Facial Geometry Reconstruction		✓
[Tran et al., 2017]	Regressing 3DMM with deep neural networks	✓	✓
[Dou et al., 2017]	End-to-end 3D face reconstruction (UH-E2FAR)		✓
[Bhagavatula et al., 2017]	3D spatial transformer network (3DSTN)	✓	✓
[Liu et al., 2017]	Dense face alignment (DeFA)	✓	✓
[Güler et al., 2017]	Dense Shape Regression (DenseReg)		✓
[Zhu et al., 2016]	Face alignment across large posess: a 3D solution (3DDFA)	✓	✓
[Jeni et al., 2015]	Dense 3D face alignment	✓	✓
[Qu et al., 2015]	Adaptative contour fitting for 3D face shape		✓
[Hassner, 2013]	Viewing real-world faces in 3D (Flow-based)		✓

Hungarian algorithm for the tracking components. We adopt the proposal in [Wojke et al., 2017], which extend the SORT algorithm by integrating appearance information using a deep descriptor. The implementation is on Python3 and TensorFlow [Abadi et al., 2016] available in [Wojke, 2018]. This strategy outperforms the face detection compared with the strategy proposed in Section 5.3.1. The detected facial bounding box in each frame is used to crop the face. Figure 6.6 illustrates sequences of facial semiology detected and cropped using this approach.

6.2.2 3D face reconstruction and region of interest definition

To achieve pose and illumination invariance, 3D information of the face is useful [Paysan et al., 2009] and the 3D surface of a face is known to be discriminative [Tran et al., 2017]. A major benefit of performing 3D face reconstruction is simplifying alignment or extracting a pose-normalised version of the input image for a fair comparison between samples [Egger et al., 2014]. However, classic 3D methods are rarely used for in real-world settings, as in such unconstrained situations methods can be unstable and yield poor representations as they are over-regularised and generic, and due to difficulty in regressing to 3D face models [Tran et al., 2017]. The most relevant approaches to obtain 3D face model reconstruction are listed in Table 6.1.

Since the proposal of the 3DMM [Blanz and Vetter, 1999], traditional statistical shape representations have used aligned 3D face shapes to learn a distribution of 3D faces [Chu et al., 2014], and are heavily reliant on the accuracy of feature point detectors [Huber et al., 2016, Jeni et al., 2015].

Figure 6.7 Selected samples of qualitative results of the 3D face reconstruction approach in the AFLW2000-3D dataset. Landmark estimation (Upper) and reconstructed shapes on the original image (Lower). Image adapted from [Feng et al., 2018a].

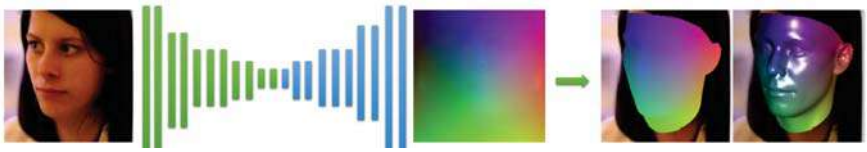
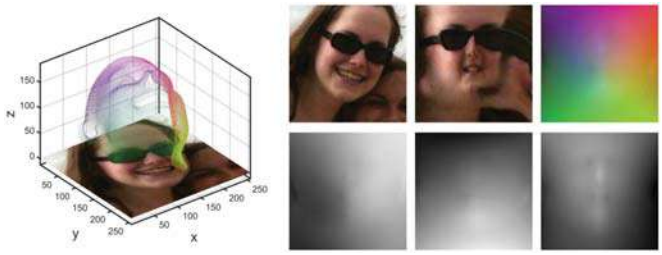


Figure 6.8 Schematic of the 3D face reconstruction model. Residual blocks and transposed convolutional layers are represented in green and blue rectangles, respectively. Image adapted from [Feng et al., 2018a].

Some methods use CNNs to learn the dense correspondence between image and 3D template to calculate the 3DMM parameters with dense constraints [Güler et al., 2017].

Landmark fitting methods [Zhu et al., 2016, Jourabloo and Liu, 2016], which use cascade CNN structure to regress the 3DMM coefficients, focus more on landmark detection, and so do not produce detailed and discriminative facial geometries (see Section 4.3.2). Other approaches have used end-to-end CNNs to regress 3DMM parameters directly from an input image [Tran et al., 2018, Dou et al., 2017, Tran et al., 2017, Tewari et al., 2017]; however, they are still model-based resulting in a limited geometry which is constrained to generate a 3D mesh from estimated parameters. Some other methods reconstruct 3D faces without 3D shape bases by warping the shape of a reference 3D model, but the structure differs when the template changes [Bhagavatula et al., 2017, Sela et al., 2017]. A recent approach is not restricted in the model space by mapping the image pixels to the full 3D facial structure via volumetric CNN regression [Jackson et al., 2017]. However, a lot of time to predict the data and a complex network structure is needed.

Based on model-free and light-weight frameworks, we adopt the state-of-the-art approach known as PRNet [Feng et al., 2018a] available in [Feng et al., 2018b] for the purpose of 3D face reconstruction due to its reported high level of performance. The reconstruction results are evaluated qualitatively by comparing the visual quality of the generated models, where the texture and the mesh preserve all details of a face's appearance, and there are no constraints caused by model parametrization. Qualitative results in a public dataset are depicted in Figure 6.7. This framework, illustrated in Figure 6.8, reconstructs a 3D face shape based on an end-to-end position map regression network, which is trained with the 300W-LP-3D dataset (see Table 3.5). Feng et



(a) Representation of the UV position map. Left: 3D plot of the input image. Right: input 2D image, extracted UV texture map, corresponding UV position map and the x, y, z channel of the UV position map.

(b) Representation of the weight mask. From left to right: UV texture map, UV position map, coloured texture map and the final weight mask.

Figure 6.9 Representation of the UV position map and weight mask. Images adapted from [Feng et al., 2018a].

al., [Feng et al., 2018a] proposed a UV position map as the representation of a full 3D facial structure, as it is illustrated in Figure 6.9(a). This position map is a 2D image which records the 3D coordinates in UV space and provides dense correspondence to the semantic meaning of each point on UV space. Based on previous works which use UV space to express the texture of faces [Bas et al., 2017], the PRN network uses the UV space to store the 3D coordinates of points from 3D face model. Therefore, the UV coordinate system, which represents a full 3D facial structure, is created based on the 3DMM, but the regression model is not constrained by the original 3DMM.

The encoder-decoder network focusses on the discriminative region to regress the UV position map from a single 2D facial image. The parameters of the network are learned using a loss function between the regressed position map and the ground truth position map based on a weight mask. This mask, depicted in Figure 6.9(b), records the weight of each point on the position map, where some regions of the face contain more discriminative features, *e.g.*, the position of the 68 landmarks has the highest weight. This strategy enhances learning in the centre of the face and neglects the impact of regions such as the neck and body. In this experiment, the weights ratio are 16, 4, 3 and 0 for subregion 1 (68 facial landmarks), subregion 2 (eye, nose, mouth), subregion 3 (other face area) and subregion 4 (neck), respectively. Given a predicted position map as $P(x,y)$ for x,y representing each pixel coordinate and the ground truth position map $\check{P}(x,y)$ and weight mask $W(x,y)$, the loss function is defined as,

$$Loss = \sum \|P(x,y) - \check{P}(x,y)\| \cdot W(x,y). \tag{6.1}$$

From the estimated facial mesh (a point cloud of size 42,867 vertices and its respective RGB colour), we extract a ROI located around the mouth and cheeks with a dimension of $[50, 140]$. The

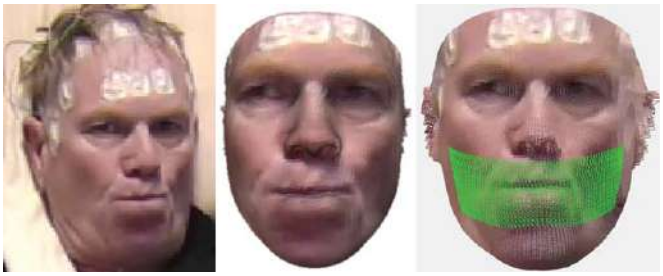


Figure 6.10 Region of interest defined from the 3D face reconstruction model in the epilepsy dataset. From left to right: original image, pose-invariant image from the 3D model, and the representation of the region of interest in green that captures the mouth semiology.

size of this ROI was defined based on visual inspection of the facial reconstruction in multiple pose-invariant images, where the main aim was to capture within the ROI the variety of motions from the mouth and cheeks. A sample of the ROI is displayed in Figure 6.10. The location of this 3D area is fixed for all frames in each sequence. As such, each *mouth model* has a spatial representation for each sequence of $[50, 140, x_i, y_i, z_i]$, where x_i, y_i, z_i are the 3D coordinates.

6.2.3 Temporal information and training of LSTMs

We capture dynamic variations of the mouth over a sequence by feeding the representations extracted from the 3D ROI to an LSTM network, which is able to learn long-term dependencies present in the sequential data [Greff et al., 2017].

The number of LSTM layers is one significant hyper-parameter to consider. We experiment with different numbers of layers and memory cells based on the feature dimensionality, and we choose the best configuration with two stacked LSTM layers, each with 128 memory cells. More complex architectures do not show significant performance gains. The outputs of the LSTM are concatenated into a single densely connected layer with a sigmoid activation function to make a single prediction for every sequence for each seizure.

As proposed in Section 4.3.3, we train the LSTM networks by optimising the binary cross entropy loss using the Adam optimiser [Kingma and Ba, 2014] with a learning factor of 10^{-3}, and the first- and second-moment decay rates of 0.9 and 0.999, respectively. We adopt a batch size of 16 and dropout with a probability of 0.35. We perform the training using 100 epochs with the default initialisation parameters (the weights of the LSTM hidden units) from Keras [Chollet, 2015].

6.3 Experimental results

6.3.1 Dataset specification

To quantify mouth semiology, the inputs of the system are short video sequences rather than a whole video, such that we obtain more data to train the system. We define a sequence as 5 consecutive

frames. For this experiment, we select seizure recordings with complex mouth modification (such as ictal pouting) in order to verify the robustness of the proposed system compared to traditional approaches. A total of 20 video clips were selected with the isolated mouth semiology, consisting of 10 videos from 2 patients with MTLE and 10 videos from 3 patients with ETLE (see Table 3.2).

6.3.2 Experimental setup

We adopt a leave-one-seizure-out cross-validation (LOZO-CV) scheme in order to validate the flexibility of the system to capture, quantify and model the variations between the two types of seizure. All sequences for a given seizure are held out as the test set, and the remaining seizures are used for training. The performance of each epilepsy type (MTLE and ETLE) is calculated as the average test accuracy of all seizures for each type of epilepsy.

To demonstrate the potential of our 3D approach, we also conduct a preliminary LOSO-CV experiment, which evaluates the entire video corpus for a test patient, who is totally excluded from the training data. The performance is calculated as the average test accuracy of all patients for each type of seizure. However, this evaluation is limited due to the limited data, and will be more informative as we incorporate more participants who exhibit complex mouth semiology.

We compare the performance of the proposed 3D model with two alternative approaches that extract spatial features coupled with the same LSTM structure. These approaches are as follows:

- We detect 2D landmarks around the mouth for the quantification of mouth motions which are represented by 20 fiducial points. The spatial representation of these points $[x_i, y_i]$ are fed to the LSTM. This approach is based on the framework proposed in [Zhang et al., 2016c] (see Section 4.3.2).
- We implement a 3D landmark detection method which contains a wide range of views using a 3D model and has been shown to outperform equivalent 2D approaches [Jourabloo and Liu, 2016]. We adopt the framework of [Bulat and Tzimiropoulos, 2017b], which is a state-of-the-art 3D landmark estimation system (see Section 4.3.2). Similarly, the 3D locations of the 20 landmarks from the mouth with a representation of $[x_i, y_i, z_i]$ (the corresponding 3D locations) are the input of the LSTM.

The approach documented in [Pediaditis et al., 2011, Pediaditis et al., 2012a], which detects the mouth and extracts time-varying features (averaging background and dense optical flow) failed on the epilepsy dataset because the strategy employed by [Pediaditis et al., 2011, Pediaditis et al., 2012a] to detect the ROI using cascade algorithms performs poorly with the unconstrained head position in the dataset used in this work as discussed in Section 4.3.1.

6.3.3 Seizure classification of mouth semiology

The proposed approach for face detection coupled with the tracking approach reached an average accuracy of 0.972 for the IoU in selected manually annotated videos. This represents an improvement

Table 6.2 Identification performance of mouth semiology.

LOZO-CV performance (Test Accuracy)			
Approach	**MTLE**	**ETLE**	**Average**
Number of Seizures	10	10	
2D landmarks + LSTM	36.8%	56.1%	46.4%
3D landmarks + LSTM	66.2%	82.1%	74.2%
3D Face (ROI) + LSTM	**84.8%**	**93.2%**	**89%**
LOSO-CV performance (Test Accuracy)			
	MTLE	**ETLE**	**Average**
Number of Patients	2	3	
2D landmarks + LSTM	20.2%	42.1%	31.15%
3D landmarks + LSTM	44.3%	71.4%	57.85%
3D Face (ROI) + LSTM	61.4%	78.2%	**69.8%**

in the face detection performance compared to region-based CNNs such as Faster R-CNN, which achieved 0.945 (see Section 5.3.3).

Each cross-validation performance scheme is reported in Table 6.2. The proposed framework was capable of achieving an average test accuracy of 89% in the LOZO-CV using the spatio-temporal information of the ROI extracted from the 3D face reconstruction. Additionally, the LOSO-CV using the 3D approach exhibited an average test accuracy of 69.8%. Figure 6.11 depicts qualitative examples of the 3D face reconstruction of images from seizures with mouth and cheeks motions, where the UV texture keeps microexpressions and shadows.

2D facial landmark detection via deep leaning was introduced for the analysis of facial expressions of patients with epilepsy in Section 4.3.2. However, this method only regresses visible points and has difficulties in dealing with large-pose changes ($\pm 90^{\circ}$) and occlusions [Zhu et al., 2016]. 3D facial landmark architectures have been proposed to handle face images with large pose variations. Nevertheless, it was observed that the methodology based only on 3D landmarks is inadequate to quantify some motions from the cheeks which are not captured by the landmarks, and are characteristics of patients with MTLE (*e.g.*, the ictal pouting semiology illustrated in the first two images of Figure 6.2(b)). Additionally, the landmark analysis is strongly affected by some landmarks that become invisible due to self-occlusion during large pose changes, where half of face is occluded. As a result, the landmark shape model struggles.

It is important to note that for both seizure types, the classification performance in the LOSO-CV improves when using the 3D face model. This demonstrates that the dense 3D face model, which already includes information on the 3D landmarks, captures more information about the changes in shape and appearance.

(a) Original image and the 3D reconstruction at different rotations.

(b) Qualitative results of the 3D face reconstruction in the epilepsy dataset. (Upper: face detected; Lower: 3D reconstruction).

Figure 6.11 Qualitative results of the 3D face reconstruction during mouth semiology.

6.4 Discussion and limitations

This chapter presents the first application of 3D face reconstruction for vision-based assessment of key isolated signs in epilepsy, such as mouth motions, which can be representative of true lateralising features to distinguish the affected brain regions. We have investigated the benefits of 3D face models and deep learning to quantify and classify seizures recorded from patients with epilepsy. Experimental results support that a 3D perspective provides additional capacity to model the variety of mouth motions that patients with MTLE and ETLE exhibit. The results demonstrate that our system is a promising option for assistive medical diagnosis from faces which can be extended to other vision-based clinical examinations, where currently automated approaches are adversely affected by challenging, unconstrained conditions of epilepsy monitoring.

With the reconstructed 3D dense model, hand-crafted features can also be estimated more accurately from the pose-invariant image than with traditional approaches [Egger et al., 2014], but this has not been the focus of this chapter. With the model's 3D perspective, we show the flexibility of the system in quantifying critical and unusual mouth semiology in natural clinical setting.

Although it is possible to implement region-based techniques such as DCNN based detectors (implementing and training state-of-the-art object detectors such as Mask-RCNN [He et al., 2017]) to detect and classify mouth motions by considering the extraction of features from mouth images [Pediaditis et al., 2011, Pediaditis et al., 2012a], we argue that this strategy has some limitations. The features extracted from the bounding box will not be valuable for performing a fair comparison between seizures due to the variety of head positions. A major benefit of performing 3D modelling is simplifying the alignment of faces for classification purposes. Additionally, detection-

based methods require a very large corpus of data of the specific mouth region to train a model, and the detection noise (jitter between frames) can affect the extraction of robust features.

These results have validated the feasibility and effectiveness of facial analysis of epilepsy using deep learning and 3D reconstruction. This opens up new opportunities in this research direction, such as investigating how to employ 3D features to determine the temporal evolution of this isolated semiology.

A clear limitation of each identification system proposed so far (face, multi-modal and mouth approaches) is their reliance on supervised learning making them unstable to potential unusual semiology. This is evident from the difference in classification performance in particular patients. For this reason, in the following chapter, we introduce a system capable of identifying aberrant epileptic seizures which alert clinicians to the occurrence of unusual events that deviate from a pre-learned database of known semiology. We aim to use this detection to perform active learning and progressively update the system with new semiologies. This avoids incorrect interpretation when classifying a new seizure using our identification systems. We are also interested to exploit the new methodology in analysing the whole body simultaneously to suggest the first marker-free system that differentiates epileptic seizures and functional neurological disorders.

Chapter 7

Identification of aberrant semiology and seizure disorders

7.1 Overview

In this chapter, we propose a new methodology capable of modelling clinical manifestations of epileptic seizures in real-life clinical settings aiming to analyse two specific and relevant diagnoses: 1) The identification of aberrant or unusual epileptic seizures; and 2) the identification of patients with functional neurological disorders and epileptic seizures, both diagnostic groups under the common label of seizure disorders in this chapter. These two diagnoses are discussed in the same chapter because they are addressed using the same region-based methodology which analyses all body motions simultaneously. We demonstrate the benefits and flexibility of our own design to assess complex and different situations that are heavily examined by neurologists. This chapter's research aim and its relationship with the thesis is illustrated in Figure 7.1.

Aberrant seizure identification: In the multi-modal strategy to classify epilepsies based on a hierarchical approach (See Section 5.3), we have validated the feasibility of quantifying clinical manifestation from patients with MTLE and ETLE. However, the classification performance is sensitive to potential aberrant or unusual semiology. We attempt to group epileptic seizures into a best-fit model, using a template of known semiology or stereotypical behaviours in the form of libraries. The libraries store feature representations of the motion exhibited by the patient during the recorded seizure. These motion libraries enable us to identify if the semiology from a new patient can fit into the status quo of the learned information, to conclude on having dissimilar findings or aberrant semiology. This process is depicted in Figure 7.2. Aberrant identification methods in epilepsy, unreported in the literature, can be very useful in identifying interesting, concerning, or unknown events. We show that this can be achieved using a pre-learned database of semiology stored in health records, motion capture libraries of spatiotemporal representations and similarities between hidden states. This system enables active learning to progressively update the system with new clinical manifestations detected within the system.

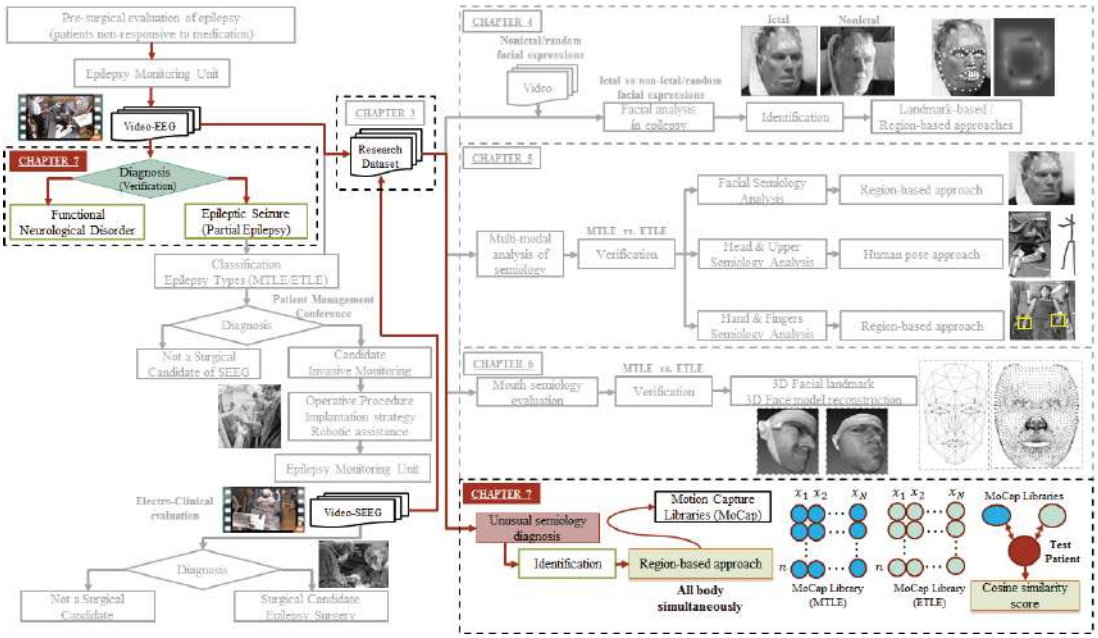


Figure 7.1 Representation of the chapters and their interconnections in the clinical evaluation of epileptic patients. Overview of the research aims in Chapter 7.

Seizure disorders identification: Another important consideration in the assessment of patients is the identification of psychogenic nonepileptic seizures or functional neurological disorders (FNDs), which mainly are manifestations of psychological distress [Benbadis and Hauser, 2000]. Functional neurological disorders are paroxysmal behaviours that simulate epileptic seizures; however, FNDs are not associated with the changes in cortical activity that characterise epilepsy [Sirven and Glosser, 1998]. A substantial proportion of patients with FND are incorrectly diagnosed with epilepsy due to overlaying clinical features during a seizure [Sirven and Glosser, 1998]. This makes the task of differentiating between patients with FND and patients with epilepsy extremely challenging. Misdiagnosis may lead to unnecessary treatment and its associated complications. Existing sensor-based and marker-based systems require physical contact with the body and are vulnerable to clinical situations such as patient position, illumination changes and motion discontinuities [Pediaditis et al., 2012b]. Marker-free systems based on computer vision and deep learning are advancing to overcome these limitations; however, there does not exist any work determining if a seizure is the result of epilepsy in the literature. To deal with this, we propose and compare two marker-free deep learning models, a landmark-based and a region-based model, both of which are capable of distinguishing between seizures. We quantify semiology by using either a fusion of reference points and flow fields, or through the complete analysis of the body.

The chapter is distributed as follows. Section 7.2 describes the aberrant epileptic seizure identification approach. Section 7.3 provides a quantitative vision-based strategy to assess epileptic seizures and functional neurological disorders.

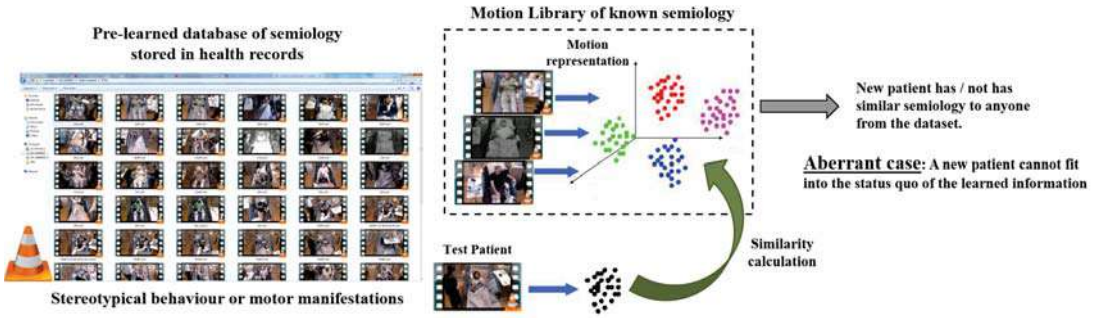


Figure 7.2 Representation of motion libraries to detect aberrant semiology. Pre-learned database of semiology stored in health records as videos is processed through deep learning architectures to create motion libraries that represent each patient behaviour. These libraries store feature representations of the motion exhibited by the patient during a seizure. The motion representation of a test patient is compared with the motion library to decide if a new video recording has aberrant semiology.

This chapter is supported by the following published and accepted manuscripts:

- **D. Ahmedt-Aristizabal**, C. Fookes, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, Aberrant Epileptic Seizure Identification: A Computer Vision Perspective, *Seizure*, 65 (2019) 65-71.
- **D. Ahmedt-Aristizabal**, S. Denman, K. Nguyen, S. Sridharan, S. Dionisio, C. Fookes, Understanding Patients' Behaviour: Vision-based Analysis of Seizure Disorders, *IEEE Journal of Biomedical and Health Informatics*, (2019). Accepted (DOI: 10.1109/JBHI.2019.2895855).

7.2 Identification of aberrant epileptic seizures

One limitation of the multi-modal approach to quantify and classify clinical manifestations is that they rely on supervised learning, which assumes the test data originates from one of the training categories. When the unseen data (*i.e.* data for a new patient) does not come from one of the training categories, the trained model is of no use. This is evident with low performance when classifying the semiology of unseen patients (LOSO-CV scheme) explained in Section 5.2.4 and Section 5.3.3. In this scenario, the ictal patterns from test patients with MTLE or ETLE are not strongly present or similar to the semiological patterns of other patients previously observed and contained within the training and validation data. Therefore, the system will show low classification accuracies for new patients with aberrant semiology, even if these patients suffer from the same condition.

In order to conduct this process, we train a deep learning system that detects, tracks, and captures clinical manifestations of known behaviours for two types of seizure MTLE and ETLE. These motions, which are represented by spatiotemporal features (characteristics of shape and motion in videos) and extracted from the trained system, are saved in motion capture (MoCap) libraries. We adopt these libraries to identify if new data samples belong to either of these known semiologies, or represent anomalous behaviours. The identification of anomalies corresponding to unusual semiology will alert clinicians to consider different diagnostic choices. Once a deviation is detected, it may be used to generate a patient-specific alert for consideration by clinicians.

This section is distributed as follows. Section 7.2.1 describes the strategy to detect aberrant epileptic seizures. In this section, the intuition and reasoning behind each phase are explained; Section 7.2.2 presents the dataset and the experimental setup. This section also illustrates the results on verifying the quantification strategy and detection of aberrant seizures; Finally, Section 7.2.3 includes the discussion of the strategy, limitations and possible future directions.

7.2.1 Strategy to identify aberrant semiology

We introduce a system that is able to determine whether a test patient has unusual semiological patterns that do not conform to known behaviours stored in health records. We design an architecture that quantifies semiology to develop MoCap libraries which are used to identify aberrant epileptic seizures. To achieve this, we propose a system with the structure as presented in Figure 7.3. Our approach is based on two key intuitions: (1) Deep learning and computer vision have revolutionised human motion understanding, producing effective motion features of particular dynamics, and (2) MoCap libraries are useful for distinguishing behaviours through simple memorisation of previously observed behaviour and similarities between features.

Each video clip captures one seizure from a patient with MTLE or ETLE, to develop two MoCap libraries (MTLE and ETLE) of known behaviours. To quantify semiology, the inputs of the system are short video sequences rather than a whole video, such that we obtain more data to train the system. We define a sequence as 25 consecutive frames. We preprocessed this dataset by detecting the patient and resizing the images in all seizures. We train a CNN-LSTM structure (See Section 4.3.3) in a supervised fashion to model the relationship between known semiologies from different epilepsy types. Then, spatiotemporal representations from all sequences are extracted from an LSTM layer to generate MoCap libraries for each type of epilepsy. Lastly, when evaluating a new seizure (test patient), we split the recorded seizure into smaller segments (sequences) and match each to the library. We adopt the cosine similarity metric to compute the similarity of each sequence from the test patient with all the sequences of each library. As a result of this, we consider a threshold of acceptance based on the total number of sequences that are similar to the library to identify aberrant behaviour.

Patient detection

We estimate the region of interest that contains the patient using the strategy for human detection proposed in Section 5.2.1. The implementation in Keras and Tensorflow [Abdulla, 2017] is based on the Mask-RCNN architecture [He et al., 2017]. Similarly, we crop and resize all images of each sequence to a resolution of 550×720 pixels.

Deep learning architecture and training

The network architecture used to create the MoCap libraries is based upon well-known cascaded networks for action recognition, and works by capturing spatiotemporal features from video sequences

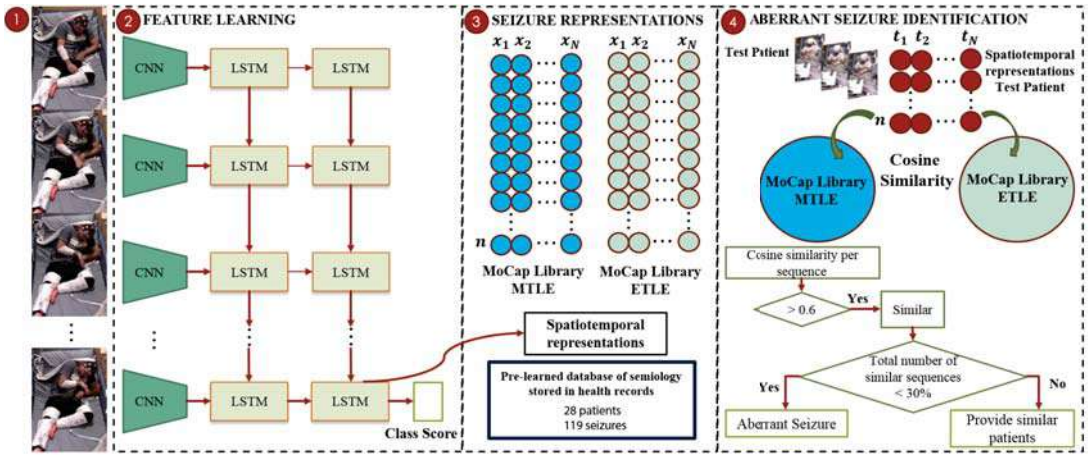


Figure 7.3 Framework proposed to identify aberrant semiology based on MoCap libraries. **1.** Detection of the patient is performed to quantify clinical manifestations from all the body simultaneously. **2.** A CNN-LSTM architecture is designed and trained to quantify and distinguish known behaviours from two types of epilepsy (MTLE and ETLE). **3.** Spatiotemporal representations are extracted and temporally-constructed into feature matrices that represent a MoCap library of each type of seizure. **4.** The cosine similarity measure is used to identify the similarity between a test patient and each MoCap library. Given this average similarity, we consider experimental boundaries to determine if a test seizure is aberrant.

to predict classes through an end-to-end deep learning model. This network is based on the CNN-LSTM architecture proposed in Section 4.3.3.

Although current computer vision approaches are moving into deeper networks, we design a shallow hybrid network using a CNN-LSTM architecture due to the limited data available. Shallow CNN architectures have also already been shown to be suitable for seizure detection with a small amount of training data in [Achilles et al., 2016c].

We exploited many insights about suitable network architectures. It could be argued that using benchmark architectures such as AlexNet [Krizhevsky et al., 2012] and VGG [Simonyan and Zisserman, 2014] to extract spatial information may be preferable over the proposed shallow CNN architecture. However, we choose to design and train our own network to enable our model to learn representations of discriminative patterns of seizure disorders using a small dataset. It is highly recommended to train small networks (150K parameters in our design) because we found experimentally that state-of-the-art networks are not suitable due to their large number training parameters (AlexNet-60M; VGG-144M) and the multiple classes they seek to categorise (the ImageNet database has 1000 object categories). In our scenario, using networks with such a large number of parameters quickly led to over-fitting, or to no significant improvement, even when using regularisation techniques such as overlap pooling, image augmentation and dropout [Pasupa and Sunhem, 2016]. Through extensive experiments, we explore different designs choices

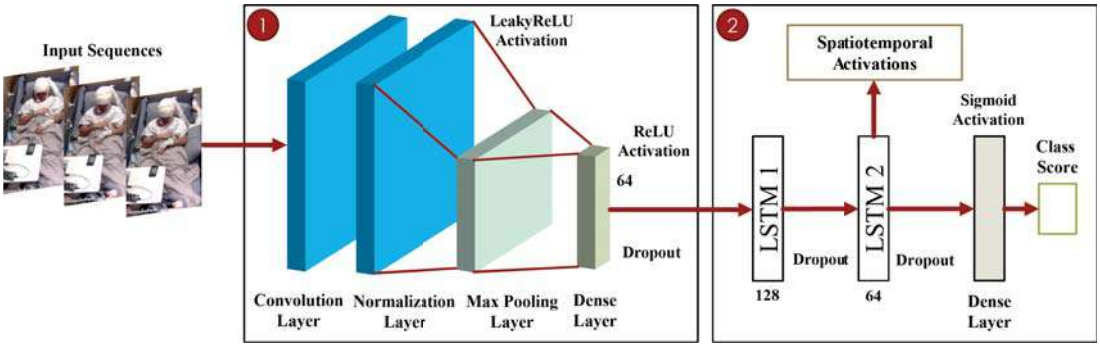


Figure 7.4 End-to-end network architecture to classify epileptic patients and extract spatiotemporal representations. **1.** A CNN architecture is used to extract spatial features. **2.** An LSTM architecture is designed to extract temporal features. We perform classification using a densely connected layer with a sigmoid activation function. Activations are extracted from the second LSTM layer (64 hidden units) to develop MoCap libraries of each type of semiology.

for our model. The design of the network architecture that shows the best performance for the task of learning spatiotemporal features is displayed in Figure 7.4.

For training and testing, all images extracted in the patient detection phase are resized to 155×200 pixels. A feature map is extracted from each input sequence by a CNN architecture containing: (1) one convolutional layer with stride 1 and 8 units, (2) one LeakyReLU activation layer, (3) one normalisation layer, (4) one max pooling layer with stride 12 and (5) one fully connected layer with ReLU activation and 64 units. Max pooling with a large $[12 \times 12]$ receptive field aims to capture rhythmic body movements of the body that are present in data, which results in extensive differences between consecutive frames. Furthermore, the large receptive fields in the max pooling operation reduces the number of parameters in the model, which is preferable when using small datasets. The CNN output is subsequently fed to a stacked LSTM architecture. We adopt an LSTM with 2 hidden layers of 128 and 64 units respectively as proposed in Section 4.3.3. Finally, the output of the second hidden recurrent layer is fed into a densely connected layer with a sigmoid activation function to describe the probability of each patient having MTLE or ETLE behaviour.

We train the CNN-LSTM network by optimising the binary cross-entropy loss. We use the Adam optimiser [Kingma and Ba, 2014] with a learning factor of 10^{-3}, and the first and second moments decay rates of 0.9 and 0.999, respectively. For regularisation, we employ dropout with a probability of 50% in the CNN and 35% in the LSTM architecture, and a batch-size of 32. We train the model over 30 epochs using the default initialisation parameters from Keras [Chollet, 2015] for initialising the weights of the hidden units. We use the Theano backend [Al-Rfou et al., 2016], and balance the training data at the sequence level using the class weight parameters.

Motion capture libraries (MoCap libraries)

The dynamics or representation of semiology to develop the MoCap within each sequence are extracted from the deep learning architecture via the LSTM layer with 64 hidden units as shown in Figure 7.4. Once the features for each sequence are extracted, the representation for each library has a dimensionality of $[2697,64]$ and $[1392,64]$ of known semiologies from patients with MTLE and ETLE, respectively. The first dimension indicates the number of sequences for all seizures, and the second dimension refers to the number of spatiotemporal features extracted from the CNN-LSTM architecture. When evaluating a new seizure, we split the test seizure into sequences and match each sequence to the library. The trained CNN-LSTM architecture is used to extract a feature vector from each sequence, and this feature vector is used to compute similarities to the libraries.

Identifying aberrant epileptic behaviour

The cosine similarity is used to measure how alike two data samples are, *i.e.* if the new seizure has/does not have similar semiological features to any sequence included in each MoCap library. The cosine similarity is a widely used metric for measuring the similarity between hidden states and it is more effective for discriminating the hidden states of deep neural networks than traditional methods using Jacquard similarity or SVMs [Fernando et al., 2018]. The similarity measure is a distance with dimensions representing features of the objects. If this distance is small, it indicates a high degree of similarity while a large distance indicates a low degree of similarity. The similarity is defined as,

$$Similarity = 1 - cos(\theta), \tag{7.1}$$

Given two vectors of attributes $A = [x_1, x_2, x_n]$ and $B = [y_1, y_2, y_n]$, the $cos(\theta)$ is the measure of the angle between the two vectors and is given by,

$$cos(\theta) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}. \tag{7.2}$$

Given m spatiotemporal feature sequences in one MoCap library and t_p spatiotemporal features sequences from a test patient, the result of the cosine similarity will have a representation of size $[m, t_p]$. The values of the cosine similarity range between -1 and 1, where 1 indicates vectors in the same direction ("similar") and -1 indicates vectors in the opposite directions ("dissimilar"). We calculate the average similarity for each sequence of the test patient with each MoCap library in order to identify aberrant behaviour. The average similarity for t_p sequences of a test patient can be represented as $[1, t_p]$. Given this average similarity per sequence, we adopt two experimental boundaries: the first one determines if a single sequence is similar to the MoCap library. The second one, determines if a seizure from test patient has aberrant clinical manifestations based on the total number of sequences from the patient's entire that are similar to each MoCap library (see Figure 7.3). To decide that a single sequence of a test patient is similar to an entry in the MoCap library, we require

a cosine similarity average score to exceed a threshold of 0.6. Then, we calculate the total number of sequences that are similar to the library, to determine a percentage of acceptance. If the percentage of acceptance, *i.e.* the total number of sequences that are similar to the MoCap library, is less than 30%, we consider the test patient to have aberrant or unusual epileptic seizures. This means that the features of the test patient do not conform to well-known behaviour. These boundary levels were defined based on visual inspection of the results, where the main aim was to identify when the similarity was unclear. If during the process of aberrant identification there is evidence of similarity between sequences of the test patient and a MoCap library, the system can indicate the patient that belong those sequences included in the dataset that shows a degree of similarity > 0.6 from strong to weak similarity.

7.2.2 Experimental results

Dataset specification

For this experiment, we select seizures with more homogeneous features and patients that exhibit stereotypical behaviour of each type of epilepsy to assist the construction of the MoCap libraries of known behaviours. We select 119 seizures from 14 patients (62 seizures) with MTLE and 14 patients (57 seizures) with ETLE. The number of seizures per patient is depicted in Table 7.2 and Table 7.3. This data represents a total of 102,225 frames or 4,089 sequences (25 consecutive video frames per sequence).

To evaluate the aberrant seizure identification approach, 5 patients separate from the group used to create the MoCap libraries of known behaviours were selected, all of whom were deemed to have unusual semiology according to the information provided in March 2018 (see Table 3.2). These seizures were selected as their clinical manifestations were not similar to other cases, or show deviations to baselines that are well described and are reproducible in the majority of diagnosed patients, *i.e.* they were different to the baseline of what most patients experience. The aberrant semiology for these patients is described as follows: Patient 1 and 2 exhibit fear expressions, Patient 3 presents swallowing motions, Patient 4 demonstrates finger snapping, and Patient 5 turns their body along the horizontal axis (See Table 7.4).

Experimental setup

A cross-validation evaluation is performed to verify the flexibility of the system to capture and quantify human motion behaviour. Based on the experimental setup proposed in Section 4.4.2, we adopt a k-fold cross-validation and a LOSO-CV.

In the k-fold cross-validation all patients of the same class are split into 70% for training, 15% for validation, 15% for testing and k is set to 10. We compare the performance of our designed CNN-LSTM architecture with alternative approaches based on two-stage architectures. The first stage extracts spatial features using well-known methods which extract human motion features, Mask-RCNN and optical flow, in the same images used in the training and test of our shallow model ($155 \times$

Table 7.1 Multifold cross-validation performance with alternative approaches.

Approach	Validation Accuracy(%)	Test Accuracy(%)	AUC
Mask-RCNN + LSTM	72%	60%	0.68
Optical Flow + LSTM	90%	75%	0.903
CNN-LSTM	**93.4%**	**90%**	**0.9703**

200 pixels). In the second stage, the spatial features are fed to our LSTM architecture which exploits the dynamic variation of these features. Using the architecture of Mask-RCNN [He et al., 2017] and the segmentation of the detected patient, spatial features are extracted from the last fully connected layer in order to identify the semiology. These features have a dimension of [1,4096], as the output of the layer in the network has 4,096 units. Each sequence of features of the movements has a dimensionality of [25,4096], capturing 25 frames, each with 4,096 features. In the case of features from optical flow, we identify the motion vectors related to the human body motions. We compute the optical flow between adjacent frames using FlowNetv2 [Ilg et al., 2017]. We use one threshold on the flow to ensure that there is motion in the frame, *i.e.* more than 10% pixels have optical flow values above zero.

The performance of established networks such as VGG16 and ResNet coupled with our LSTM design, shows AUC values under 0.5, which indicates that the models were not training. For this reason, these architectures are not considered further to compare the performance of the proposed system.

In the LOSO-CV scheme, we confirm the sensitivity and specificity of each MoCap library to distinguish seizures from the two different groups of known epilepsy. For each patient diagnosed with the same type of epilepsy as the MoCap library, we compare the acceptance of the patient with the CNN-LSTM architecture prediction.

In order to test the capability of each MoCap library to identify aberrant seizures based on the cosine similarity, we selected 5 patients in this experiment that exhibited seizures with unusual clinical manifestations as discussed in the dataset specifications.

Verification of MoCap libraries and identification of aberrant seizures

The CNN-LSTM model achieved an average AUC of 0.9703 in the *k*-fold cross-validation scheme. This result outperforms alternative strategies based on baseline approaches which extract human motion features. The performance of these models is summarised in Table 7.1.

Table 7.2 and Table 7.3 display the identification performance conducted for each patient group using each MoCap library, and compares the LOSO-CV performance of the CNN-LSTM architecture for patients of the same group. The "Proportion sequences" column indicates the proportion of the entire corpus that is from each subject. The identification using the trained architecture is shown in the "Test Accuracy CNN-LSTM" column, which reached an average accuracy of 66.48% and 62.19% for patients with MTLE and ETLE, respectively. These results represent a large variability in the

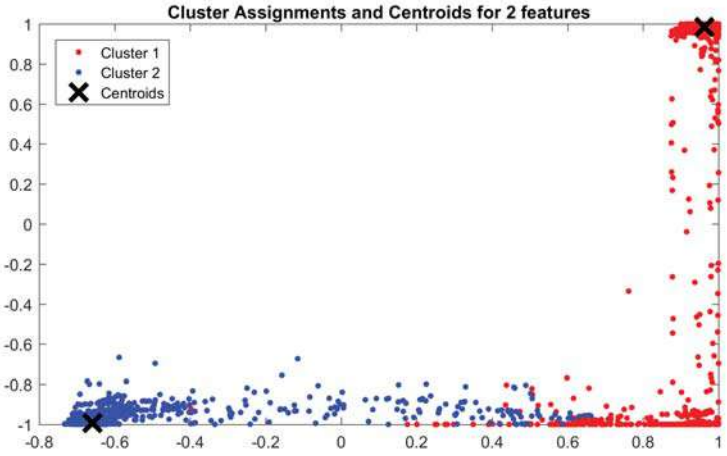


Figure 7.5 Selected sample of clustering epilepsy types by selecting the first two spatiotemporal features from the combined MoCap library. Cluster 1: ETLE (Red dots). Cluster 2: MTLE (Blue dots)

verification of each patient because of the high variation in the data when classifying the semiology of particular patients, which is one of the significant limitations of such strategies discussed in Chapter 4 and 5.

On the other hand, the identification of seizures based on the proposed MoCap libraries and average similarities show more consistent results, as can be seen by the acceptance test of each patient. We have included the average of the cosine similarity of all sequences for each test patient. A label of "Yes" indicates that the patient is considered similar to the library because the percentage of sequences from all seizures with a cosine similarity > 0.6 is greater than 30%. Similarly, a label of "No" indicates that the patient is unrelated to the dataset. The acceptance level of the MTLE MoCap library indicates that 13 of 14 patients were correctly categorised as MTLE and 12 of 14 patients with ETLE were unrelated with the library (Sensitivity: 92.85%, Specificity: 85.71%). Similarly, 11 of 14 patients with ETLE and 13 of 14 patients with MTLE were properly identified with the ETLE MoCap (Sensitivity: 78.57%, Specificity: 92.85%). These results confirm that the MoCap libraries are more robust for identifying clinical presentations than a trained classifier approach, and can be used for the experimental phase of identifying unusual seizures.

To visualise the potential of the features included in the MoCap libraries, we conducted a clustering experiment. We created a combined MoCap library that contains feature from the two epilepsy types of known behaviour with a representation of [4089, 64], where 4,089 is the number of sequences and 64 the number of features. Figure 7.5 shows the clustering of the first two spatiotemporal features. The obtained results show that sequences from MTLE and ETLE were correctly clustered with an accuracy of 90.39% and 89.79%, respectively. This demonstrates that the underlying semiological patterns captured by each class are distinctive.

Table 7.2 Identification performance based on the MTLE MoCap library. We compare MTLE patients data to the MTLE MoCap Library and compute their similarity to an MTLE model trained using the CNN-LSTM method. For completeness, we also compare each of the ETLE sequences to the MTLE MoCap library.

Patient MTLE	Number Seizures	Proportion sequences (%)	Test Accuracy CNN-LSTM (%)	Average Cosine Similarity	Test Acceptance MoCap Library	Patient ETLE	Average Cosine Similarity	Test Acceptance MoCap Library
M1	4	6.12	66.40	0.8390	Yes	E1	-0.1017	No
M2	3	5.47	78.63	0.8382	Yes	E2	-0.3180	No
M3	3	2.38	100	0.8390	Yes	E3	-0.2490	No
M4	8	15.90	48.24	0.5549	Yes	E4	0.0150	No
M5	5	11.08	43.46	0.5420	Yes	E5	0.1050	No
M6	3	4.96	53.77	0.3490	Yes	E6	-0.2540	No
M7	6	10.38	50.00	0.3349	Yes	E7	0.3290	Yes
M8	2	1.03	63.64	0.6248	Yes	E8	-0.0100	No
M9	7	12.39	77.36	0.7125	Yes	E9	0.1930	No
M10	1	2.34	100	0.8390	Yes	E10	0.2700	No
M11	5	7.48	36.25	0.1117	No	E11	0.1580	No
M12	10	13.93	46.64	0.4620	Yes	E12	-0.2850	No
M13	4	4.63	85.86	0.7120	Yes	E13	-0.3115	No
M14	1	1.92	80.49	0.7380	Yes	E14	0.2865	Yes
Average			**66.48**		**13/14**			**12/14**

Table 7.3 Identification performance based on the ETLE MoCap library. We compare ETLE patients data to the ETLE MoCap Library and compute their similarity to an ETLE model trained using the CNN-LSTM method. For completeness, we also compare each of the MTLE sequences to the ETLE MoCap library.

Patient ETLE	Number Seizures	Proportion sequences (%)	Test Accuracy CNN-LSTM (%)	Average Cosine Similarity	Test Acceptance MoCap Library	Patient MTLE	Average Cosine Similarity	Test Acceptance MoCap Library
E1	3	4.48	96.67	0.8059	Yes	M1	-2.0050	No
E2	7	7.46	100	0.8390	Yes	M2	-1.9000	No
E3	3	5.22	100	0.8390	Yes	M3	-1.7580	No
E4	1	2.76	67.30	0.5248	Yes	M4	-0.1060	No
E5	7	12.31	40	0.4915	Yes	M5	0.0180	No
E6	2	1.57	100	0.8390	Yes	M6	0.1090	No
E7	3	7.31	45.71	0.2415	Yes	M7	0.2120	No
E8	4	11.94	63.75	0.8215	Yes	M8	-0.1080	No
E9	6	17.09	40.17	0.4798	Yes	M9	-0.0045	No
E10	5	4.93	36.36	0.1011	No	M10	-0.1790	No
E11	6	11.49	30.52	0.0569	No	M11	0.1060	Yes
E12	4	6.27	57.36	0.8390	Yes	M12	0.0050	No
E13	3	4.18	55.36	0.8390	Yes	M13	-2.1060	No
E14	3	2.99	37.5	0.0958	No	M14	-1.8090	No
Average			**62.19**		**11/14**			**13/14**

The performance when analysing test patients with aberrant semiology is displayed in Table 7.4. The system successfully identifies 4 of the 5 seizures categorised as aberrant semiology correctly. This result allows us to demonstrate that a simple strategy based on MoCap libraries and similarities between features is a promising technique to identify clinical manifestations that can be unusual compared to well-described semiology. From the five selected patients in this experiment, patient 3 was considered similar to the MTLE library, which indicates a false positive detection. Although for this test patient the average similarity of all sequences was low (0.2017), the total number of sequences that were similar to the MoCap library was 34.5%, which is higher than the defined threshold of 30%. For this reason, the test patient was not considered as an aberrant epileptic seizure. In this situation, we have indicated the patients from the MoCap library that were found to have similar motions to Patient 3 based on the level of similarity. It is possible to argue that the swallowing motion of Patient 3 has a similar dynamic to the subtle mouth motions of patients M11 and M12.

Table 7.4 Aberrant epileptic seizure identification based on MoCap libraries.

Test Patient	Number Seizures	Label	MTLE MoCap Library		ETLE Mocap Library		Possible Similar Patients
			Average Cosine Similarity	Is an aberrant patient?	Average Cosine Similarity	Is an aberrant patient?	
1	1	Aberrant	0.1117	**Yes**	-0.0190	**Yes**	
2	1	Aberrant	0.0056	**Yes**	-0.1240	**Yes**	
3	1	Aberrant	**0.2017**	No	-0.1390	**Yes**	**M11, M12**
4	1	Aberrant	0.0124	**Yes**	0.1240	**Yes**	
5	1	Aberrant	0.1070	**Yes**	-0.2540	**Yes**	

7.2.3 Discussion and limitations

We have developed an automated system that is able to identify aberrant seizures utilising visual cues from known behaviours (*e.g.* facial expression, limb posturing, repetitive movements, etc). When the semiology of a new patient does not fit the status-quo of learned information, would be considered as having dissimilar findings, thus aberrant semiology.

Semiology evaluation is dependent on observer experience and training. Such experience is developed over a long period by clinicians and through the comparison of past individual patients. In this contribution, the knowledge learned through previous patients is generalised and described in terms of MoCap libraries of known behaviours, that were extracted from a trained shallow deep learning architecture.

From the experimental evaluations conducted, we have shown that the presented approach for semiology identification, *i.e.*, detection and MoCap library construction, is efficient in both computation and architecture to analyse challenging imaging conditions typical of an EMU. By using a shallow network, the system can be more flexible and simple to transfer to low-end devices such as portable or embedded devices. Our study shows that the analysis of patients considering all body motions simultaneously is viable using the existing monitoring technology in the hospital.

As it was confirmed in the results of multi-modal approaches in Chapter 5, the performance of LOSO-CV schemes is strictly related to semiological patterns contained in the dataset. This is evident from the difference in classification performance between patients based on the deep learning architecture. The obtained results showed that the identification performance ranges from 36.25% to 100% for MTLE and from 30.52% to 100% for ETLE. On the other hand, the MoCap library performance has revealed a consistent performance by categorising the majority of the patients correctly with an average sensitivity of 85.71% and specificity of 89.28% between the two libraries. By demonstrating that is possible to detect outlier semiologies, it is possible to perform active learning by including these features in a system trained to classify epileptic seizures such as the hierarchical multi-modal approach (see Section 5.3) to avoid incorrect interpretation during diagnosis.

We have demonstrated the benefits of our CNN-LSTM architecture and argue that it has advantages over baseline approaches which extract human motion features such as Mask-RCNN [He et al., 2017] and optical flow [Ilg et al., 2017]. Approaches that have proposed the use of optical flow estimation with deep networks for action recognition have shown encouraging results. Although this technique can be used to measure seizures that involve limb and head movements, we argue that they are unsuitable for detecting subtle movements related to semiology from facial modifications and hand

motions. Additionally, the strategy of using the segmentation mask [He et al., 2017] of the detected patient can be unsuitable because this segmentation cannot capture the totality of the patient's body with a fine-tuned model.

We argue that the identification of anomalies is essential to alert clinicians to the occurrence of unusual events that deviate from the majority of examples recorded in the hospital. However, this work is far from being able to replace the expertise of clinical practise. It is an attempt to use a novel approach based on computer vision to take on a complex area such as seizure semiology. The proposed techniques as it currently stands, is not for precise localization purposes or to define the exact underlying epileptic network, but rather to detect semiologies which may be considered "outliers", thereby triggering the need for further investigations. For example, in the case of an MRI lesion and epilepsy, the lesion itself may be unreliable and semiology would be the key to proceed with the diagnosis.

We would also argue that the proposed approach is flexible enough to support finer granularity in term of diagnostic assistance as more data becomes available. Our methodology can be extended to create motion libraries with more specific localisations. With our current system, lateralising signs (head version, figure four position, contralateral dystonia, post ictal nose wipe, etc) were not seen in all of the seizures and lateralising features varied between patients making it difficult to isolate these specific features. At present we have insufficient data to investigate this, however, we have shown that the approach to identify aberrant seizures is promising, infering that with more data the system can achieve greater utility.

This work is by no means a complete solution to semiology assessment, but rather a completely novel method, unreported in the literature, to tackle this highly complex area through further research. We expect that our results will provide the basis of technological development for encoding semiology in the form of MoCap libraries. This would enable more efficient transfer learning between clinical experts and hospitals, avoiding the ethical problem of using identifiable information of patients. When a new patient is presented, his/her seizure can be easily searched for a match in the database conditioned on the behaviour in the MoCap library. It may also help to encourage the adoption of new treatment targets based on recommendations directly from the similarity between patients that were successfully diagnosed.

7.3 Identification of seizure disorders

One-third of patients with epilepsy will not respond to standard anti-epileptic drugs (AEDs) [Wiebe et al., 2001], and resective surgery remains the best option for them. A proportion of these patients, 10-20%, have psychogenic non-epileptic seizures (PNES), also known more recently as functional neurological disorders (FNDs). In this paper, we refer to both diagnostic groups epileptic seizures (ES) and FND, under the common label of seizure disorders. Seizure semiology is an essential component for patient monitoring, and it can be used effectively to differentiate between epileptic and non-epileptic seizures [Noachtar, 2003]. However, FND is commonly mistaken for ES due to overlaying clinical features during a seizure, as both can manifest with rhythmic movements [Sirven and Glosser, 1998]. Accurate classification of FND and ES is required to allow patients to receive appropriate treatment.

Functional neurological disorders have a psychological origin and are categorised as a type of conversion disorder. FND is recognised when epileptiform activity is absent before, during and after a clinical event with a seizure. A normal electroencephalography (EEG) recording, however, does not rule out epilepsy, as some ES, such as simple partial epileptic seizures, can have scalp-negative EEG findings [Devinsky et al., 2011]. FND and ES can be associated with convulsions and/or alterations in behaviour and consciousness. The appearance of a seizure with preserved consciousness frequently leads to the erroneous diagnosis of FND [Cascino, 2002]. Misdiagnosis of seizure disorders exerts a financial and emotional burden on the patient and medical system. Misdiagnosed patients with FND are unnecessarily prescribed AEDs with the potential for adverse effects [Syed et al., 2011]. AEDs do not cure FND, and AED toxicity may worsen their disorder. As such, differentiating between seizure disorders is an uncomfortable and often frustrating challenge were quantitative analysis may help to develop objective criteria to support this analysis.

To assess motor phenomena during seizures and to differentiate FND from ES, some authors have proposed non-camera approaches such as accelerometers (ACM) [Kusmakar et al., 2016, Gubbi et al., 2016]. From the acceleration signals (three-axis motion), multiple time and frequency domain features can be extracted. However, these sensors need maintenance (*e.g.* data synchrony, calibration, batteries) and need to be attached to the body such that they do not detach with seizure-induced movements [Cunha et al., 2016a]. Additionally, there is a need to assess limb-free movements such as facial modifications or hand automatisms (blinking, eyeball deviation, smacking, grimacing, head version, thumb adduction, and fumbling), which requires the detection of fine motions [Noachtar and Peters, 2009].

Video analytic systems are suitable for seizure detection and assessment, but their potential is yet to be fully exploited. For this scenario, motion trajectories obtained from marker-based techniques have shown promising performance in differentiating FND and ES seizures [Chen et al., 2009]. However, these methods are based on reflective material being manually attached to key body parts to quantify motions, have limitations on the camera position, and the patient-attached reflectors can cause discomfort.

The techniques proposed in this thesis so far aim to address the limitation in quantifying clinical manifestations to support the complex diagnosis of epileptic seizures by providing objective information. However, strategies that exploit recent advances in computer vision to distinguish seizure disorders are currently not available.

This section consolidates and evaluates the performance of all deep learning techniques discussed in previous chapters to effectively and efficiently quantify seizures from videos and distinguish between ES and FND, where ES consists of both mesial temporal and extra-temporal lobe epilepsy. To address this problem, we propose and compare two different approaches. The first approach is a fusion method of landmark-based techniques, which uses fine-tuned benchmark architectures to detect and track movements from human pose estimation, combined with features extracted from dense optical flow. In the second approach, we use the designed end-to-end architecture introduce in Section 7.2.1 to classify seizures by analysing the entire body simultaneously eliminating the need to use multiple architectures to assess isolated semiologies.

This section is distributed as follows. Section 7.3.1 describes the two approaches to quantify and detect seizure disorders. The intuition and reasoning behind each approach are also explained; Section 7.3.2 introduces the dataset, the experimental setup and discusses the results in distinguishing seizure disorders; Finally, Section 7.3.3 includes the discussion of the strategy, limitations and possible future directions.

7.3.1 Strategies to identify seizure disorders

We propose two approaches that quantify body movements to classify seizure disorders by using either a fusion of reference points and flow fields, or through analysis of the entire body. We compare and explore the difference between the two approaches. An abstract view of the two approaches proposed is illustrated in Figure 7.6.

The first approach, a landmark-based approach, aims to create an architecture that reflects the semiology with a combination of large and fine motions. We fuse two different feature types: a set of features extracted from time-varying signals of 2D key-points on the body across the sequence (see Section 5.2.2); and a set of features extracted from a convolutional architecture applied to a dense optical flow representation of the input image. We further process the combined features with an LSTM to encode temporal information in the sequence to estimate the seizure type with a soft-max activation. For the second approach, a region-based approach, we train the CNN-LSTM architecture proposed in Section 7.2.1 to model each type of semiology. This extracts a spatio-temporal representations of sequences and performs classification also using a dense layer with a soft-max activation.

Each video clip captures one seizure from a patient with ES or FND. To quantify semiological features, the inputs of the system are short video sequences rather than a whole video, such that we obtain more data to train the system. We define a sequence as five consecutive frames. Input videos are captured at 25 frames per second, and we downsample them by a factor of 5 to extract more

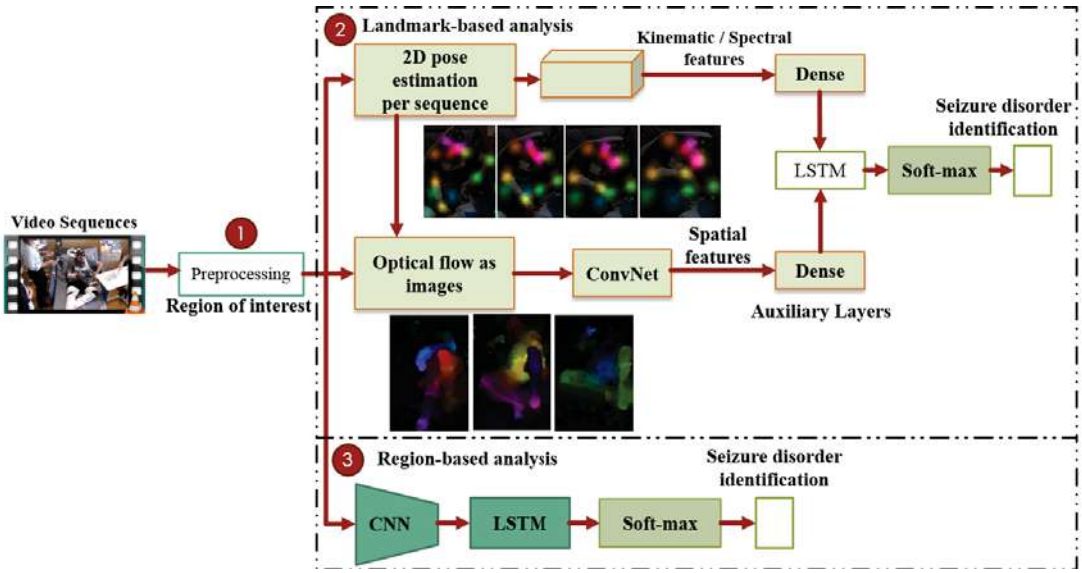


Figure 7.6 Approaches proposed for identifying seizure disorders. **1.** Region of interest definition: patient detection is performed with a region-based CNN object detection architecture to guide further analysis to the patient region of interest only. **2.** Landmark-based analysis: 2D patient poses are localised and tracked per sequence to quantify the head and upper limb movements. Kinematic and spectral features are extracted from the joint trajectories. Pose information and optical flow are used to calculate flow vectors of the human body and deep features are extracted from images of the flow fields with a ConvNet. Features from two channels are merged using additional dense layers. An LSTM is used to extract temporal relations between sequences and perform classification. **3.** Region-based analysis: a CNN-LSTM architecture is designed and trained to classify the type of seizure disorder using sequences of cropped and resized images of the patient detected. The output of the system is represented by the classification accuracy of each patient's symptomatology.

pronounced motion and better recognise seizures. As such, each 5 frame sequence captures 1 second of motion. The region of interest is defined with the patient detected using the approach explained in Section 5.2.1. The cropped and resized images to a resolution of 550×720 pixels are used to compute each approach.

Landmark-based and optical flow approach

Quantifying a person's posture, head and limb articulation is convenient for understanding semiology. The joints estimated from the 2D human pose architecture and the optical flow stored as images are used to extract spatial features, where the long-range dependencies of these features are modelled by the training of a light-weight recurrent neural network.

Extracting features from 2D human pose estimation: We aim to overcome constraints associated with marker-free systems by leveraging recent advances in deep learning for estimating 2D human pose in videos being recorded with current video technology in hospitals. Our purpose is to enhance 2D pose prediction for seizure disorder analysis by extracting consistent poses in sequences using pose

Table 7.5 Comparative of benchmark techniques for 2D pose estimation in videos.

Author	Title	Penn Action	JHMDB
[Luo et al., 2018]	LSTM PM	97.7	93.6
[Song et al., 2017]	Thin-Slicing Network	96.5	92.1

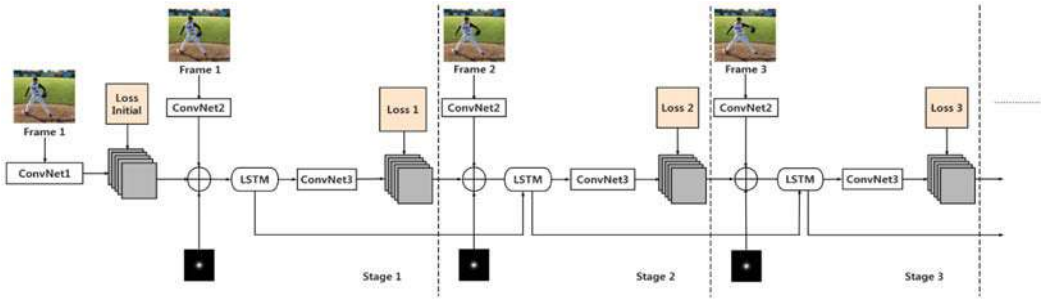


Figure 7.7 Schematic of the pose estimation network in videos. The network consists of T stages which are the number of frames in each sequence. *ConvNet1* is the first stage for initialisation and *ConvNet2* is a CNN network for extracting features. The *LSTM* modules take multiple steps to forget the old memory, absorb new information and create the output. *ConvNet3* produce prediction for each frame based on the outputs from LSTM. Image adapted from [Luo et al., 2018].

tracking approaches [Andriluka et al., 2017]. Following the research area of unified deep learning architectures, we adopt a light-weight model known as LSTM Pose Machines [Luo et al., 2018] which improves the pose tracking performance based on [Song et al., 2017] (see Section 5.2.2). The last one has limitations on the optical flow extraction and the long-range motion blur. Table 7.5 illustrates the improved performance of the proposed technique in public datasets.

LSTM Pose Machines [Luo et al., 2018], which is motivated by [Gkioxari et al., 2016, Wei et al., 2016], allows keypoint localisation under partial occlusion by capturing the geometric relationships between joints over time, increasing the stability of joint predictions on moving bodies. This architecture replaces the multi-stage CNNs [Wei et al., 2016] with ConvNets and LSTMs, and learns the temporal dependencies between video frames.

The recurrent pose machine based on an LSTM unit and derived from the CPM model [Wei et al., 2016] can be formulated as,

$$\begin{aligned} b_t &= g(\mathscr{L}(\mathscr{F}(X_t))), t = 1, \\ b_t &= g(\mathscr{L}(\mathscr{F}(X_t) \oplus b_{t-1})), t = 2, 3, ..., T, \end{aligned} \tag{7.3}$$

where $X_{t(1 \leqslant t \leqslant T)}$ are consecutive frames from a video sequence, $g(\cdot)$ are the identical generators across all stages, $\mathscr{L}$ is the LSTM's memory, $\mathscr{F}(\cdot)$ is a ConvNet used to extract features from input images, T is the length of frames in each sequence and b_t represents the produced belief maps matched with frame $t \in \{1, 2, ..., T\}$.

Figure 7.7 depicts the structure of the LSTM Pose Machine. Consecutive frames in a video sequence are sent to the network as input in different stages. The network consists of T stages, where

T is the number of frames in each sequence ($T = 5$, as per 5 frames per sequence). According to the architecture proposed in [Luo et al., 2018], in each stage, one frame is fed to the network. *ConvNet1* aims to process raw input and the *ConvNet2* ($\mathscr{F}(\cdot)$) is used in all stages to produce preliminary belief maps, where the results are concatenated with new inputs and are fed to the LSTM module. Then, outputs from the LSTM will pass to *ConvNet3*($g(\cdot)$) which generates predictions for each frame. In this scenario, the weights of the ConvNets and LSTM are shared across stages to reduce the number of parameters.

We adopt the pre-trained model on two large-scale video pose estimation benchmarks: Penn Action [Zhang et al., 2013] and JHMDB [Jhuang et al., 2013] datasets; which provides sequences of images under rare poses and strong camera motions (see Table 3.8).

The pose estimation predicts the X and Y coordinates of eight key body-parts (Nose, Neck, Right-Left Shoulder, Right-Left Elbow, and Right-Left Wrist) for each video frame. We generate X-axis and Y-axis movement trajectories for each detected joint per sequence (5 consecutive frames). Figure 7.8 shows qualitative performance of the pose estimation in video sequences as a set of heat maps of the estimated joints.

We extract 25 features for each trajectory and sequence as follows:

- 17 kinematic features: the velocity, acceleration and jerk (the derivative of acceleration) over time are computed, and for each of these signals, we measure the standard deviation, median, mean, maximum, minimum. This yields a total of 15 features per sequence and joint. The total covered distance (sum of the Euclidean distance of the consecutive points) and movement displacement (Euclidean distance between the initial and final position) are 2 further features calculated as proposed in [Cunha et al., 2016a].
- 8 spectral features: for each sequence, we calculate the power spectral density of the displacement and velocity, where the X and Y component of the movement signal are combined before the spectral estimation is computed. From these signals, we measure the entropy, peak magnitude, the sum of the spectrum, and spectral half point (the frequency that divides spectral power into equal halves) as suggested in [Li et al., 2017].

Overall, each defined sequence of features has a dimensionality of $[1,200]$, capturing 25 kinematic and spectral features for each of the eight key joints.

Extracting features from optical flow of the human body: This approach incorporates information from an optical flow in order to model the temporal dependency among video frames. Since the goal of using optical flow is to characterise only the patient motion, we need to supress background motion or motion from hospital equipment, *e.g.* monitoring equipment, sheets, pillows. Following the strategy in [Zhang et al., 2018] and using the 2D localisation of the human body skeleton estimated, we identify the motion vectors related only to the human body. This technique is opposite to existing approaches for pose estimation that use optical flow and the body joint locations in adjacent frames to infer the body joints in the current frame.

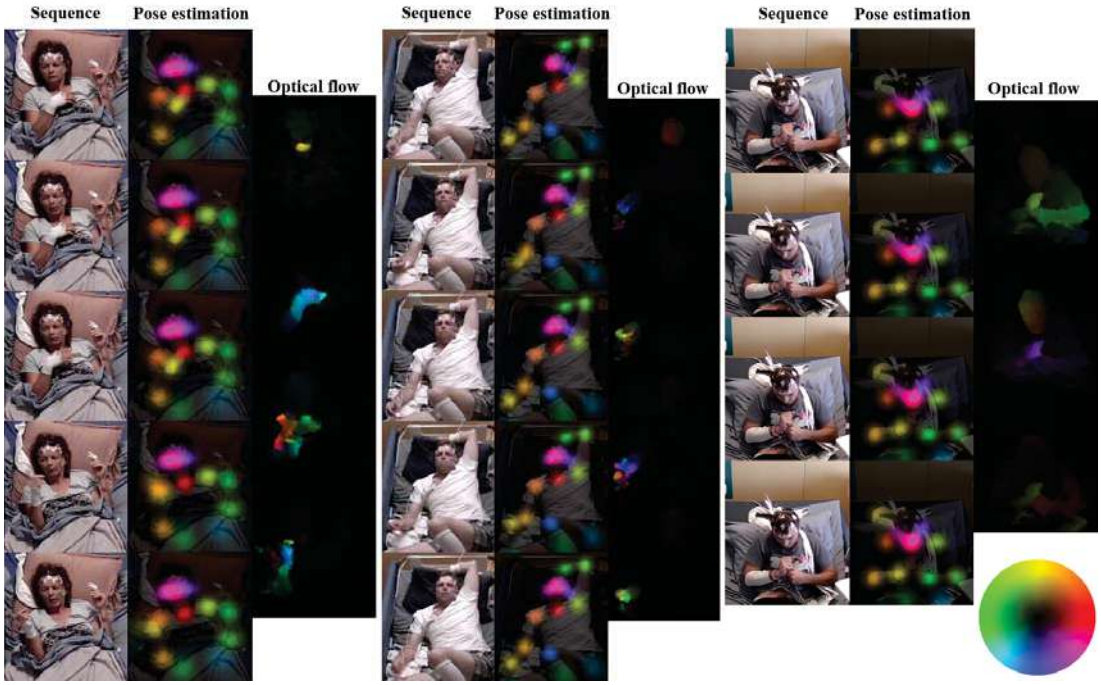


Figure 7.8 Qualitative results of pose and optical flow estimation in a sequence of the epilepsy dataset. From left to right: Patient 1 (ETLE) exhibits mouth motion, blinking and slow right hand movement; Patient 2 (MTLE) experiences subtle mouth motions and waving of the right hand; Patient 3 (MTLE) presents small displacements by trembling both hands. The 2D pose estimation is represented by the heatmaps of the predictions. The pixels in the output map of the optical flow are coloured according to the provided colour map (low right corner). This encodes the direction of movement and the colour intensity reflects the magnitude. All images are cropped to the region of interest.

We compute the optical flow between adjacent frames using [Brox et al., 2004], with the python wrapper for dense optical flow from [Pathak et al., 2017], which allows the analysis of small displacements with fast computation. We use one threshold on the flow to ensure that there is motion in the frame, *i.e.* more than 10% pixels have optical flow values above zero. The underlying accuracy of the optical flow in this seizure environment is beyond the scope of this work due in part to the challenges in obtaining optical flow ground truth, and the fact that we are employing established optical flow techniques. We store optical flow as images by thresholding at $[-20, 20]$ and the horizontal and vertical components are rescaled to the range $[0, 255]$ to simplify later processing. We concatenate images of the optical flow fields generated for each sequence of 5 frames (4 optical flow frames). Qualitative performance of the optical flow in frames from sequences is displayed in Figure 7.8. Then, we use a multi-layer CNN pre-trained for pose estimation [Luo et al., 2018] to extract spatial features from the last fully connected layer, which has 4096 units. Each set of features extracted from the optical flow sequence has a dimensionality of $[4, 4096]$ (4 flow maps of 4,096 features each).

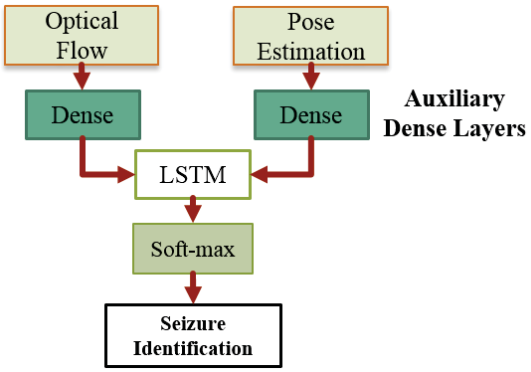


Figure 7.9 Simplified diagram of the fusion strategy in the landmark-based approach.

Fusion and training of LSTMs in the landmark-based approach: We capture dynamics of the patient's body over a sequence by merging the pose (kinematic and spectral features) and optical flow (CNN features from optical flow) representations. Merged data is fed to a two-stage LSTM model.

To merge these features, we adopt a fusion at the feature level (see Section 5.2.3) where a single learning phase handles features from both modes. Figure 7.9 illustrates the simplified diagram of the fusion strategy. Each sequence of features (pose-$[1,200]$ and optical flow-$[4,4096]$) are fed to auxiliary dense layers with 8 cell units, then, the output of each dense layer (a list of tensors) is concatenated using a merge layer from Keras [Chollet, 2015], as input to the LSTM layer. We experimented with various numbers of layers and memory cells based on the small dimensionality of the features, and we choose to use two hidden layers of 128 and 64 memory cell units respectively, as proposed in Section 4.3.3. Following the LSTM layers, a densely connected layer with a soft-max activation function makes a prediction of the seizure disorder for each sequence. LSTM training is carried out by optimising the categorical cross entropy loss using the Adam optimiser [Kingma and Ba, 2014] with a learning rate of 10^{-3}, and decay rates for the first and second moments of 0.9 and 0.999, respectively. We adopt a batch size of 32 and perform the model training for 20 epochs using the default weight initialisation parameters from Keras [Chollet, 2015]. The light-weight LSTM architecture is quick to train and has a low computational complexity, *e.g.* 10 min average training time on one Tesla M40 24GB GPU.

Region-based analysis and training

The region-based method is based on the hybrid CNN-LSTM architecture proposed in Section 7.2.1, which can be applied to the task of seizure disorder identification. The architecture, which is presented in Figure 7.10, is also trained in a supervised fashion, where the output of the second hidden recurrent layer is fed into a densely connected layer with a soft-max activation function to predict the probability of each patient having FND or ES. Similarly, to train the CNN-LSTM network, we follow the strategy documented in Section 4.3.3, nevertheless, in this situation, we use data from 3 different classes: epileptic seizures from patients with MTLE and ETLE, and seizures from patients with FND. We

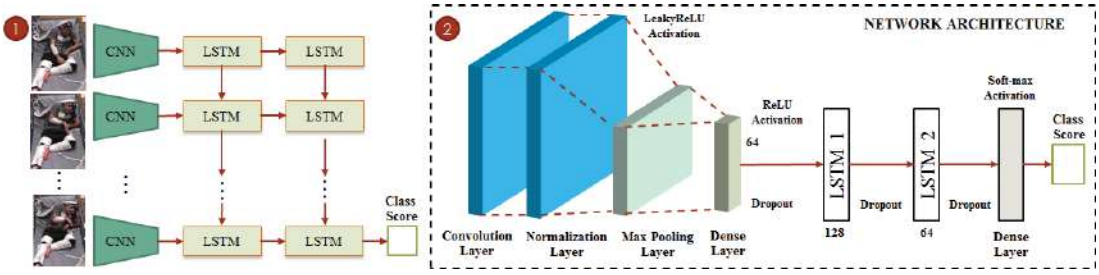


Figure 7.10 The region-based approach to distinguish seizure disorder. (See Section 7.2.1 for information about this design).

adopt a batch size of 15 and train the model over 50 epochs using the default initialisation parameters from Keras [Chollet, 2015].

7.3.2 Experimental results

Dataset specifications

To evaluate the identification of seizure disorders, we analysed 10 videos from 5 patients diagnosed with FND according to the information provided in June 2018 (see Table 3.2). To compare these motor manifestations with epileptic seizures, we preserve the balance of the data between the two types of epilepsy (MTLE and ETLE) and the small data from FND by selecting a subset of the available epilepsy cases (see Table 3.2). We ensure that the database selected captures diverse and varied types of clinical manifestations. Therefore, a total of 50 videos were considered for this experiment: 10 videos from 5 patients with FND and 40 videos from 30 patients with ES (15 patients with MTLE and 15 with ETLE). The data used to quantify seizure disorders represents a total of 15,000 frames (downsampled by 5 from 75,000 frames) or 3,000 sequences (each represented by 5 frames). Patients from the experimental dataset with FND and ES (MTLE + ETLE) are demographically representative and similar to other patients with pharmacoresistant seizure disorders previously described in the literature.

Experimental setup

We evaluate the robustness of the two approaches, landmark-based and region-based, to quantify and classify seizures disorders through a LOSO-CV scheme as proposed in Section 4.4.2. The performance of each seizure disorder (ES and FND) is calculated as the average test accuracy of all patients for each type of seizure. We provide average accuracies for each category of epileptic seizures (MTLE and ETLE) in order to validate which type of epilepsy is more challenging due to similarity with non-epileptic seizures.

Our purpose is to demonstrate that seizures with similar semiological patterns from patients that experience different symptomatology (ES or FND) are sufficiently reliable to categorise video sequences. In our scenario, the proposed algorithms are being used to support clinical decision making

Table 7.6 LOSO-CV performance for seizure disorder identification showing a comparison of average test accuracies between landmark-based and region-based approaches.

Seizure Disorder	Number of Patients	Landmark-based approach (ATA)			Region-based approach (ATA)
		Pose estimation	Optical Flow	Fusion	
FND	5	89.2%	45.2%	85.6%	86.7%
ES (MTLE)	15	41.6%	67.1%	61.4%	73.8%
ES (ETLE)	15	37.2%	71.4%	57.2%	78.2%
Average		56%	61.2%	**68.1%**	**79.6%**

ATA: Average Test Accuracy; FND: Patients with functional neurological disease; ES (MTLE): Patients with epileptic seizure from mesial temporal lobe epilepsy; ES (ETLE): Patients with epileptic seizure from extra-temporal lobe epilepsy.

off-line, as practised in epilepsy centres. However, the issue of developing the system in one hospital and deploying it in another to assess other clinicians performing patient diagnosis is out our scope; and as such we do not compare to human-based diagnostic accuracy.

Identification of seizure disorders (LOSO-CV)

Table 7.6 illustrates the performance of the LOSO-CV for each approach. Regarding the landmark-based approach, we demonstrate the potential of the fusion strategy for assessing seizure disorders, by displaying the individual sensing modality for the pose estimation and optical flow. The fusion proposed in the landmark-based approach achieved an average of 68.1%, confirming that feature fusion improves performance over single-modal techniques (landmark only reaches 56% and optical flow only achieves 61.2%). This is likely due to the optical flow identifying complementary information, which captures behaviours such as head turning, hand motions or trembling (see Figure 7.8) which cannot be detected through pose estimation.

For the region based approach, the LOSO-CV reached an average accuracy of 79.6%. This performance gain is likely due to the region-based approach capturing different types of motions that are distinctive in epileptic seizures (ES) such as facial modifications.

7.3.3 Discussion and limitations

In this contribution, we have developed the first marker-free system that analyses seizure disorders using visual cues from behaviour. Our study has confirmed that by capitalising on recent advances in deep learning to capture patients' behaviour, quantifying seizures under highly varied healthcare conditions is promising, infering that with more data the system can achieve greater utility. The proposed technique as it currently stands is not for precise brain localization purposes, but rather to identify the difference between seizure disorders.

According to Table 7.6, it was observed that the classification accuracy for ES (MTLE and ETLE) increases considerably compared to the landmark-based fusion approach, however, the accuracy for FND remains similar. Patients with ES also commonly exhibited facial modifications including blinking, chewing automatisms and expressions of fear or disgust. They also experience semiology

from hand and finger movements such as finger claw position, tapping or grabbing of objects and snapping the fingers [Noachtar and Peters, 2009]. Although the landmark-based approach that combines human pose and optical flow estimations can be used to quantify vigorous and slow rhythmic movements, it is ill-suited to detecting subtle movements relating to facial semiology and finger motions.

From the experimental evaluations, we show the flexibility of our shallow end-to-end deep architecture in quantifying clinical manifestations in a challenging and natural clinical setting by analysing the entire body simultaneously using the existing video monitoring technology in the hospital. The light-weight network has great potential for low-end devices such as portable and embedded systems. As a possible direction for further investigation, it is worth evaluating the computational cost and performance of multi-scale methods [Neverova et al., 2014], which analyse data at a variety of scales to capture a wider range of relevant features and may enhance the detection of fine movements from mouth, eyes and fingers.

We confirmed that our design can also support the important evaluation of dissociative attacks; however, the region-based approach highlights a significant clinical limitation in evaluating an isolated semiology. The system cannot determine the discriminative features that differentiate disorders and their relationship with specific body movements, or what portion of the body should be observed for diagnosis. Clearly, this limitation can be addressed using the multi-modal and mouth approaches discussed in previous chapters. Therefore, in order to validate the experimental results obtained so far in a future clinical deployment, we encourage researchers to develop an integrated system that exploits the benefits of each approach in distinguishing seizure disorders, isolated semiologies (research aim 1,2,3) and unusual seizures (research Aim 4). This future work is illustrated in Chapter 9.

The integrated system proposed can be useful in capturing and quantifying epileptic seizures. Nevertheless, it is not designed to represent the evolution of semiology, *i.e.* to identify the stepwise progression of clinical features. Thus, we have considered that developing a system that can provide a flow of signals that highlights the changes of semiology and the most common clinical events can be powerful to support the diagnosis of epilepsy. In this chapter, we identified that deep learning approaches eliminate the need for feature engineering but provide a single result that reflects the classifier's decision and confidence. Therefore, in the following chapter, we introduce the development of a system that allows the visualisation and analysis of the dynamic changes in semiology over an entire seizure using a compact image representation known as motion signature.

Chapter 8

Motion signatures and electrical analysis

8.1 Overview

In this chapter, we illustrate the flexibility of a novel approach to visualise and diagnose the full expression of semiology and the advantages of deep learning to effectively learn and model discriminative temporal patterns from EEG sequential data.

Semiology refers to the study of patient behaviour and movement, and the analysis of the temporal evolution of these clinical features during epileptic seizures is also powerful to support the diagnosis of epilepsy. Recent advances in video analytics have been helpful in capturing and quantifying epileptic seizures. Nevertheless, the automated representation of the evolution of semiology, as examined by neurologists, has not been appropriately investigated. The analysis of seizure evolution, which aims to identify the presence or absence of certain movement features (including the order in which they occurred) and the dynamic changes in movement frequency and amplitude during a seizure, is a major component of epilepsy patient assessment [Noachtar and Peters, 2009].

Approaches that use statistical information [Cunha et al., 2016a], provide quantitative movement parameters by considering the totality of semiology duration *i.e.* one metric describes the motion for the full length of the seizure; and as such, this method does not capture information on the semiology changes. The approaches discussed in previous chapters have shown promising results in quantifying and distinguishing epileptic seizures eliminating the need for feature engineering. However, they don't provide clinicians with intuitive tools to support the assessment of seizure evolution. These deep learning based approaches cannot determine the discriminative features that differentiate semiology and their relationship with specific body movements, or what portion of the body should be observed for diagnosis. These systems also provide a single result that reflects the classifier's decision and (to an extent) confidence. Additionally, by analysing short video sequences rather than a whole video of the seizure, the systems are limited when distinguishing frames related to the clinical onset from those which show the propagation of semiology. Overall, the task of developing a computer-based tool that may analyse the stepwise progression of clinical features, which is the scope of this chapter, has neither been considered nor reported in the literature.

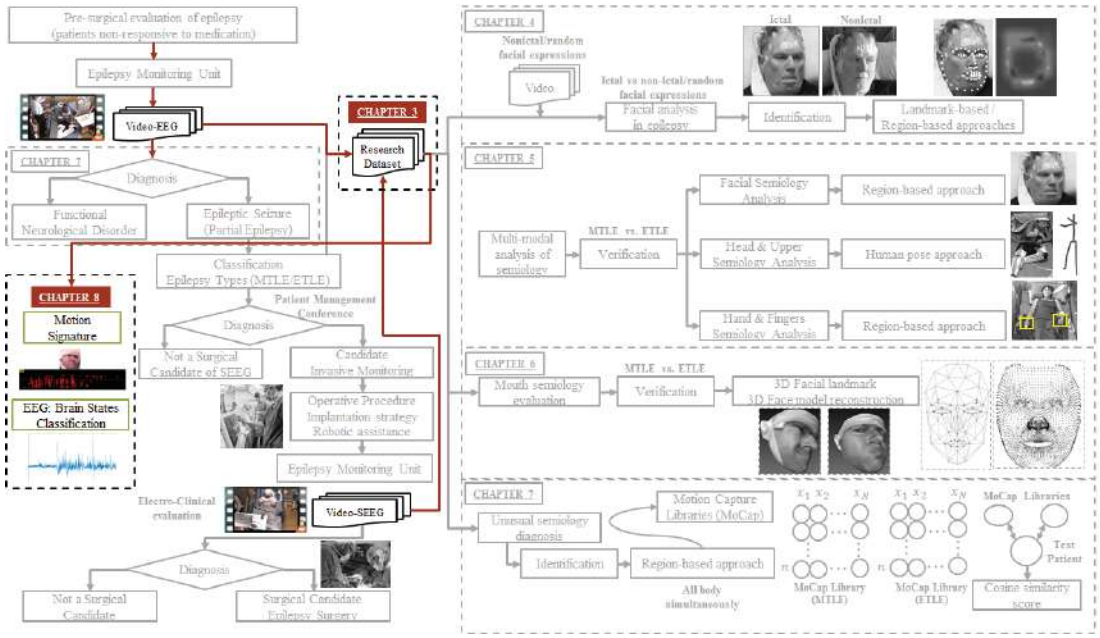


Figure 8.1 Representation of the chapters and their interconnections in the clinical evaluation of epileptic patients. Overview of the research aims in Chapter 8.

Motivated by the significance of analysing the evolution of semiology, in this chapter we provide a system that visualises the dynamic changes in semiology over an entire seizure, which we term a *motion signature*. This computer-aided system presents quantitative information from the representation of semiology as a flow of signs and the correlation between body parts in a real-life clinical setting.

We envision that this system may form the bases for further research to assess electro-clinical features based on multi-sensory feature techniques that can fuse visual and time-series signals (electroencephalography) [Owens and Efros, 2018]. For this reason, we are also interested in providing insights on the benefits of deep learning systems to analyse brain electrical data (EEG signals) by conducting preliminary experiments to quantify and classify physiological brain states. However, the process of quantifying the electro-clinical correlation or the approach to analyse changes in the motion signature and electrophysiological information simultaneously is out of the scope of this thesis. It is important to relate semiology to the underlying EEG discharge; nevertheless, in some situations, the EEG may not show any changes, despite profound semiology (especially in extra-temporal epilepsies) [Chauvel and McGonigal, 2014]. This chapter's research aim and its relationship with the thesis is illustrated in Figure 8.1.

The chapter is distributed as follows. Section 8.2 describes the system capable of computing motion signatures from video recordings of face and hand semiology to provide quantitative information of semiology as a flow of signs. Section 8.3 provides preliminary experiments of a

deep learning strategy to detect different pathological brain states: healthy, interictal and ictal EEG signals.

This chapter is supported by the following published and accepted manuscripts:

- **D. Ahmedt-Aristizabal**, M. Saquib Sarfraz, S. Denman, K. Nguyen, C. Fookes, S. Dionisio, R. Stiefelhagen, Motion Signatures for the Analysis of Seizure Evolution in Epilepsy, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2019.

- **D. Ahmedt-Aristizabal**, K. Nguyen, S. Sridharan, C. Fookes, Deep Classification of Epileptic Signals, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2018, 332-335.

8.2 Motion signatures for the analysis of seizure evolution

Seizures characterised by motor manifestations are analysed and classified on the basis of the type of motor symptomatology. Simple motor seizures are defined by unnatural movements, which can be divided into myoclonic, clonic, tonic, versive, and tonic-clonic seizures; depending on the duration of the muscle contraction, the rhythmicity of movement repetition, and the muscle involved (*e.g.* asymmetric posture, flexion of the neck, abduction of both arms, turning of eyes and head to one side). In more complicated motor manifestations, on the other hand, patients may experience movements that appear natural and involve different body segments (*e.g.* manual and oral automatisms such as chewing, swallowing, smacking the lips and fumbling) [Noachtar and Peters, 2009, Blume et al., 2001]. Seizure manifestations may vary from repetitive rhythmic movement of trunks, limbs or hands such as whole body rocking or manipulation of an object to a more extreme form of presentation with excessive amounts of amplitude, speed, and acceleration [Bonini et al., 2014, Leung et al., 2008, Pfänder et al., 2002, Ridley, 1994].

The analysis of seizure evolution is an important step to evaluate electro-clinical patterns of a seizure, *i.e.* a close observation of clinical features (semiology) and their relation to the region primarily or secondarily involved in the epileptic discharge, allowing a temporal-spatial profile of the seizure's origin and propagation patterns [Bonini et al., 2014, Chauvel and McGonigal, 2014].

We design a system which provides an overview of the motion patterns observed and can support the assessment of patients independent of the motion rate and range, and the amount of data available. We adopt a framework that aims to capture semiology from video recordings and provides interpretable signals of the motion as presented in Figure 8.2. We exploit the results of previous chapters to detect body regions for isolated clinical manifestations (face and hand semiology), and to extract representations of motion between consecutive frames. Then, we estimate the patterns of apparent motion in a defined sequence at pixel-level by computing the optical flow between consecutive frames (*i.e.* a displacement vector assigned to each pixel position). This motion representation is used to estimate the motion signature, which captures the spatial location of semiology and the temporal

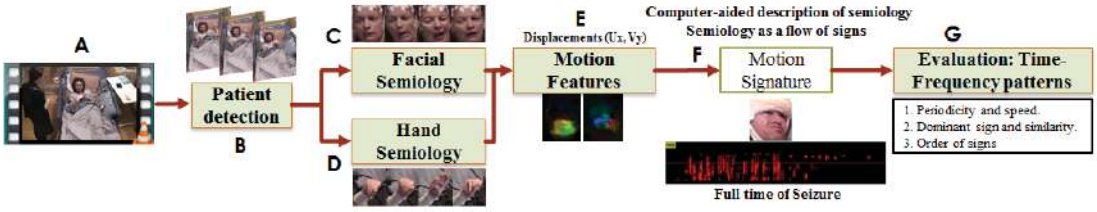


Figure 8.2 Framework proposed that captures the motion dynamics of semiology and creates motion signatures that represent the evolution of epileptic seizures. **A.** Video recording during seizures. **B.** A region of interest is defined to improve the detection of isolated clinical manifestations: face and hands. **C.** Sequences of facial semiology are created based on face detection and tracking. **D.** Sequences of hand semiology are produced via pose estimation techniques in videos. **E.** Extraction of motion features is performed using the optical flow from aligned sequences of consecutive images. **F.** Sequences of flow vectors are used to compute the motion signature for the full length of semiology in terms of time and position. **G.** The computer-aided tool visualises semiology as a flow of signs to provide quantitative information to support the diagnosis of patients.

relation between frames. The motion signature (the semiology as a signal) highlights the motion history of the seizure and correlates types of semiology in the case that a patient experiences multiple semiologies simultaneously. We show how this representation can be used in a clinical environment by providing quantitative information using signal processing techniques.

8.2.1 Region of interest definition: patient, face and hand detection

The patient is detected using the strategy proposed in Section 5.2.1, to crop all images with the region of interest as is shown in Figure 8.2B. Then, from these cropped images of the patient, we extract sequences that capture facial modifications and hand semiology.

To detect the face during epileptic seizures, we adopt the implementation explained in Section 6.2.1, which used the fine-tuned detector from [Hu and Ramanan, 2017] coupled with the tracking strategy of [Wojke et al., 2017]. Figure 8.2C illustrates sequences of facial semiology detected and cropped using this approach.

We capture hand and finger semiology by detecting both hands automatically. Based on the architecture explained in Section 5.3.2, we detect each hand using region-based methods that detect hand-bounding boxes in challenging healthcare conditions. In this scenario, the position of the wrist and elbow are used to approximate the hand location [Simon et al., 2017]. This approach, which is heavily constrained by the predicted pose, allows us to capture fast motions involving upper limb translations (*e.g.* waving), and helps ensure that the fingers of the patient are not located outside the bounding box due to fast movements during a seizure. We aim to extract accurate poses in sequences using pose tracking approaches to estimate hand-bounding box more consistently localised with reduced jitter between frames. Although the proposal for pose estimation in videos discussed in Section 7.3.1 has impressive performance, we opt to select a lightweight, yet highly effective approach that is not limited computationally in analysing long sequences. Different from

Figure 8.3 Selected sequences of hand semiology created using the hand detection strategy. Top: subtle finger motions; Middle: fast hand motions (waving); Bottom: copped images with background removed.

increasing the stability of joint predictions between 5 or 25 consecutive frames, we are interested in capturing the relationships between joints over time in the full length of the seizure. We adopt the architecture of [Girdhar et al., 2018], which is a two-stage approach where first uses a 3D Mask R-CNN [He et al., 2017] to predict the human pose, then implements a lightweight optimization that links the predictions in time. We use a model trained on the PoseTrack dataset to detect the patient wrist and elbow in each frame with the implementation available in [Girdhar, 2018]. Once the pose estimation is performed, we estimate the location of each hand and crop the images to a fixed bounding-box size of 120×120 pixels. The extracted hand-bounding box captures all motions related to the hand and fingers, but also includes information pertaining to background motions such as movements in the bedding, cables and monitoring equipment. To suppress background motion, we adopt a simple strategy of skin segmentation using thresholds adapted for the illumination conditions of our dataset. This algorithm is implemented in OpenCV [Bradski et al., 2000] based on the HSV colour space. Samples of detected hands and the background removal are depicted in Figure 8.3.

8.2.2 Extraction of motion features in sequences

We adopt an optical flow based strategy to capture important information from each type of semiology, including the spatial arrangement of body parts and the rate of change of the arrangement [Horn and Schunck, 1981].

Prior to computing the optical flow, successive frames are geometrically aligned which uses pixel-to-pixel matching by warping the images relative to each other and comparing the pixel intensity values using the enhanced correlation coefficient (ECC) [Evangelidis and Psarakis, 2008]. The benefits of using the ECC strategy is the inference speed due to the simplicity of the iterative scheme to solve the optimization as a linear problem. We use the Euclidean transformation model where the aligned image is a rotated and shifted version of the first image. This method gives good results under various changes in brightness and contrast available on OpenCV release 3.0 [Bradski et al., 2000]. Once all successive pairs of images are aligned, we resize all images to a resolution of height $H = 224$ and width $W = 224$ pixels and compute the optical flow.

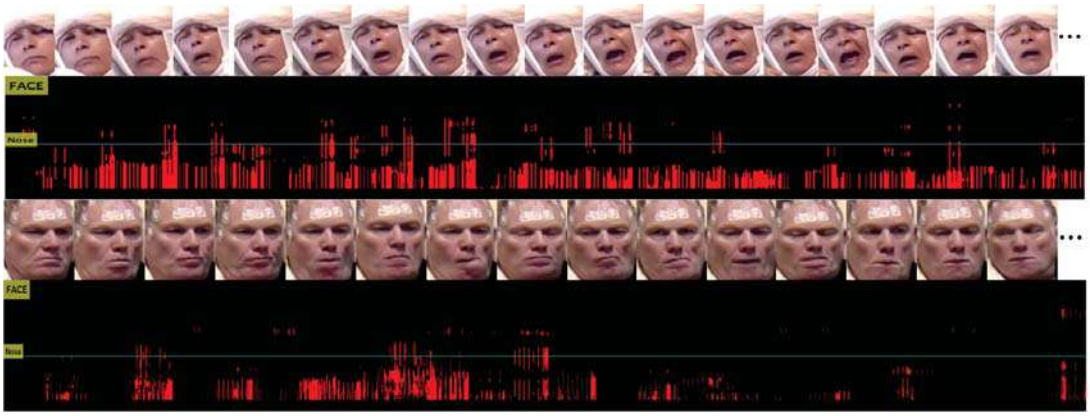


Figure 8.4 Selected samples of motion signatures representing facial semiology. Upper: Patient DG with fast mouth and eye movement (ETLE-opercular-upper bank). Lower: Patient PU with subtle or slow lower mouth movements (MTLE-lower areas).

The optical flow is computed between adjacent frames using FlowNet v2 [Ilg et al., 2017], which is a coarse-to-fine approach that uses stacking CNNs for optical flow refinement allowing the robust analysis of small displacements. We use one threshold on the flow to ensure that there is motion in the frame, *i.e.* more than 10% pixels have optical flow values above zero. Finally, we obtain $N-1$ optical flow maps for N frames, where each flow map has horizontal and vertical (u and v) components.

8.2.3 Constructing the proposed motion signature

To analyse the evolution of a seizure as a flow of signs, we develop a compact image representation of semiology, using the optical flow information, which illustrates the location, variance and periodicity of motion. To have an intuitive understanding of our proposal, consider two flow maps of size $H \times W$ (the same size as the original image/frame). We can measure the change in the motion patterns between them by computing an absolute difference. This should only be high at spatial locations where there was strong movement. We can summarise this change along a given direction in this difference flow map. We sum the obtained difference values along the horizontal direction W, thereby getting a $H \times 1$ motion profile between two flow maps. This motion profile represents spatial motion from top to bottom of the image along the W direction. If we keep computing such motion profiles between successive optical flow maps and stack them together, we will have a *motion signature* that represents the temporal change of motion along the horizontal x-axis over time, and captures the spatial motion of body parts (*e.g.*, eyes, mouth, etc.) along the vertical y-axis. Together, one can see the motion from top to bottom of the image and how it is progressing over time.

To obtain a stable estimate of temporal motion change, we define a temporal window of length L over successive optical flow maps. The sequence of L flow maps in this window is used to capture the motion change in the corresponding image frames. In the sequence, the motion profile (as described above) is computed between all combinations of optical flow maps. To understand it better, if our

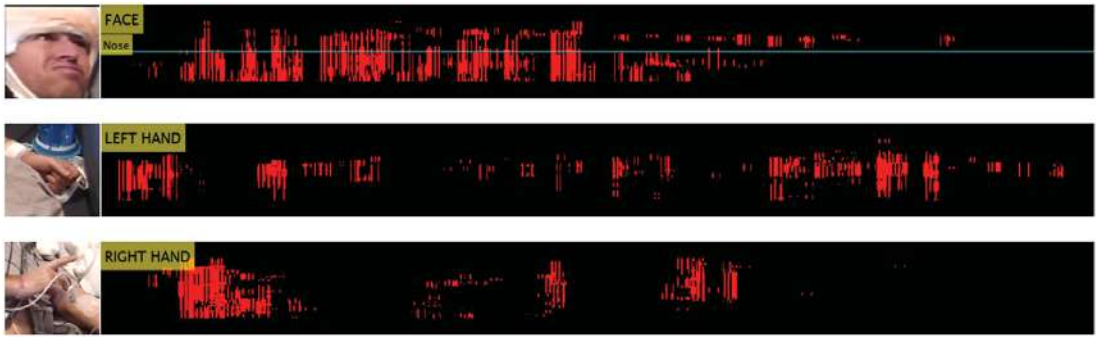


Figure 8.5 Selected samples of a motion signature representing face and hand semiology simultaneously. Patient MB with fast cheek and subtle hand movements is diagnosed with MTLE.

sequence window length $L = 4$, we will compute the motion profile between flow map 1 and 2, 3 and 4, similarly flow map 2 with 1, 3, 4 and so on. Therefore, for a sequence length L we will have a total of $L \times (L-1)$ motion profiles. In our case for $L = 4$ we generate 12 motion profiles. Using such all combinations provides a more stable motion progression and proves to be more robust with respect to any noise such as misalignment between frames. Since we have a defined sequence window to compute such motion profiles we can use the average of the values in the obtained motion profiles (a scalar) as a threshold to only keep real motion. Such a binarization of motion profiles helps to determine if a motion segment is related to a clinical manifestation or noise from the optical flow estimation. This strategy allows us to use it as an online monitoring tool of the evolution of motion *e.g.* by using a buffer of $L+1$ frames.

Figure 8.4 illustrates the motion signature computed from two patients that experienced facial semiology and Figure 8.5 displays the motion signature of a patient that exhibited face and hand semiology simultaneously. The motion of each displacement vector is summarised and represented by a red vertical bar distributed in the x-axis from the start to the full expression of the clinical manifestation and the y-axis corresponds to the spatial location of the motion in the input frame. Black means no seizure related motion.

For the diagnosis of facial semiology, we consider that the location of the nose divides the face into two regions: upper and lower areas. The upper area corresponds to motions in upper facial regions (eyes and eyebrows), while the lower area is related to mouth and chin motions as depicted in Figure 8.4. The continuous line in the facial semiology signature represents the average location of the nose during the full expression of semiology, in order to appreciate the difference between motions from upper and lower facial regions. We compute the location of the nose by adopting the state-of-the-art facial landmark estimation system [Bulat and Tzimiropoulos, 2017b] (see Section 4.3.2). This location is considered as the average estimation of the six facial landmarks that represent the lower nose as depicted in Figure 8.6. To provide and display the information in an appealing way, motion signatures are saved in a video format that allows the user to visualise the motion and the visible image for the entire seizure simultaneously.

Figure 8.6 Representation of the landmarks used to estimate the average location of the nose during a seizure and qualitative results in selected images from the epilepsy dataset (Patient DG, PU and MB).

8.2.4 Experimental results

Experimental setup

To demonstrate the capability of the system to quantify semiology based on motion signatures, we show how this representation can be used in a clinical environment by providing quantitative information from the representation of semiology as a flow of signs. We compute a flow that highlights the most common events and the changes within the recorded seizure, and we show how time-frequency properties computed from the motion signatures can support the assessment by:

- Analysing the motion signature itself, *e.g.* which is the dominant and most frequent sign, blinking or mouth motion.
- Applying frequency-based analysis to identify periodic motions, and showing how we can also use autocorrelation to support this process, to illustrate if each semiology is periodic or a single episode, and it's speed.
- Using power spectrum analysis to quantify the strength of periodic components to determine the dominant semiology (face or hand).
- Displaying the order of signs as a stepwise progression, which is very important as it allows the analysis of the underlying seizure spread.

Analysis of motion signatures of semiology

Identifying dominant and frequent signs in facial semiology: The motion signatures of three seizures from three selected patients are represented as images in Figure 8.4 and Figure 8.5. To calculate time-frequency properties of each signature, we represent the image as a one-dimensional signal, which contains information of the motion location. Using the one dimensional signal of the facial motion, we compute histograms to quantify the number of events recorded in each face location. These histograms are shown in Figure 8.7. The x-axis represents the face location (from lower face to upper face in a scale of 0-224 pixels) and shows the average location of the nose. From the histograms, we can see that for the three patients the dominant sign is mouth semiology. Patient PU has motions in the lower mouth area, Patient DG has motion in the lower and upper mouth and Patient MB has more motions in the upper parts of the mouth (cheeks).

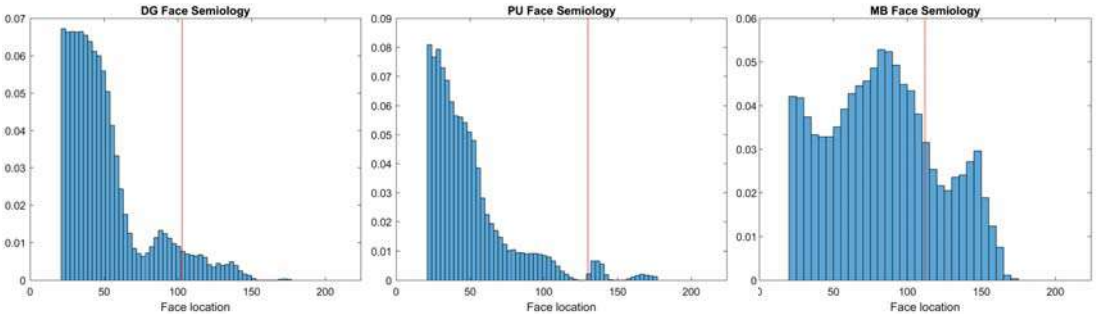


Figure 8.7 Visualisation of dominant motions in the face through normalised histograms. Patients exhibit semiology in the eyes, mouth, chin and cheeks. The vertical line represents the average nose location.

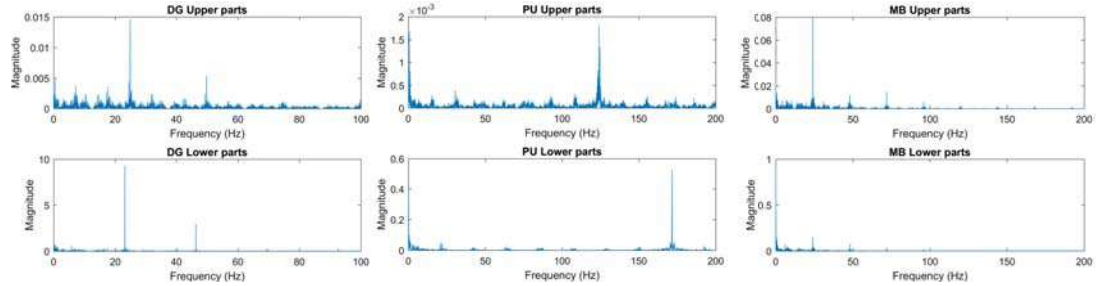


Figure 8.8 Visualisation of the periodogram for upper and lower facial regions for each patient.

Identifying periodic motions in facial semiology: We analyse the periodicity and speed of a signal via spectral analysis based on the power spectral density (PSD) [Fulop and Fitz, 2006] by computing the periodogram, which is given as the discrete-time Fourier transformation of the auto-correlation function. Figure 8.8 illustrates the periodogram for the upper and lower facial regions of each patient. Analysing the lower facial parts for patients DG and PU, the spectral analysis shows statistically significant periods and harmonics, or cycles in the data that stand out from the background noise. However, for Patient MB, there are no clear dominant oscillations or periodicity in the motion of the lower face. For the upper face for the three patients, there is considerable noise that affects the identification of cyclic behaviour, and they also show several spurious peaks that are likely caused by noise. Overall, Patient DG exhibits cyclic behaviour in the lower face with a frequency of approximately 25 Hz while Patient PU shows cyclic behaviour in the lower face with a frequency of approximately 170 Hz. These dominant oscillations allow us to confirm that the speed of the motion for Patient PU is higher than Patient DG.

We confirm the analysis of periodicity of the fundamental spikes in the frequency domain with the autocorrelation of the signal in the time domain. The autocorrelation of a periodic signal has the same cyclic characteristics as the signal itself. Thus, autocorrelation can help verify the presence of periodic behaviour and determine the period [Vlachos et al., 2005]. If the data is periodic, it should

Table 8.1 Autocorrelation of the motion signatures in the time-domain.

Patient	Upper Face	Lower Face	Patient MB	
DG	0.3938	**3.2935** ^{P}	Face	0.8684
PU	0.3745	**2.2653** ^{P}	Left Hand	**1.3860** ^{P}
MB	0.6182	0.6204	Right Hand	0.6666

A value greater than one means the signal has high correlation once the lag time matches the period and can be considered periodic. *P*: Periodic.

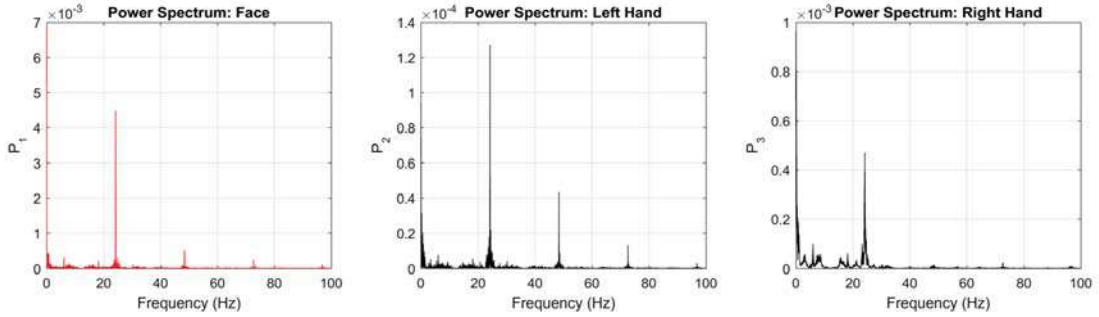


Figure 8.9 Visualisation of the power spectrum for face and hands semiology in patient MB.

have high correlation once the lag time matches the period. As shown in Table 8.1, we can confirm that the mouth motions of Patients DG and PU are periodic.

Dominant signs and periodicity considering face and hand semiology: To evaluate dominant signs, we also can estimate the power of each frequency by computing the power spectrum (PS). The PS of a time-domain signal is the distribution of power contained within the signal over frequency, based on a finite set of data. Considering the total average power as the sum of the power of all the frequency components of the signal, it can be estimated that the mouth motions are more powerful in Patient DG (2.1063) than Patient PU (1.0581). For Patient MB, who experiences face and hands semiology (Figure 8.5), we compute the total average power of the PS as shown in Figure 8.9. It is possible to confirm that the facial motion is the dominant sign in the semiology according to the average power: Face (1.0271), right hand (0.0300) and left hand (0.0021). Only in the left-hand signal are dominant oscillations clear, with spikes in the other signals the result of noise. This periodicity of the signals is confirmed with the autocorrelation results (see Table 8.1), where the left hand can be considered to have a cycle of subtle motion. Considering all motions from the face compared with each hand it is possible to find matching frequencies. It can be seen in Figure 8.5 that the signals have a similar component at 24 Hz. The order of signs in patient MB is illustrated in Figure. 8.10, which is important as it allows the analysis of underlying seizure spread.

Analysing the entire body simultaneously: The motion signature can be also implemented to analyse the whole body simultaneously, and to evaluate isolated semiology such as the complex motor behaviour of body turning [Leung et al., 2008]. Figure 8.11 illustrates the motion signature for this

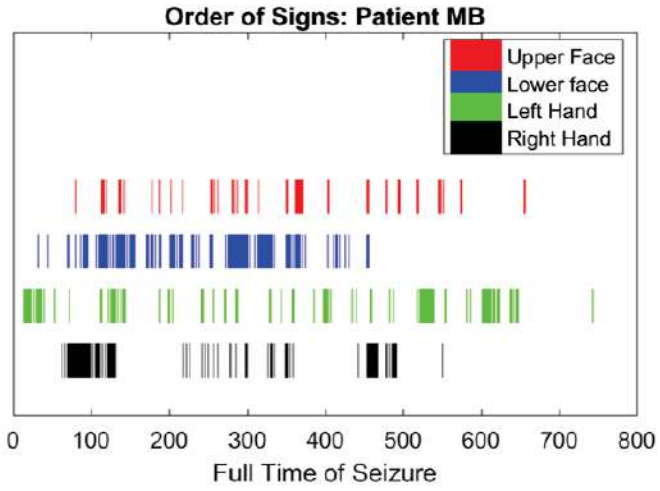


Figure 8.10 Representation of the order of signs as a stepwise progression from the motion signature of patient MB (Figure 8.5).

type of semiology where it is possible to appreciate when the rotation happens and how long it lasts for.

8.2.5 Discussion

In this first part of the chapter, we present a novel and intuitive computer-aided tool to support the expertise of clinical practitioners in the complex area of seizure semiology. The motion signatures are flexible and provide diagnostic assistance when analysing videos in real-life healthcare conditions, presenting semiology as a flow of signs. This strategy enables the use of simple and robust time-frequency techniques to evaluate seizure recordings and isolate repeating patterns. The approach for assistive medical diagnosis in assessing video recordings of seizures, quantifying the dominance, correlation, and motion evolution of semiology from different body parts, has not been previously documented.

One drawback of our system is the reliance on the accurate detection of the regions we monitor for semiology (face and hand), and their alignment and the extraction of flow information; thereby triggering the need for further investigation. However, the system is flexible and the performance of the motion detection and quantification can be easily improved (Figure 8.2 C,D,E) by incorporating new computer vision approaches. For example, it is worth evaluating the computational cost of considering image registration using deep convolutional techniques [DeTone et al., 2016] which have comparable or better accuracy than feature-based or direct methods.

In this work, we have presented an efficient, in both computation and architecture, computer vision approach to capture motion signatures of face and hand semiology, to provide a diagnostic tool to clinicians to evaluate the evolution of clinical manifestations in patients with epilepsy. The motion signatures of epileptic seizures provide relevant features to the physician and a way to intuitively

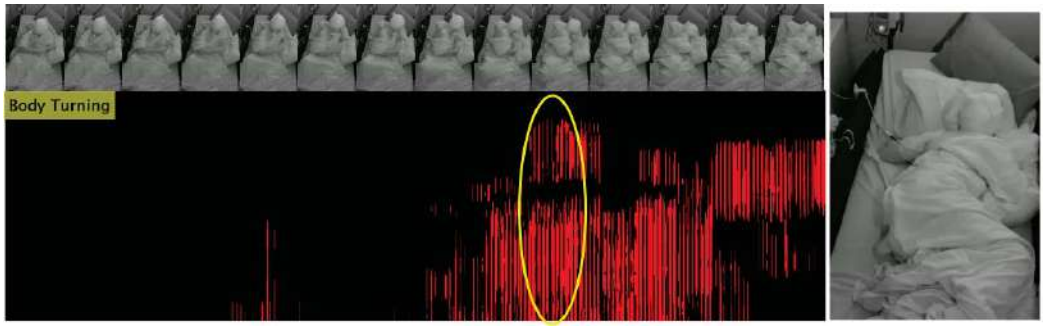


Figure 8.11 Motion signature of the isolated semiology known as body turning. The yellow circle represents the specific moment of the body turning during a seizure.

assess the patient's movement, which is helpful for proper disease management. We expect that a computer-aided tool visualising semiology as a signal could support the electro-clinical analysis that neurologists perform, to aid the progression to successful surgery in patients who are drug-resistant to epilepsy. Finally, the simplicity of our method may enable the diagnosis of patients based on online and real-time monitoring of patients' behaviour.

8.3 Electrophysiological analysis in epilepsy

Electrophysiological observation plays a major role in epilepsy evaluation. However, human interpretation of brain signals is subjective and prone to misdiagnosis. Automating this process, especially seizure detection relying on scalp-based EEG and intracranial EEG, has been the focus of research over recent decades. Nevertheless, its numerous challenges have inhibited a definitive solution. Inspired by recent advances in deep learning, in this section we describe a new classification approach for EEG time series based on recurrent neural networks via the use of an LSTM network. The proposed deep network effectively learns and models discriminative temporal patterns from EEG sequential data. Especially, the features are automatically discovered from the raw EEG data without a pre-processing step, eliminating humans from the laborious feature design task. Our light-weight system has a low computational complexity and reduced memory requirement for large training datasets. This work reinforces the benefits of deep learning to be further attended in clinical applications and neuroscientific research.

In this section, we first present a literature review on automated approaches to assess brain electrical activity in patients with epilepsy using traditional and deep learning techniques, which has been the focus of the majority of research in epilepsy compared to semiology. Then, we introduce our proposed system that captures and classifies patterns from different pathological brain states.

8.3.1 Traditional analysis of brain electrical activity in epilepsy

EEG has long been considered the gold standard for the diagnosis of seizures. The panel of available methods is very wide depending on the aim, such as detection and prediction of seizures, and recognition and classification of EEG patterns. These approaches follow the traditional framework of supervised machine learning techniques for the tasks of feature extraction (time and frequency domains [Qu and Gotman, 1997, Anusha et al., 2012], wavelet transform [Ayoubian et al., 2013, Gajic et al., 2014], energy [Husain and Rao, 2012], and non-linear techniques [Ghosh-Dastidar et al., 2007] and classification or regression (support vector machine [Satapathy et al., 2016, Boashash and Ouelha, 2016] and artificial neural networks [Fergus et al., 2016, Juarez-Guerra et al., 2015, Nigam and Graupe, 2004, Tzallas et al., 2009]. High levels of accuracy, sensitivity, and specificity have been achieved using public datasets such as the CHB-MIT Scalp EEG dataset [Goldberger et al., 2000] and the University of Bonn dataset [Andrzejak et al., 2001]. However, some proposals are difficult to compare because researchers used their own dataset and the distribution is restricted by legal considerations. Although many approaches are robust, none of the existing developments is universally accepted because the performance in clinical scenarios has not been satisfactory. Significant work is still needed to yield expert-level evaluation, specifically in the understanding of the epileptiform activity [Wendling et al., 2016, Fergus et al., 2016, Antoniades et al., 2016], and by generalising representations that are invariant to inter- and intra-subject differences.

The detection and analysis of epileptiform spikes, that is, the ictal onset or the first EEG changes in a seizure, are of importance in the presurgical evaluation of epilepsy and the localization of the epileptogenic network. Characteristics of the EEG ictal activity are a rapid low-voltage discharge with a marked increase of signal frequency and a significant variability in waveforms between patients [Wendling et al., 2003]. Diverse automatic localisation attempts have been made using template-based methods [Wilson and Emerson, 2002], temporal and spatial information [Ramabhadran et al., 1999, Black et al., 2000] wavelet decomposition and generalised Gaussian model [Quintero-Rincón et al., 2016], spectral features [Nasehi and Pourghassem, 2013], and transient events in interictal EEG recordings [Tzallas et al., 2006]. Multimodal methods have supported clinical cases in the absence of epileptiform discharges [Bourien et al., 2005]. For instance, these algorithms estimate seizure onset detection using EEG and functional MRI [Hunyadi et al., 2013], or epilepsy-specific voltage maps of EEG-correlated with hemodynamic changes [Grouiller et al., 2011]. Ding et al. [Ding et al., 2007] proposed an approach of EEG/MEG time series that characterised seizures in space, time, and frequency domains, and distinguished ictal onset from the propagation. Fernandes et al. [Fernandes et al., 2005] quantitatively described the topography and morphology of the epileptiform transients of EEG and MEG showing statistical differences, which indicates that automatic spike detection should take into account the complementary sensitivities of the two techniques.

In regards to SEEG analysis, Gavaret et al. [Gavaret et al., 2009] and Koessler et al. [Koessler et al., 2010] evaluated the accuracy of ictal source localisation using scalp EEG and validated it with SEEG signals. Bartolomei et al. [Bartolomei et al., 2008] chose specific patterns to quantitatively evaluate the degree of epileptogenicity using spectral and temporal properties in SEEG. Wendling et al. [Wendling et al., 2009], defined a temporal–spatial profile of the seizure origin and propagation, which automatically identified the subset of brain structures involved in the generation of intracerebral interictal spikes. Mierlo et al. [Mierlo et al., 2013] validated in SEEG records a methodology that showed connectivity patterns of the seizure onset to localise the ictal onset network accurately, using time-variant effective connectivity and graph analysis on the electrode contacts.

The performance of traditional detection approaches relies heavily on expert knowledge to design the signal features employed. However, there is no warranty that these hand-crafted features are optimal for the chosen task, especially in the complex scenario of brain electrical activity. A major question to be asked is whether the feature engineering can be conducted automatically to discover the optimal features directly from the data, without the need for human-expert knowledge, and domain knowledge. The recent advances in deep learning could be the answer to this question. The main difference between traditional machine learning techniques and deep learning is in the feature engineering as shown in Figure 8.12. In deep learning algorithms, feature engineering is automatically learned from the training data, not by human assumption, leading to natural and effective signal representation and superior performance [Thodoroff et al., 2016].

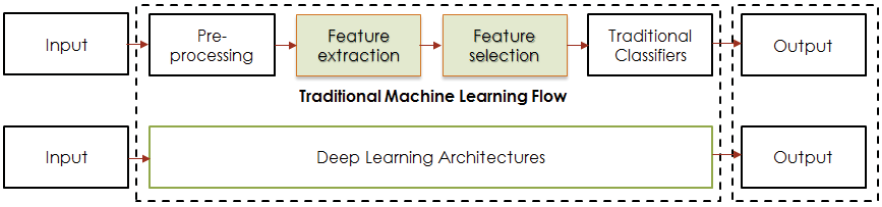


Figure 8.12 Common framework of traditional machine learning and deep learning techniques for brain electrical analysis.

8.3.2 Deep learning techniques for electrophysiological analysis in epilepsy

Deep learning techniques has revolutionised computer vision through end-to-end learning, that is, learning from the raw data, and applying it them to time-series data is gaining increasing attention [Längkvist et al., 2014]. Thorough reviews of deep learning algorithms for healthcare applications based on physiological signals such as brain signals have been documented in [Craik et al., 2019, Faust et al., 2018, Bozhkov, 2016]. For example, these architectures have been implemented in EEG signals for biometrics [Ma et al., 2015], EEG decoding and visualisation [Schirrmeister et al., 2017], EEG-based emotion recognition [Jirayucharoensak et al., 2014], sleep stage scoring based on single-channel EEG [Sors et al., 2018], and modelling cognitive events [Bashivan et al., 2015]. These methodologies address the challenges of machine learning techniques based on temporal and frequency domains, wavelet transforms, or energy analysis and are robust to analyse high-dimensional data with a poor signal-to-noise ratio and considerable variability between individual subjects and recording session [Stober et al., 2015]. All these established findings can apply to the interpretation of EEG signals and how deep learning could benefit electrophysiological studies in epilepsy.

Deep learning architectures within cognitive neuroscience and specifically for processing EEG recordings for epilepsy evaluation have been very limited so far [Craik et al., 2019]. Thodoroff et al. [Thodoroff et al., 2016], assessed the capacity of deep neural architectures to learn a patient-independent representation from EEG to detect seizures. This approach simultaneously captured spectral, temporal, and spatial patterns from an image-based representation of EEG and applied LSTMs, which allows a further analysis of the brain areas involved during the seizure event. Figure 8.13 displays the network proposed in [Thodoroff et al., 2016]. The recurrent architecture depends on the previous and future elements, which is appropriate because neurologists analyse EEG from the past and future windows. However, the performance of sensitivity was low because deep learning models are sensitive to training and parametrization using a small dataset. Lin et al. [Lin et al., 2016] and Viyaratne et al. [Vidyaratne et al., 2016], using deep learning, implemented automatic detection of epileptic seizures by learning the complex and non-stationary epileptic EEG signals and performed feature extraction without relying on methods that are supervised and require domain-specific expertise. The proposal is capable of learning more abstract and high-level representations, which allows for discovering significant differences between seizure and

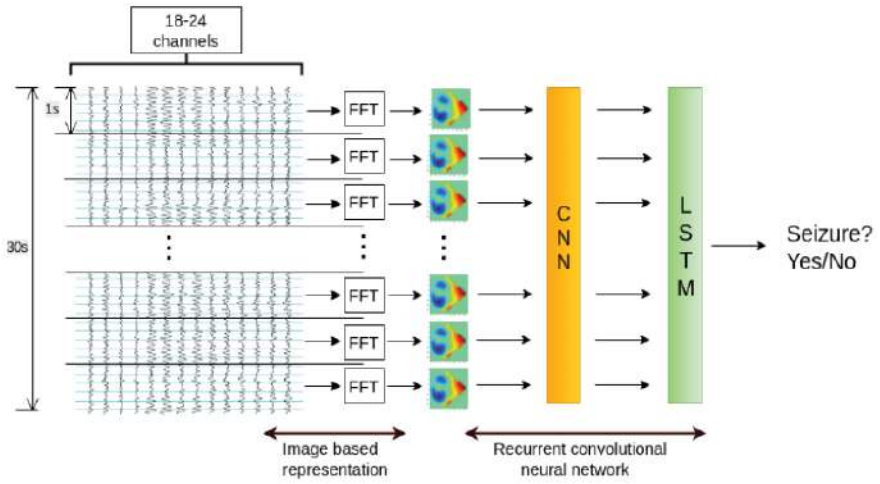


Figure 8.13 Recurrent convolutional neural network using image-based representation of EEG. Image adapted from [Thodoroff et al., 2016].

normal EEG signals. Similarly, Cılasun and Yalçın [Cılasun and Yalçın, 2016] evaluated CNN to detect a seizure, which outperforms previous works using SVM and ANN in public EEG datasets. This model eliminates the requirement of preprocessing and features reduction compared with traditional techniques. In the field of interictal epileptic discharge detection in scalp EEG. Page et al. [Page et al., 2016] proposed personalised model based on CNNs to scalp-based EEG seizure detection. Though the personalised models produce satisfactory results, in real world, it becomes impractically complex in collecting and developing the model for each patients EEG data separately. Antoniades et al. [Antoniades et al., 2016] demonstrated that CNN could learn the intracranial spike waveform patterns. Johansen et al. [Johansen et al., 2016] implemented CNN to learn the discriminative features of epileptiform spikes and detect them for diagnosing epilepsy. Although CNN performance outperforms benchmark classifiers, the model's size was limited in the number of parameters, which is a drawback because large-scale CNN models could increase the performance of detection.

Despite the benefits of these deep learning approaches, there are two major limitations: 1) Current methods either pre-process the raw data into some other forms before being fed into a deep learning architectures such as the Convolutional Neural Network (CNN); and 2) they use very deep and complex networks which have millions of parameters to be trained [Acharya et al., 2017] and require very large training datasets, which are usually not available in the clinical scenarios.

In order to address these limitations, we investigate the plausibility of using deep learning architectures that are capable of both abstracting high-order features with limited training data and classifying them according to the physiological brain state and achieve state-of-the-art performance. We propose a light-weight LSTM network that retains the benefits of deep models. Our system achieves high performance with fast run-time and reduced need for large datasets. Unlike currently

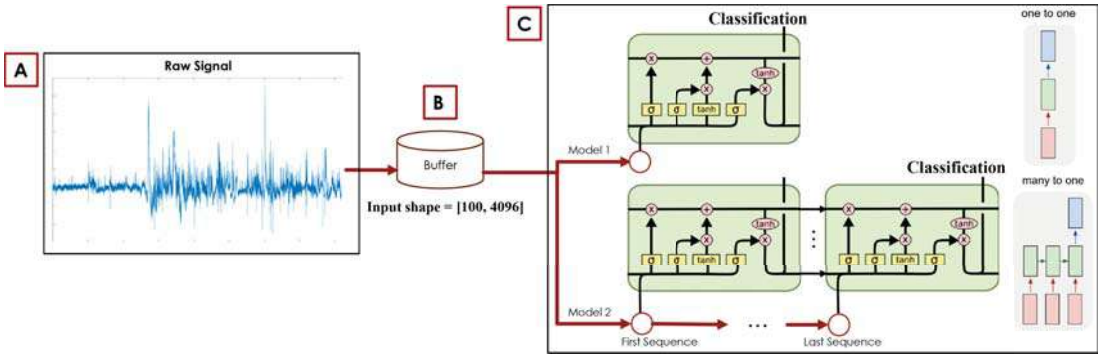


Figure 8.14 The proposed deep framework to identify brain electrical activity of epileptic seizures. **A.** The raw samples for each type of brain-state are concatenated without pre-processing. The value of amplitude of the signal is considered as a single representation of the segment size. **B.** The temporal evolution of the signal is analysed using the complete length of the signal, which indicates a total number of 4,096 segments and 100 samples for each type. **C.** The feature sequence is fed to a Long-Short-Term-Memory (LSTM) structure to exploit the temporal relation between segments and to predict brain-states signals. Two LSTM models were considered to analyse the EEG activity.

used machine learning methods for EEG, the proposed network processes the raw data directly, without any transformation to the original EEG recordings and exploit the temporal patterns through the use of LSTMs. By automatically exploiting and discovering features from the temporal data, the proposed network can extract robust and reliable patterns to classify epileptic signals.

8.3.3 Strategy to identify epileptic signals

The aim of this research is to compare properties of brain electrical activity from different recording regions and from different pathological brain states, *i.e.* to classify healthy, inter-ictal and ictal EEG signals. To achieve this, we propose a deep framework which receives the raw EEG signals and extracts temporal features using an end-to-end training scheme based on an LSTM architecture. Unlike conventional signal-processing techniques where the features are hand-crafted and the signals are pre-processed, our method automatically learns the inherent characteristics of seizure data. The block diagram of the proposed deep learning system is displayed in Figure 8.14.

Dataset specification for this experiment

The experimental data we have used to validate our system is from the publicly available dataset from the Department of Epileptology, University of Bonn [Andrzejak et al., 2001]. The dataset includes five sets (denoted from A to E) with a total of 100 EEG samples for each set. Each sample is a single channel EEG recorded at 173.6 Hz with 23.6 seconds of duration. Thus, the sample length of each sample is 4,096. Set A and B were recorded using scalp EEG from five healthy volunteers (healthy state) with eyes open and closed respectively. Set C, D and E, from five epileptic patients prior to surgery diagnosed with Temporal Lobe Epilepsy, were recorded using depth electrodes implanted

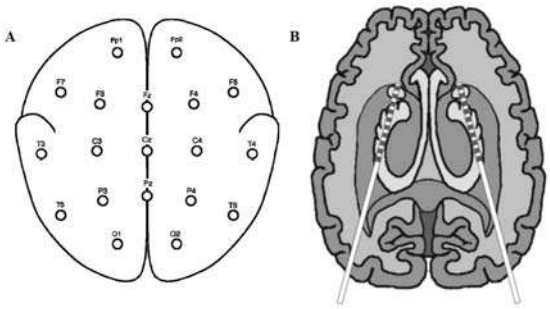


Figure 8.15 Representation of the location of electrodes in the Bonn dataset. **A.** Scalp EEG according to the internation 10-20 system. **B.** Intracranial electrodes implanted for presurgical evaluation of epilepsy patients. Image adapted from [Andrzejak et al., 2001].

Table 8.2 Proposed LSTM architectures for the analysis of EEG signals.

	Model 1: One to One		**Model 2: Many to One**	
Layer Type	Output	Parameters	Output	Parameters
Input	(4097,1)		(4097,1)	
$LSTM_1$	(4097,64)	16,896	(4097,128)	66,560
$Dropout_1$			(4097,128)	
$LSTM_2$			(64)	49,408
$Dropout_2$			(64)	
$Dense_1$	(-1)	65	(1)	65
Total		**16,961**		**116,033**

symmetrically into the hippocampal formation. Set C and D were during seizure-free intervals, where set D was recorded from the epileptogenic zone (Inter-Ictal state or between seizures) and set C from the opposite brain hemisphere. Finally, set E described the recordings of the epileptogenic zone during an epileptic seizure (Ictal state). Figure 8.15 displays the electrodes used to record each type of signal, scalp EEG and intra-cranial EEG.

Deep learning architecture

We adopt two models to analyse the information of the public dataset and the architectures are selected according to the performance of the model for each pair-set. The model one-to-one indicates that from one single layer, the model estimates one single output. On the other hand, the model many-to-one, refers to multiple stack LSTMs that infer one output. Table 8.2 displays the specific LSTM models adopted in the classification process. We obtained the best performance with a network configured with one single layer with 64 hidden units (Model 1) and with 2 hidden layers of 128 and 64 hidden units, respectively (Model 2). For each model, we perform classification using a soft-max layer. We tested more complex architectures but the performance gain is not significant. More complicated architectures have more capability to model complicated signals, but practical clinical implementation

would be affected; hence light-weight architectures with one or two layers could yield very accurate results in the experimental data. Therefore, our models are lightweight, with on the order of less than 17,000 trainable parameters in the case of Model 1. Once the model has been trained, the temporal features that lead to one or another prediction depending on brain state are extracted. These features illustrate the specific structures that should exist in a signal to trigger a specific classification.

8.3.4 Experimental results

Experimental setup

The proposed network is employed to classify six pairs of EEG recordings. These pairs are illustrated in Table 8.3. For instance, the classification between set A and E refers to the verification of healthy volunteers with eyes open and ictal EEG signals. The complete temporal sequence for each set has an input shape of $[100, 4096]$. This illustrates 100 samples, each of them with 4096 segments.

We adopted a k-fold cross-validation to verify the generalisation and robustness of the proposed architecture. For this evaluation, the samples of each set are randomly split into 70% for training, 20% for validation and 10% for testing. The difference between the validation and test samples is that the last one is not seen during the training phase. The validation and test accuracy of the framework is computed as the average performance of each fold (10-folds in this experiment). The performance of the classification task can also be expressed using sensitivity, specificity, precision and the area under the curve (AUC) values.

Training of the LSTM networks is carried out by optimising the binary cross entropy loss function. The model is optimised with the Adam optimizer with a learning factor of 10^{-3}, and decay rate of first and second moments as 0.9 and 0.999, respectively. Batch size set to 4 and dropout with a probability of 0.35, for Model 2, are considered to reduce the overfitting in deep neural networks when dealing with a small training data. We perform the model training using 20 epochs and use the default initialization parameters from Keras [Chollet, 2015] for initialising the weights of the LSTM hidden units.

Classification of brain states

The multi-fold cross-validation average performance is displayed in Table 8.3. The deep framework was capable of achieving an average of 95.54% in the validation accuracy and an average area under the curve of 0.9582 between all the sets pairs. The validation accuracy and error are shown in Figure 8.16 and Figure 8.17, respectively. This demonstrates that the learned features showed clear differences in dynamical properties of brain electrical activity from different physiological brain states.

We can appreciate that the proposed framework achieves a significantly high accuracy of classification with the proposed light-weight deep learning architecture, which has a low computational cost (*e.g.* 4.5 sec average of training time and 200MB of RAM on a 2.6GHz CPU for each set-pair).

Table 8.3 Multi-fold cross-validation performance for epileptic signals identification and comparative with an alternative approach.

Sets	Type Model	Validation Accuracy (%)	Test Accuracy (%)	Test Sensitivity (%)	Test Specificity (%)	Test Precision (%)	AUC	Validation Accuracy (%) [Lin et al., 2016]
A and E	1	**99.50**	**97.00**	96.00	98.00	98.09	**0.9820**	**95.50**
B and E	1	94.75	92.50	91.00	94.00	94.27	0.9850	92.50
C and E	1	97.25	92.00	95.00	89.00	90.06	0.9650	91.67
D and E	1	96.50	91.00	95.00	87.00	89.06	0.9510	93.34
A and D	2	**90.25**	**82.00**	82.00	82.00	84.78	**0.9030**	**86.42**
B and D	2	95.00	93.00	92.00	93.00	93.00	0.9630	N.A
Average		**95.54**	**91.25**	91.83	90.50	91.50	0.9582	

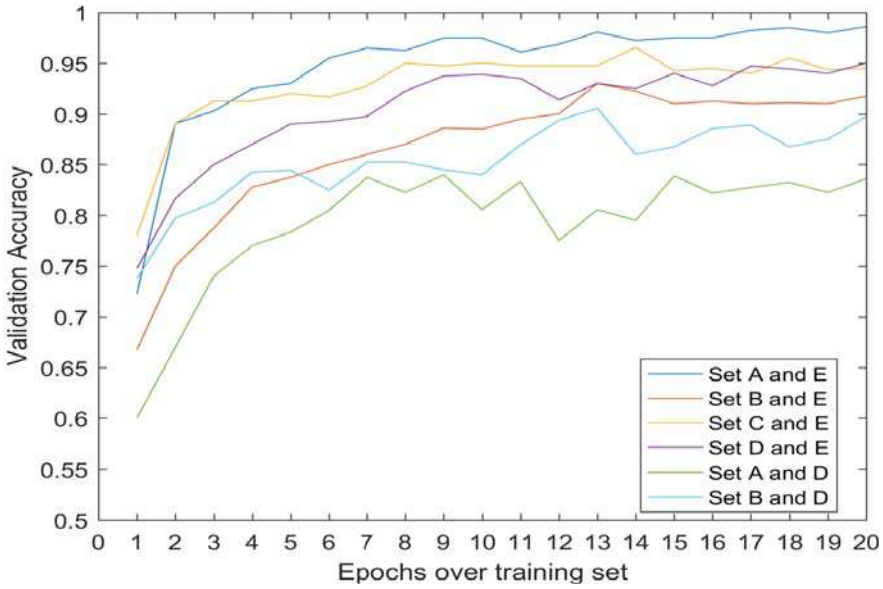


Figure 8.16 Validation accuracy performance of all defined sets.

As illustrated Table II, this performance outperformed the results of a deep learning model based on the same data reported in [Lin et al., 2016], where 90% of the data was used for training in comparison with only 70% of the data in our training. Additionally, our results have reached similar high performance compared to state-of-the-art algorithms based on powerful feature extraction techniques and robust classifiers [Li et al., 2018b], which rely on specific expert knowledge and manual extraction of data. In the experiments, the highest accuracy is obtained with the pair Set A-E, while the lowest is Set A-D. This result was expected because the dynamical properties of the signals from the epileptogenic zone between seizures are more similar to healthy EEG segments than to ictal signals.

8.3.5 Discussion

We have investigated the benefits of a recurrent deep learning framework to classify EEG segments from epileptic signals. We adopt an LSTM network to extract temporal patterns in the frame sequences. Experimental results, confirms that our computationally efficient models can achieve a very high

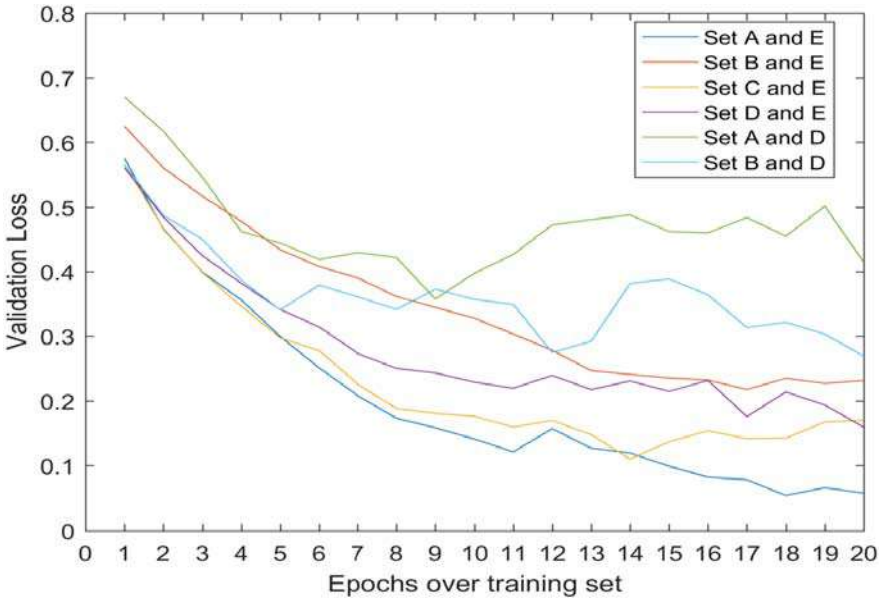


Figure 8.17 Validation error performance of all defined sets.

degree of accuracy. The proposed approach demonstrates the capability of recurrent models to learn a general representation of a seizure event directly and automatically from the raw data. The fast run-time coupled with extremely sparse use of computing resources makes our model desirable for real-time use, for such low computation devices as wearable sensors, which could enhance the diagnosis and treatment planning for patients that experience epilepsy. This is an area requiring further investigation to analyse complex signals from surgical monitoring such as SEEG recordings.

Undoubtedly the electrophysiological investigation has advanced in considerable proportions; nonetheless, clinical researchers have shown the necessity of the electroclinical approaches to analyse the clinical signs and their relationship with the electrical onset and the spreading activity. Furthermore, there are only few studies on epileptiform activity using SEEG recordings. Deep learning techniques have started to be deployed on brain signals for the purpose of epilepsy evaluation, and they could be further utilised to adapt automatic systems to each subject individually to compensate for differences in spatial patterns due to anatomical differences or variance in electrode positions. This represents the ability to learn a general representation of a seizure event, which leads to significant improvements in cross-patient evaluation performance. Furthermore, deep learning may advance the prognostic value of SEEG significantly, in part as it can learn a hierarchical feature representation from raw data automatically. This methodology could be designed to preserve the spatial, spectral, and temporal structure of brain activity, which leads to finding features that are less sensitive to variations and distortions within each dimension.

Chapter 9

Conclusions

Semiology refers to the study of patient behaviour and movement, and their temporal evolution during epileptic seizures. Understanding semiology provides clues to the cerebral networks underpinning the epileptic episode and is a vital resource in the pre-surgical evaluation aiming to achieve seizure freedom. We are motivated to support the diagnostic precision by providing quantitative information from video monitoring recordings. This thesis has presented research in computer vision and machine learning to take on a complex area such as seizure semiology. We have conducted investigations to ascertain if the recent advances in video analytics are robust and flexible to support epileptologists in the clinical diagnosis of patients with epilepsy.

A thorough literature review on the automated analysis of clinical manifestations was conducted and documented in [Ahmedt-Aristizabal et al., 2017]. Although several approaches have been proposed to quantify motions from upper limbs and the head during epileptic seizures, the performance is affected under natural clinical settings which may exhibit variable illumination conditions, partial occlusions and changing viewpoints. There has been considerably less work on analysing facial expressions and providing preliminary experiments implementing multi-modal approaches, *i.e.* to evaluate semiology from different body parts simultaneously. Recent breakthroughs in computer vision and deep learning offer an exciting avenue for clinical application; however, most existing automated approaches to assess semiology eschew deep features. Thus, ongoing research into improving the automatic analysis of semiology is still continuing.

We have proposed to assess semiology from different perspectives using video monitoring recordings from patients undergoing Phase I workup for their drug-resistant epilepsy with collaboration with Dr. Sasha Dionisio and his team at the Mater Hospital in Brisbane, Australia (a tertiary referral public epilepsy surgery centre). Different components were analysed including facial semiology and the isolated mouth semiology, motion from upper limbs and hands, the identification of unusual epileptic seizures, the recognition of non-epileptic seizure disorders and the visualisation of the evolution of semiology as a flow of signs. Within each component, original contributions have been presented to take on a complex area of movement analysis to assist with seizure localization.

To carry out experiments and evaluate our research contributions, we developed a dataset of seizures captured as part of the routine long-term video monitoring. A total number of 161 video clips were captured from 17 patients with mesial temporal lobe epilepsy (MTLE), 17 patients with extra-temporal lobe epilepsy (ETLE), and 5 patients with functional neurological disorder during 3 years of research. Publicly available benchmark datasets in human motion evaluation were also used in the research dataset to support the quantification of semiology.

A novel facial semiology quantification strategy, *i.e.* detection and recognition of facial modifications under challenging natural conditions of a hospital setting was proposed and published in [Ahmedt-Aristizabal et al., 2018b]. We adopted a robust human and face detection approach to detect the patient's face and to quantify facial expressions using a landmark-based and a region-based approach. The landmark-based method is based on a detector of facial landmarks for measurement and quantification of facial motions over time. In the region-based method, spatio-temporal features are extracted to model the variability in morphological and contextual factors of the whole face in consecutive video frames. The use of a region-based method employing a combination of Convolution Neural Networks (CNNs) and a Long Short-Term-Memory (LSTM) network provided a more robust performance to capture the dynamic change of facial physical structure in unconstrained conditions of head pose and illumination changes. Promising results illustrated that quantitative facial evaluation based on deep learning is possible and can differentiate facial semiology from patients with MTLE from spontaneous expressions during routine monitoring. The proposed deep learning model automatically learns spatio-temporal features from raw data, which reduces the need for feature engineering, one of the most time-consuming phases of traditional machine learning in practice.

Deep learning frameworks used to evaluate human behaviour were also found useful to analyse multiple clinical manifestations. We proposed the first in the literature fusion approach documented in [Ahmedt-Aristizabal et al., 2018d] and a hierarchical multi-modal approach published in [Ahmedt-Aristizabal et al., 2018a, Pemasiri et al., 2019a] to quantify and classify seizures from patients with MTLE and ETLE using clinical manifestation from the face, head, upper limbs, hands and fingers motions. The chief contribution of this work was to exploit the potential of deep learning architectures in detecting and capturing the motion of specific body parts. Region-based techniques were used to detect the patient, face and hands during a seizure, while 2D and 3D human pose estimation techniques in videos were adopted to capture motion from the patient's head and upper limbs. The results of this investigation heavily re-enforced the benefits of transfer feature learning of the techniques implemented, which aims to adapt models from one domain to another with minimal data for the new domain, *i.e.* learning across datasets. The extraction of deep features, which represent the motion of the patient during a seizure, enabled the automatic differentiation of epilepsy types in clinical environments. We expect that a larger-scale training dataset that better captures the wide variety of semiological patterns that can occur will result in significantly improved performance. The fusion strategy may not be applicable to every patient because not all patients experience facial and body semiology simultaneously. However, the hierarchical approach showed that it is more suitable to assess the patient because the decision of the system is independent of the type of semiology. It

was also shown that by implementing the proposed techniques it is feasible to analyse videos using the existing camera infrastructure present in Epilepsy Monitoring Units, avoiding the requirement for additional specialist equipment such as depth cameras or wearable sensors. The most obvious distinction of our contribution is the modularity and flexibility of the system, allowing to enhance the performance by replacing each module of human motion quantification with new, more accurate and robust computer vision approaches as they continue to advance in the field.

The use of 3D reconstruction of the face and deep learning techniques were also investigated to detect and quantify the isolated mouth semiology and documented in [Ahmedt-Aristizabal et al., 2019c]. This activity is heavily examined by neurologists but has not been quantified properly. Current proposed computer vision based techniques are unable to accurately quantify mouth motions, which are heavily examined by neurologists to distinguish between seizure types. The proposed system exploited the detailed 3D dense reconstruction from sequences of 2D images, making this, the first-of-its-kind experiment in the epileptic research area. 3D facial landmark can handle large pose variations compared with 2D landmarks, but they are still inadequate to quantify some motions from the cheeks. The introduction of a 3D perspective retains rich information about the shape and appearance of faces, simplifying alignment for comparison between image sequences under challenging conditions. This approach also enhanced computer vision techniques to model and analyse the diverse types of mouth motions exhibited during epileptic seizures compared to purely 2D image-based approaches. With the reconstructed 3D model, hand-crafted features can also be estimated more accurately from the pose-invariant image, but this has not been the focus of this research.

Contributions were also made in the design of a region-based architecture (CNN-LSTM) capable of modelling clinical manifestation to assess two specific and relevant diagnoses: the identification of aberrant or unusual epileptic seizures and the identification of patients with functional neurological disorders or psychogenic non-epileptic seizures. The first identification system for aberrant epileptic seizures is recommended and provided in [Ahmedt-Aristizabal et al., 2019b], using a simple strategy of motion capture (MoCap) libraries extracted from our end-to-end deep learning architecture design, and similarities to pre-learned semiology (past patient cases stored in health records with stereotypical behaviour). The identification of anomalies is essential to alert clinicians to the occurrence of unusual events that deviate from the majority of examples recorded in the hospital. We organised epileptic seizures into a best-fit model, using a template from a pre-learned database of known semiology in the form of libraries (feature representations of motion during a seizure). An aberrant semiology was considered when its properties do not fit the status-quo of learned information. This contribution enhanced the analysis of semiology by enabling active learning to progressively update the quantification strategy of seizures with these new semiology features detected within the system. Additionally, this technique provides the basis of encoding semiology in the form of motion libraries allowing an efficient transfer learning between clinical experts and hospitals with unidentifiable information of patients. The potential of the proposed architecture which captured all body motion simultaneously was also used to investigate the first application of vision-based

techniques to differentiate overlaying clinical features during epileptic and non-epileptic seizures different from traditional non-camera (accelerometers) and marker-based techniques as discussed in [Ahmedt-Aristizabal et al., 2019a]. The analysis of the entire body is compared with a fusion approach of reference points (body landmarks) and flow fields. It was shown that the region-based approach reached a better performance due to the ability to capture subtle movements relating to facial semiology and finger motions.

The final contributions of the thesis was the proposal of the first of its kind computer-aided system that captures the dynamics of semiology as a flow of signals enabling the visualisation and diagnosis of the seizure, *e.g.* semiology evolution and correlation between body parts in a real-life clinical setting as documented in [Ahmedt-Aristizabal et al., 2019d]. The system may form the bases for further research to evaluate electroclinical patterns of a seizure, *i.e.* a close observation of clinical features (semiology) and their relation to the region primarily or secondarily involved in the epileptic discharge. Preliminary investigation towards the development of a light-weight deep network was also proposed to provide insights into the benefits of deep learning systems to analyse electrographic patterns (EEG signals) as published in [Ahmedt-Aristizabal et al., 2018c]. Results showed the capability of recurrent models to learn a general representation of a seizure event directly from the raw data to quantify and classify physiological brain states.

Future work

Research has been presented in this thesis which has addressed the problem of modelling the mechanisms of semiological differences in patients with epilepsy with objective and quantitative motions analysis. Nevertheless, the underlying mechanics of clinical semiology can be much harder to interpret. We have shown the benefits of the recent explosion of artificial intelligence techniques in video analytics in capturing and quantifying semiology during epileptic seizure which is an area that has seen limited advances in the literature. These research contributions are unique and provide important supplementary and unbiased data to assess semiology. They are a vital complementary resource in the era of seizure-based detection through electrophysiological data. This thesis successfully demonstrates a basis for ongoing significant breakthroughs in the field of epilepsy. The results documented in this thesis are far from being able to replace the expertise of clinical practise. The proposed techniques as it currently stands, is not for precise localization purposes or to define the exact underlying brain epileptic network, but rather to provide objective information of the motion recorded, thereby triggering the need for further investigations.

Some possible future directions for research include,

- Analysis and development of a closed-loop system that integrates each proposed architectures in this thesis combined with the current technology of the hospital to evaluate the performance in patients under diagnosis. We aim to exploit the robust performance of the region-based system in analysing the whole body simultaneously to distinguish seizure disorders (Section 7.3), with the flexibility of the modular multi-modal approach (Section 5.3) and the mouth semiology

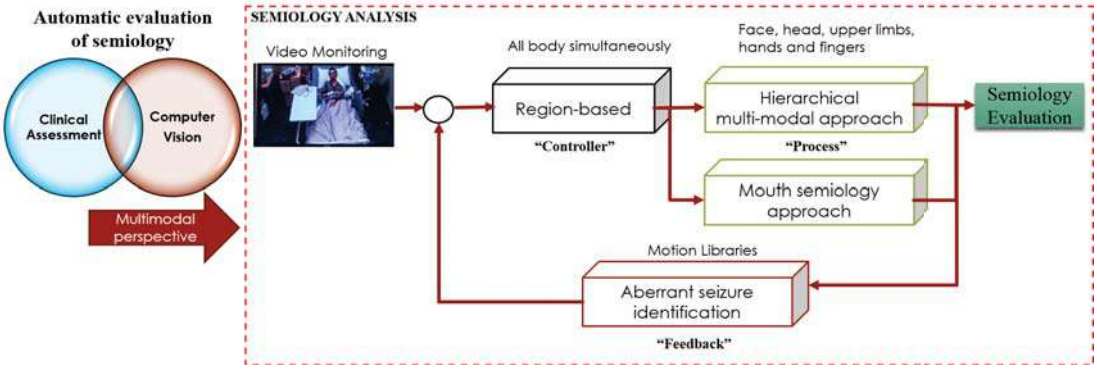


Figure 9.1 Block diagram of the closed-loop system to assess semiology. The "Controller" will provide to the "Process" only seizure disorders that are categorised as epileptic seizures. Then, the "Process" will capture and provide the classification results of each type of semiology (face, mouth, head and upper limbs, and hand and fingers) to clinical experts to evaluate a general condition of the symptomatology of the patient. This trained model will have a "Feedback" which will be responsible to conduct active learning and progressively update the system with new semiologies identified.

strategy (Section 6.2) to evaluate isolated semiologies. This implementation can be updated with aberrant motion representations detected within the system (Section 7.2). Figure 9.1 illustrates the integration of each proposed approach.

- The motion signatures proposed in Chapter 8 to analyse the seizure evolution is a powerful tool to evaluate clinical manifestations because it picks up dominant features and divides them into speed, periodicity and amount of events. However, further analysis and evaluation to capture the differences in motions should be considered, which is the key in clinical work.

- Development of a methodology that could jointly learn across visually observed semiology and electrical patterns from Stereo-EEG recordings, expected to differ quantitatively between epilepsies, and to predict the linked subsets of brain networks involved in a seizure event as it is shown in Figure 9.2. The research is based on the hypothesis that seizures with similar intracerebral EEG changes and motor symptoms involve neuronal activity within the same specific brain networks and are sufficient to categorise patients with a specific type of epilepsy. We have demonstrated the advantages and robust performance of implementing deep learning architectures to analyse video detection systems and brain electrical activity, outperforming traditional techniques and solving current limitations. Therefore, we aim to assess electro-clinical representations based on self-supervised multi-sensory systems [Owens and Efros, 2018]. We expect to implement an early fusion method where only one learning phase handling all multi-sensory features. To confirm the brain anatomical areas, a strategy can be sought by correlating the electrical activity recorded with the anatomical localisation of each contact and electrode. Developing strategies for clustering and statistical

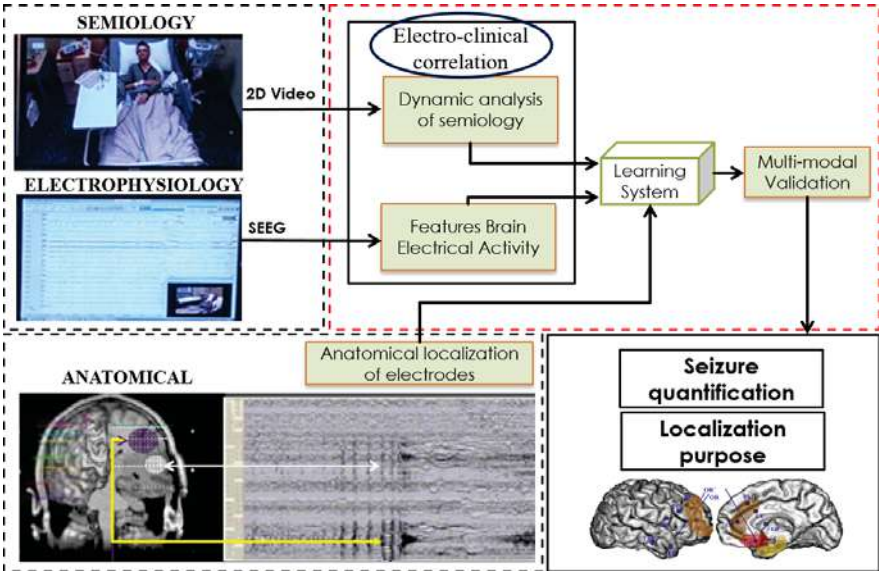


Figure 9.2 Proposed methodology of the learning system and validation of the anatomical, electrical, and clinical features.

techniques will generate anatomical features as an input of the multi-sensory architecture to provide a correlation between electro-clinical features and involved anatomical areas.

- Provide an information guide as to what is the underlying areas of cortical information, which is a highly complex analysis because of the interplay between cortical regions, sub-cortical regions, brainstem, spinal cord, frequency changes and because we still cannot fundamentally explain the movement.

Bibliography

[Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: a system for large-scale machine learning. In *Proceedings of the Operating Systems Design and Implementation (OSDI)*, volume 16, pages 265–283.

[Abbasnejad et al., 2017] Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., and Lucey, S. (2017). Using synthetic data to improve facial expression analysis with 3D convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1609–1618.

[Abdulla, 2017] Abdulla, W. (2017). Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN.

[Acharya et al., 2017] Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., and Adeli, H. (2017). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in biology and medicine*.

[Achilles et al., 2016a] Achilles, F., Choupina, H., Loesch, A., Cunha, J., Remi, J., Vollmar, C., Tombari, F., Navab, N., and Noachtar, S. (2016a). Ep 114. uncovering epileptic seizures–a feasibility study for the semiological analysis of hidden patient motion during epileptic seizures. *Clinical Neurophysiology*, 127(9):e289.

[Achilles et al., 2016b] Achilles, F., Ichim, A.-E., Coskun, H., Tombari, F., Noachtar, S., and Navab, N. (2016b). Patient mocap: Human pose estimation under blanket occlusion for hospital monitoring applications. In *Proceeding of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 491–499.

[Achilles et al., 2016c] Achilles, F., Tombari, F., Belagiannis, V., Loesch, A. M., Noachtar, S., and Navab, N. (2016c). Convolutional neural networks for real-time epileptic seizure detection. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–6.

[Adams et al., 1997] Adams, R. D., Victor, M., Ropper, A. H., and Daroff, R. B. (1997). Principles of neurology.

[Agarwal et al., 2012] Agarwal, S., Mierle, K., et al. (2012). Ceres solver.

[Aggarwal and Cai, 1999] Aggarwal, J. K. and Cai, Q. (1999). Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440.

[Ahlberg, 2001] Ahlberg, J. (2001). Candide-3-an updated parameterised face.

[Ahmedt-Aristizabal et al., 2019a] Ahmedt-Aristizabal, D., Denman, S., Nguyen, K., Sridharan, S., Dionisio, S., and Fookes, C. (2019a). Understanding patients' behaviour: Vision-based analysis of seizure disorders. *IEEE Journal of Biomedical and Health Informatics*.

[Ahmedt-Aristizabal et al., 2018a] Ahmedt-Aristizabal, D., Fookes, C., Denman, S., Nguyen, K., Sridharan, S., and Dionisio, S. (2018a). A hierarchical multi-modal system for motion analysis in epileptic patients. *Epilepsy & Behavior*, 87:46–58.

[Ahmedt-Aristizabal et al., 2019b] Ahmedt-Aristizabal, D., Fookes, C., Denman, S., Nguyen, K., Sridharan, S., and Dionisio, S. (2019b). Aberrant epileptic seizure identification: A computer vision perspective. *Seizure*, 65:65–71.

[Ahmedt-Aristizabal et al., 2017] Ahmedt-Aristizabal, D., Fookes, C., Dionisio, S., Nguyen, K., Cunha, J. P. S., and Sridharan, S. (2017). Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey. *Epilepsia*, 58(11):1817–1831.

[Ahmedt-Aristizabal et al., 2018b] Ahmedt-Aristizabal, D., Fookes, C., Nguyen, K., Denman, S., Sridharan, S., and Dionisio, S. (2018b). Deep facial analysis: A new phase I epilepsy evaluation using computer vision. *Epilepsy & Behavior*, 82:17–24.

[Ahmedt-Aristizabal et al., 2018c] Ahmedt-Aristizabal, D., Fookes, C., Nguyen, K., and Sridharan, S. (2018c). Deep classification of epileptic signals. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 332–335.

[Ahmedt-Aristizabal et al., 2018d] Ahmedt-Aristizabal, D., Nguyen, K., Denman, S., Sridharan, S., Dionisio, S., and Fookes, C. (2018d). Deep motion analysis for epileptic seizure classification. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 3578–3581.

[Ahmedt-Aristizabal et al., 2019c] Ahmedt-Aristizabal, D., Nguyen, K., Denman, S., Sridharan, S., Dionisio, S., and Fookes, C. (2019c). Vision-based mouth motion analysis in epilepsy: A 3D perspective. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*.

[Ahmedt-Aristizabal et al., 2019d] Ahmedt-Aristizabal, D., Sarfraz, M. S., Denman, S., Fookes, C., Dionisio, S., and Stiefelhagen, R. (2019d). Motion signature for the analysis of seizure evolution in epilepsy. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*.

[Akhter and Black, 2015] Akhter, I. and Black, M. J. (2015). Pose-conditioned joint angle limits for 3D human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455.

[Al-Rfou et al., 2016] Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., et al. (2016). Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*.

[Andriluka et al., 2017] Andriluka, M., Iqbal, U., Milan, A., Insafutdinov, E., Pishchulin, L., Gall, J., and Schiele, B. (2017). Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Andriluka et al., 2014] Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693.

[Andrzejak et al., 2001] Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907.

[Angermueller et al., 2016] Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878.

[Antoniades et al., 2016] Antoniades, A., Spyrou, L., Took, C. C., and Sanei, S. (2016). Deep learning for epileptic intracranial EEG data. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

[Anusha et al., 2012] Anusha, K., Mathews, M. T., and Puthankattil, S. D. (2012). Classification of normal and epileptic EEG signal using time & frequency domain features through artificial neural network. In *Proceeding of the International Conference on Advances in Computing and Communications (ICACC)*, pages 98–101.

[Arain, 2017] Arain, A. M. (2017). EEG and semiology in focal epilepsy. In *Epilepsy Board Review*, pages 109–113.

[Arel et al., 2010] Arel, I., Rose, D. C., Karnowski, T. P., et al. (2010). Deep machine learning-a new frontier in artificial intelligence research. *IEEE computational intelligence magazine*, 5(4):13–18.

[Ataoğlu et al., 2015] Ataoğlu, E. E., Yıldırım, İ., and Bilir, E. (2015). An evaluation of lateralizing signs in patients with temporal lobe epilepsy. *Epilepsy & Behavior*, 47:115–119.

[Aupy et al., 2018] Aupy, J., Noviawaty, I., Krishnan, B., Suwankpakdee, P., Bulacio, J., Gonzalez-Martinez, J., Najm, I., and Chauvel, P. (2018). Insulo-opercular cortex generates oroalimentary automatisms in temporal seizures. *Epilepsia*, 59(3):583–594.

[Australia, 2016] Australia, E. A. (2016). Facts and statistics about epilepsy. https://www.epilepsy.org.au/resources/for-media/facts-statistics-about-epilepsy.

[Ayoubian et al., 2013] Ayoubian, L., Lacoma, H., and Gotman, J. (2013). Automatic seizure detection in SEEG using high frequency activities in wavelet domain. *Medical engineering & physics*, 35(3):319–328.

[Baars and Gage, 2012] Baars, B. J. and Gage, N. M. (2012). *Fundamentals of cognitive neuroscience: a beginner's guide*. Academic Press.

[Bagdanov et al., 2011] Bagdanov, A. D., Del Bimbo, A., and Masi, I. (2011). The florence 2D/3D hybrid face dataset. In *Proceedings of the ACM workshop on Human gesture and behavior understanding*, pages 79–80.

[Baltrusaitis et al., 2013] Baltrusaitis, T., Robinson, P., and Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 354–361.

[Baltrušaitis et al., 2014] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2014). Continuous conditional neural fields for structured regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 593–608.

[Baltrušaitis et al., 2016] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Proceeding of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10.

[Baltrušaitis et al., 2018] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2018). Using a single RGB frame for real time 3D hand pose estimation in the wild. In *Proceeding of the IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[Barron et al., 1994] Barron, J. L., Fleet, D. J., and Beauchemin, S. S. (1994). Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77.

[Bartolomei et al., 2008] Bartolomei, F., Chauvel, P., and Wendling, F. (2008). Epileptogenicity of brain structures in human temporal lobe epilepsy: a quantified study from intracerebral EEG. *Brain*, 131(7):1818–1830.

[Bas et al., 2017] Bas, A., Huber, P., Smith, W. A., Awais, M., and Kittler, J. (2017). 3D morphable models as spatial transformer networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 904–912.

[Bashivan et al., 2015] Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, page abs/1511.06448.

[Baysal-Kirac et al., 2015] Baysal-Kirac, L., Rémi, J., Loesch, A. M., Hartl, E., Vollmar, C., and Noachtar, S. (2015). Eye movements differ between ictal ipsilateral and contralateral head turning. *Epilepsy research*, 114:73–77.

[Belagiannis and Zisserman, 2017] Belagiannis, V. and Zisserman, A. (2017). Recurrent human pose estimation. *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*.

[Belhumeur et al., 2013] Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940.

[Bellantonio et al., 2016] Bellantonio, M., Haque, M. A., Rodriguez, P., Nasrollahi, K., Telve, T., Escarela, S., Gonzalez, J., Moeslund, T. B., Rasti, P., and Anbarjafari, G. (2016). Spatio-temporal pain recognition in CNN-based super-resolved facial images. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*.

[Benbadis and Hauser, 2000] Benbadis, S. R. and Hauser, W. A. (2000). An estimate of the prevalence of psychogenic non-epileptic seizures. *Seizure-European Journal of Epilepsy*, 9(4):280–281.

[Benbadis et al., 1996] Benbadis, S. R., Kotagal, P., and Klem, G. H. (1996). Unilateral blinking a lateralizing sign in partial seizures. *Neurology*, 46(1):45–48.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

[Berg et al., 2010] Berg, A. T., Berkovic, S. F., Brodie, M. J., Buchhalter, J., Cross, J. H., van Emde Boas, W., Engel, J., French, J., Glauser, T. A., Mathern, G. W., et al. (2010). Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE commission on classification and terminology, 2005–2009. *Epilepsia*, 51(4):676–685.

[Bewley et al., 2016] Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). Simple online and realtime tracking. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 3464–3468. IEEE.

[Bhagavatula et al., 2017] Bhagavatula, C., Zhu, C., Luu, K., and Savvides, M. (2017). Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 7.

[Black et al., 2000] Black, M. A., Jones, R. D., Carroll, G. J., Dingle, A. A., Donaldson, I. M., and Parkin, P. J. (2000). Real-time detection of epileptiform activity in the EEG: a blinded clinical trial. *Clinical EEG and Neuroscience*, 31(3):122–130.

[Blanz and Vetter, 1999] Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194.

[Bleasel et al., 1997] Bleasel, A., Kotagal, P., Kankirawatana, P., and Rybicki, L. (1997). Lateralizing value and semiology of ictal limb posturing and version in temporal lobe and extratemporal epilepsy. *Epilepsia*, 38(2):168–174.

[Blume et al., 2001] Blume, W. T., Lüders, H. O., Mizrahi, E., Tassinari, C., van Emde Boas, W., and Engel Jr, Ex-officio, J. (2001). Glossary of descriptive terminology for ictal semiology: report of the ILAE task force on classification and terminology. *Epilepsia*, 42(9):1212–1218.

[Boashash and Ouelha, 2016] Boashash, B. and Ouelha, S. (2016). Automatic signal abnormality detection using time-frequency features and machine learning: A newborn EEG seizure case study. *Knowledge-Based Systems*, 106:38–50.

[Bonini et al., 2014] Bonini, F., McGonigal, A., Trébuchon, A., Gavaret, M., Bartolomei, F., Giusiano, B., and Chauvel, P. (2014). Frontal lobe seizures: from clinical semiology to localization. *Epilepsia*, 55(2):264–277.

[Bourien et al., 2005] Bourien, J., Bartolomei, F., Bellanger, J., Gavaret, M., Chauvel, P., and Wendling, F. (2005). A method to identify reproducible subsets of co-activated structures during interictal spikes. Application to intracerebral EEG in temporal lobe epilepsy. *Clinical Neurophysiology*, 116(2):443–455.

[Bozhkov, 2016] Bozhkov, L. (2016). Overview of deep learning architectures for classifying brain signals. *KSI Transactions on Knowledge Society*, 9:54–59.

[Bradski et al., 2000] Bradski, G. et al. (2000). The opencv library. *Doctor Dobbs Journal*, 25(11):120–126.

[Brox et al., 2004] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 25–36.

[Bulat and Tzimiropoulos, 2016a] Bulat, A. and Tzimiropoulos, G. (2016a). Human pose estimation via convolutional part heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 717–732.

[Bulat and Tzimiropoulos, 2016b] Bulat, A. and Tzimiropoulos, G. (2016b). Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild 3DFAW challenge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 616–624.

[Bulat and Tzimiropoulos, 2017a] Bulat, A. and Tzimiropoulos, G. (2017a). Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 4.

[Bulat and Tzimiropoulos, 2017b] Bulat, A. and Tzimiropoulos, G. (2017b). How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 4.

[Burgos-Artizzu et al., 2013] Burgos-Artizzu, X. P., Perona, P., and Dollár, P. (2013). Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1513–1520.

[Burkert et al., 2015] Burkert, P., Trier, F., Afzal, M. Z., Dengel, A., and Liwicki, M. (2015). Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*, page abs/1509.05371.

[Cai et al., 2018] Cai, Y., Ge, L., Cai, J., and Yuan, J. (2018). Weakly-supervised 3D hand pose estimation from monocular RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682.

[Camps et al., 2018] Camps, J., Samà, A., Martín, M., Rodríguez-Martín, D., Pérez-López, C., Arostegui, J. M. M., Cabestany, J., Català, A., Alcaine, S., Mestre, B., et al. (2018). Deep learning for freezing of gait detection in parkinson's disease patients in their homes using a waist-worn inertial measurement unit. *Knowledge-Based Systems*, 139:119–131.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Carreira et al., 2016] Carreira, J., Agrawal, P., Fragkiadaki, K., and Malik, J. (2016). Human pose estimation with iterative error feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742.

[Cascino, 2002] Cascino, G. D. (2002). Video-EEG monitoring in adults. *Epilepsia*, 43(s3):80–93.

[Cendes, 2005] Cendes, F. (2005). Mesial temporal lobe epilepsy syndrome: an updated overview. *Journal of Epilepsy and Clinical Neurophysiology*, 11(3):141–144.

[Charles et al., 2016] Charles, J., Pfister, T., Magee, D., Hogg, D., and Zisserman, A. (2016). Personalizing human video pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3063–3072.

[Chauvel and McGonigal, 2014] Chauvel, P. and McGonigal, A. (2014). Emergence of semiology in epileptic seizures. *Epilepsy & Behavior*, 38:94–103.

[Chen and Ramanan, 2017] Chen, C.-H. and Ramanan, D. (2017). 3D human pose estimation= 2D pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 6.

[Chen et al., 2016] Chen, D., Hua, G., Wen, F., and Sun, J. (2016). Supervised transformer network for efficient face detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 122–138.

[Chen et al., 2009] Chen, L., Yang, X., Liu, Y., Zeng, D., Tang, Y., Yan, B., Lin, X., Liu, L., Xu, H., and Zhou, D. (2009). Quantitative and trajectory analysis of movement trajectories in supplementary motor area seizures of frontal lobe epilepsy. *Epilepsy & Behavior*, 14(2):344–353.

[Chollet, 2015] Chollet, F. (2015). Keras.

[Chu et al., 2014] Chu, B., Romdhani, S., and Chen, L. (2014). 3D-aided face recognition robust to expression and pose variations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1899–1906.

[Cılasun and Yalçın, 2016] Cılasun, M. H. and Yalçın, H. (2016). A deep learning approach to EEG based epilepsy seizure determination. In *Proceedings of the IEEE Conference on Signals Processing and Communication Application (ICSPCA)*, pages 1573–1576.

[Cireşan et al., 2012] Cireşan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3642–3649.

[CMU, 2017] CMU (2017). CMU graphics lab motion capture database. *Access on: http://mocap.cs.cmu.edu*.

[Cootes et al., 1994] Cootes, T. F., Taylor, C. J., and Lanitis, A. (1994). Active shape models: Evaluation of a multi-resolution method for improving image search. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 1, pages 327–336.

[Craik et al., 2019] Craik, A., He, Y., and Contreras-Vidal, J. L. P. (2019). Deep learning for electroencephalogram (eeg) classification tasks: A review. *Journal of neural engineering*.

[Cunha et al., 2016a] Cunha, J. P. S., Choupina, H. M. P., Rocha, A. P., Fernandes, J. M., Achilles, F., Loesch, A. M., Vollmar, C., Hartl, E., and Noachtar, S. (2016a). Neurokinect: a novel low-cost 3Dvideo-EEG system for epileptic seizure motion quantification. *PloS one*, 11(1):e0145669.

[Cunha et al., 2012] Cunha, J. P. S., Paula, L. M., Bento, V. F., Bilgin, C., Dias, E., and Noachtar, S. (2012). Movement quantification in epileptic seizures: a feasibility study for a new 3D approach. *Medical engineering & physics*, 34(7):938–945.

[Cunha et al., 2013] Cunha, J. P. S., Rémi, J., Vollmar, C., Fernandes, J. M., Gonzalez-Victores, J. A., and Noachtar, S. (2013). Upper limb automatisms differ quantitatively in temporal and frontal lobe epilepsies. *Epilepsy & Behavior*, 27(2):404–408.

[Cunha et al., 2016b] Cunha, J. P. S., Rocha, A. P., Choupina, H. M. P., Fernandes, J. M., Rosas, M. J., Vaz, R., Achilles, F., Loesch, A. M., Vollmar, C., Hartl, E., et al. (2016b). A novel portable, low-cost kinect-based system for motion analysis in neurological diseases. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 2339–2342.

[Cunha et al., 2009] Cunha, J. P. S., Vollmar, C., Fernandes, J., and Noachtar, S. (2009). Automated epileptic seizure type classification through quantitative movement analysis. In *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, pages 1435–1438.

[Cunha et al., 2003] Cunha, J. P. S., Vollmar, C., Li, Z., Fernandes, J., Feddersen, B., and Noachtar, S. (2003). Movement quantification during epileptic seizures: a new technical contribution to the evaluation of seizure semiology. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, volume 1, pages 671–673.

[Cuppens et al., 2012a] Cuppens, K., Bonroy, B., Van de Vel, A., Ceulemans, B., Lagae, L., Tuytelaars, T., Van Huffel, S., and Vanrumste, B. (2012a). Automatic video detection of nocturnal epileptic movement based on motion tracks. In *Proceedings of the international conference on bio-inspired systems and signal processing (Biosignals)*, pages 342–345.

[Cuppens et al., 2012b] Cuppens, K., Chen, C.-W., Wong, K. B.-Y., Van de Vel, A., Lagae, L., Ceulemans, B., Tuytelaars, T., Van Huffel, S., Vanrumste, B., and Aghajan, H. (2012b). Using spatio-temporal interest points (STIP) for myoclonic jerk detection in nocturnal video. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 4454–4457.

[Cuppens et al., 2010] Cuppens, K., Lagae, L., Ceulemans, B., Van Huffel, S., and Vanrumste, B. (2010). Automatic video detection of body movement during sleep based on optical flow in pediatric patients with epilepsy. *Medical & biological engineering & computing*, 48(9):923–931.

[Dantone et al., 2013] Dantone, M., Gall, J., Leistner, C., and Van Gool, L. (2013). Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3041–3048.

[Das et al., 2015] Das, N., Ohn-Bar, E., and Trivedi, M. M. (2015). On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 2953–2958.

[De la Torre et al., 2015] De la Torre, F., Chu, W.-S., Xiong, X., Vicente, F., Ding, X., and Cohn, J. (2015). Intraface. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, volume 1, pages 1–8.

[De Tisi et al., 2011] De Tisi, J., Bell, G. S., Peacock, J. L., McEvoy, A. W., Harkness, W. F., Sander, J. W., and Duncan, J. S. (2011). The long-term outcome of adult epilepsy surgery, patterns of seizure remission, and relapse: a cohort study. *The Lancet*, 378(9800):1388–1395.

[Deng et al., 2016] Deng, Z., Li, K., Zhao, Q., and Chen, H. (2016). Face landmark localization using a single deep network. In *Proceedings of the Chinese Conference on Biometric Recognition (CCBR)*, pages 68–76.

[Deng et al., 2017] Deng, Z., Li, K., Zhao, Q., Zhang, Y., and Chen, H. (2017). Effective face landmark localization via single deep network. *arXiv preprint arXiv:1702.02719*.

[DeTone et al., 2016] DeTone, D., Malisiewicz, T., and Rabinovich, A. (2016). Deep image homography estimation. *arXiv preprint arXiv:1606.03798*.

[Devinsky et al., 2011] Devinsky, O., Gazzola, D., and LaFrance Jr, W. C. (2011). Differentiating between nonepileptic and epileptic seizures. *Nature Reviews Neurology*, 7(4):210.

[Ding et al., 2007] Ding, L., Worrell, G. A., Lagerlund, T. D., and He, B. (2007). Ictal source analysis: localization and imaging of causal interactions in humans. *Neuroimage*, 34(2):575–586.

[Dittmar et al., 2017] Dittmar, C., Denzler, J., and Gross, H.-M. (2017). A feedback estimation approach for therapeutic facial training. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 141–148.

[do Carmo Vilas-Boas and Cunha, 2016] do Carmo Vilas-Boas, M. and Cunha, J. P. S. (2016). Movement quantification in neurological diseases: Methods and applications. *IEEE reviews in biomedical engineering*, 9:15–31.

[Donahue et al., 2015] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634.

[Dou et al., 2017] Dou, P., Shah, S. K., and Kakadiaris, I. A. (2017). End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–26.

[Drover et al., 2018] Drover, D., Chen, C.-H., Agrawal, A., Tyagi, A., and Huynh, C. P. (2018). Can 3D pose be learned from 2D projections alone? *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*.

[Dumas et al., 2009] Dumas, B., Lalanne, D., and Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In *Human machine interaction*, pages 3–26.

[Egger et al., 2014] Egger, B., Schönborn, S., Forster, A., and Vetter, T. (2014). Pose normalization for eye gaze estimation and facial attribute description from still images. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, pages 317–327.

[Esteva et al., 2017] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115.

[Evangelidis and Psarakis, 2008] Evangelidis, G. D. and Psarakis, E. Z. (2008). Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865.

[Fang et al., 2018] Fang, Z., Leung, H., and Choy, C. S. (2018). Spatial temporal GRU convnets for vision-based real time epileptic seizure detection. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1026–1029.

[Farfade et al., 2015] Farfade, S. S., Saberian, M. J., and Li, L.-J. (2015). Multi-view face detection using deep convolutional neural networks. In *Proceeding of the ACM International Conference on Multimedia Retrieval (ICMR)*, pages 643–650.

[Faust et al., 2018] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: a review. *Computer methods and programs in biomedicine*.

[Felzenszwalb and Huttenlocher, 2012] Felzenszwalb, P. F. and Huttenlocher, D. P. (2012). Distance transforms of sampled functions. *Theory of computing*, 8(1):415–428.

[Feng et al., 2018a] Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018a). Joint 3D face reconstruction and dense alignment with position map regression network. *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Feng et al., 2018b] Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018b). Joint 3D face reconstruction and dense alignment with position map regression network (PRNet). https://github.com/YadiraF/PRNet.

[Feng et al., 2017] Feng, Z.-H., Kittler, J., Awais, M., Huber, P., and Wu, X. (2017). Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 2106–2115.

[Fergus et al., 2016] Fergus, P., Hussain, A., Hignett, D., Al-Jumeily, D., Abdel-Aziz, K., and Hamdan, H. (2016). A machine learning system for automated whole-brain seizure detection. *Applied Computing and Informatics*, 12(1):70–89.

[Fernandes et al., 2005] Fernandes, J. M., da Silva, A. M., Huiskamp, G., Velis, D. N., Manshanden, I., de Munck, J. C., da Silva, F. L., and Cunha, J. P. S. (2005). What does an epileptiform spike look like in MEG? comparison between coincident EEG and MEG spikes. *Journal of clinical neurophysiology*, 22(1):68–73.

[Fernando et al., 2018] Fernando, T., Denman, S., Sridharan, S., and Fookes, C. (2018). Tracking by prediction: A deep generative model for mutli-person localisation and tracking. In *Proceeding of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1122–1132.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

[Fisher et al., 2005] Fisher, R., Boas, W., Blume, W., Elger, C., Genton, P., Lee, P., and Engel, J. (2005). Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the international bureau for epilepsy (IBE). *Epilepsia*, 46(4):470–472.

[Fogarasi et al., 2007] Fogarasi, A., Tuxhorn, I., Janszky, J., Janszky, I., Rásonyi, G., Kelemen, A., and Halász, P. (2007). Age-dependent seizure semiology in temporal lobe epilepsy. *Epilepsia*, 48(9):1697–1702.

[Fulop and Fitz, 2006] Fulop, S. A. and Fitz, K. (2006). Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America*, 119(1):360–371.

[Gajic et al., 2014] Gajic, D., Djurovic, Z., Di Gennaro, S., and Gustafsson, F. (2014). Classification of EEG signals for detection of epileptic seizures based on wavelets and statistical pattern recognition. *Biomedical Engineering: Applications, Basis and Communications*, 26(02):1450021.

[Gammulle et al., 2017] Gammulle, H., Denman, S., Sridharan, S., and Fookes, C. (2017). Two stream LSTM: A deep fusion framework for human action recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 177–186.

[Gavaret et al., 2009] Gavaret, M., Trébuchon, A., Bartolomei, F., Marquis, P., Mcgonigal, A., Wendling, F., Regis, J., Badier, J.-M., and Chauvel, P. (2009). Source localization of scalp-EEG interictal spikes in posterior cortex epilepsies investigated by HR-EEG and SEEG. *Epilepsia*, 50(2):276–289.

[Gelziniene et al., 2008] Gelziniene, G., Endziniene, M., Vaiciene, N., Magistris, M., and Seeck, M. (2008). Presurgical evaluation of epilepsy patients. *Medicina*, 44(8):585–592.

[Ghasemi et al., 2016] Ghasemi, A., Denman, S., Sridharan, S., and Fookes, C. (2016). Discovery of facial motions using deep machine perception. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–7.

[Ghosh-Dastidar et al., 2007] Ghosh-Dastidar, S., Adeli, H., and Dadmehr, N. (2007). Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. *IEEE transactions on biomedical engineering*, 54(9):1545–1551.

[Gil-Nagel and Risinger, 1997] Gil-Nagel, A. and Risinger, M. W. (1997). Ictal semiology in hippocampal versus extrahippocampal temporal lobe epilepsy. *Brain*, 120(1):183–192.

[Girdhar, 2018] Girdhar, R. (2018). Detect-and-track: Efficient pose estimation in videos. https://github.com/facebookresearch/DetectAndTrack.

[Girdhar et al., 2018] Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., and Tran, D. (2018). Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 350–359.

[Girshick, 2015] Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

[Girshick et al., 2014] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587.

[Gkioxari et al., 2016] Gkioxari, G., Toshev, A., and Jaitly, N. (2016). Chained predictions using convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 728–743.

[Goldberger et al., 2000] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning (adaptive computation and machine learning series). *Adaptive Computation and Machine Learning series*, page 800.

[Gou et al., 2016] Gou, C., Wu, Y., Wang, F.-Y., and Ji, Q. (2016). Shape augmented regression for 3D face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 604–615.

[Greff et al., 2017] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.

[Gross et al., 2010] Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and Vision Computing*, 28(5):807–813.

[Grouiller et al., 2011] Grouiller, F., Thornton, R. C., Groening, K., Spinelli, L., Duncan, J. S., Schaller, K., Siniatchkin, M., Lemieux, L., Seeck, M., Michel, C. M., et al. (2011). With or without spikes: localization of focal epileptic activity by simultaneous electroencephalography and functional magnetic resonance imaging. *Brain*, 134(10):2867–2886.

[Gubbi et al., 2016] Gubbi, J., Kusmakar, S., Rao, A. S., Yan, B., O'Brien, T., and Palaniswami, M. (2016). Automatic detection and classification of convulsive psychogenic nonepileptic seizures using a wearable device. *IEEE journal of biomedical and health informatics*, 20(4):1061–1072.

[Guide, 1998] Guide, M. U. (1998). The mathworks. *Inc., Natick, MA*, 5:333.

[Güler et al., 2018] Güler, R. A., Neverova, N., and Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Güler et al., 2017] Güler, R. A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., and Kokkinos, I. (2017). Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 5.

[Guo et al., 2017] Guo, Z., Shen, M., Duan, L., Zhou, Y., Xiang, J., Ding, H., Chen, S., Deussen, O., and Dan, G. (2017). Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 135–138.

[Hasani and Mahoor, 2017] Hasani, B. and Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3D convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2278–2288.

[Hassner, 2013] Hassner, T. (2013). Viewing real-world faces in 3D. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3607–3614.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

[He et al., 2014] He, K., Zhang, X., Ren, S., and Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 346–361.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Henkel et al., 2000] Henkel, A., Winkler, P., and Noachtar, S. (2000). Ipsilateral blinking: a rare lateralizing seizure phenomenon in temporal lobe epilepsy. *Epileptic disorders*, 1(3):195–8.

[Hidalgo et al., 2018] Hidalgo, G., Cao, Z., Simon, T., Wei, S.-E., Joo, H., and Y, S. (2018). Openpose. https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[Hinton et al., 2012] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

[Holland, 2018] Holland, K. (2018). Epilepsy: Facts, statistics, and you. https://www.healthline.com/health/epilepsy/facts-statistics-infographic.

[Holte et al., 2012] Holte, M. B., Tran, C., Trivedi, M. M., and Moeslund, T. B. (2012). Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of selected topics in signal processing*, 6(5):538–552.

[Horn and Schunck, 1981] Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.

[Hu, 2018] Hu, P. (2018). Tiny face detector. https://github.com/peiyunh/tiny.

[Hu and Ramanan, 2017] Hu, P. and Ramanan, D. (2017). Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530.

[Hu et al., 2004] Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352.

[Huber et al., 2016] Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W. J., Ratsch, M., and Kittler, J. (2016). A multiresolution 3D morphable face model and fitting framework. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*.

[Hunyadi et al., 2013] Hunyadi, B., Tousseyn, S., Mijović, B., Dupont, P., Van Huffel, S., Van Paesschen, W., and De Vos, M. (2013). Ica extracts epileptic sources from *f*MRI in EEG-negative patients: a retrospective validation study. *PloS one*, 8(11):e78796.

[Husain and Rao, 2012] Husain, S. J. and Rao, K. S. (2012). Epileptic seizures classification from EEG signals using neural networks. In *Proceeding of the International Conference on Information and Network Technologies (ICINT)*, volume 37, pages 269–273.

[Ilg et al., 2017] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

[Ilyas et al., 2018] Ilyas, C. M. A., Nasrollahi, K., Moeslund, T. B., Rehm, M., and Haque, M. A. (2018). Facial expression recognition for traumatic brain injured patients. *SCITEPRESS Digital Library*, page 1.

[Insafutdinov et al., 2017] Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., and Schiele, B. (2017). Arttrack: Articulated multi-person tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Insafutdinov et al., 2016] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 34–50.

[Ionescu et al., 2014] Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339.

[Iqbal et al., 2018a] Iqbal, U., Doering, A., Yasin, H., Krüger, B., Weber, A., and Gall, J. (2018a). A dual-source approach for 3D human pose estimation from single images. *Computer Vision and Image Understanding*.

[Iqbal et al., 2017a] Iqbal, U., Garbade, M., and Gall, J. (2017a). Pose for action-action for pose. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 438–445.

[Iqbal et al., 2017b] Iqbal, U., Milan, A., and Gall, J. (2017b). Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Iqbal et al., 2018b] Iqbal, U., Molchanov, P., Breuel Juergen Gall, T., and Kautz, J. (2018b). Hand pose estimation via latent 2.5 D heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134.

[Jackson et al., 2017] Jackson, A. S., Bulat, A., Argyriou, V., and Tzimiropoulos, G. (2017). Large pose 3D face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1031–1039.

[Jain and Learned-Miller, 2010] Jain, V. and Learned-Miller, E. G. (2010). FDDB: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*.

[Jaiswal and Valstar, 2016] Jaiswal, S. and Valstar, M. (2016). Deep learning the dynamic appearance and shape of facial action units. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8.

[Jeni et al., 2015] Jeni, L. A., Cohn, J. F., and Kanade, T. (2015). Dense 3D face alignment from 2D videos in real-time. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, volume 1, pages 1–8.

[Jeni et al., 2016] Jeni, L. A., Tulyakov, S., Yin, L., Sebe, N., and Cohn, J. F. (2016). The first 3D face alignment in the wild 3DFAW challenge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 511–520.

[Jhuang et al., 2013] Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M. J. (2013). Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3192–3199.

[Ji et al., 2013] Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

[Jia et al., 2014] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678.

[Jiang and Learned-Miller, 2017] Jiang, H. and Learned-Miller, E. (2017). Face detection with the faster R-CNN. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 650–657.

[Jin et al., 2017] Jin, S.-Y., Su, H., Stauffer, C., and Learned-Miller, E. G. (2017). End-to-end face detection and cast grouping in movies using erdös-rényi clustering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5286–5295.

[Jin and Tan, 2017] Jin, X. and Tan, X. (2017). Face alignment in-the-wild: A survey. *Computer Vision and Image Understanding*, 162:1–22.

[Jirayucharoensak et al., 2014] Jirayucharoensak, S., Pan-Ngum, S., and Israsena, P. (2014). EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014.

[Jobst et al., 2000] Jobst, B. C., Siegel, A. M., Thadani, V. M., Roberts, D. W., Rhodes, H. C., and Williamson, P. D. (2000). Intractable seizures of frontal lobe origin: clinical characteristics, localizing signs, and results of surgery. *Epilepsia*, 41(9):1139–1152.

[Johansen et al., 2016] Johansen, A. R., Jin, J., Maszczyk, T., Dauwels, J., Cash, S. S., and Westover, M. B. (2016). Epileptiform spike detection via convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 754–758. IEEE.

[Johnson and Everingham, 2011] Johnson, S. and Everingham, M. (2011). Learning effective human pose estimation from inaccurate annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1465–1472.

[Joo et al., 2018] Joo, H., Simon, T., and Sheikh, Y. (2018). Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329.

[Jourabloo and Liu, 2015] Jourabloo, A. and Liu, X. (2015). Pose-invariant 3D face alignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3694–3702.

[Jourabloo and Liu, 2016] Jourabloo, A. and Liu, X. (2016). Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4188–4196.

[Jourabloo and Liu, 2017] Jourabloo, A. and Liu, X. (2017). Pose-invariant face alignment via CNN-based dense 3D model fitting. *International Journal of Computer Vision*, 124(2):187–203.

[Juarez-Guerra et al., 2015] Juarez-Guerra, E., Alarcon-Aquino, V., and Gomez-Gil, P. (2015). Epilepsy seizure detection in EEG signals using wavelet transforms and neural networks. *Springer International Publishing*, pages 261–269.

[Kalitzin et al., 2012] Kalitzin, S., Petkov, G., Velis, D., Vledder, B., and da Silva, F. L. (2012). Automatic segmentation of episodes containing epileptic clonic seizures in video sequences. *IEEE Transactions on Biomedical Engineering*, 59(12):3379–3385.

[Kanazawa et al., 2018] Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J. (2018). End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131.

[Karayiannis and Tao, 2003] Karayiannis, N. B. and Tao, G. (2003). Extraction of temporal motion velocity signals from video recordings of neonatal seizures by optical flow methods. *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 1:874–877.

[Karayiannis et al., 2006] Karayiannis, N. B., Tao, G., Frost, J. D., Wise, M. S., Hrachovy, R. A., and Mizrahi, E. M. (2006). Automated detection of videotaped neonatal seizures based on motion segmentation methods. *Clinical Neurophysiology*, 117(7):1585–1594.

[Karpathy et al., 2014] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874.

[Kibler and Durand, 2011] Kibler, A. B. and Durand, D. M. (2011). Orthogonal wave propagation of epileptiform activity in the planar mouse hippocampus in vitro. *Epilepsia*, 52(9):1590–1600.

[Kim et al., 2017] Kim, D. H., Baddar, W., Jang, J., and Ro, Y. M. (2017). Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*.

[King, 2012] King, D. (2012). Dlib c++ library. *Access on: http://dlib. net*.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Klare et al., 2015] Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., and Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939.

[Koessler et al., 2010] Koessler, L., Benar, C., Maillard, L., Badier, J.-M., Vignal, J. P., Bartolomei, F., Chauvel, P., and Gavaret, M. (2010). Source localization of ictal epileptic activity investigated by high resolution EEG and validated by SEEG. *Neuroimage*, 51(2):642–653.

[Kohavi et al., 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145.

[Köstinger et al., 2011] Köstinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2144–2151.

[Kotagal et al., 1995] Kotagal, P., Lüders, H. O., Williams, G., Nichols, T. R., and McPherson, J. (1995). Psychomotor seizures of temporal lobe onset: analysis of symptom clusters and sequences. *Epilepsy research*, 20(1):49–67.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the Neural Information Processing Systems (NIPS)*, pages 1097–1105.

[Kuhn, 1955] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

[Kuhner et al., 2017] Kuhner, A., Schubert, T., Maurer, C., and Burgard, W. (2017). An online system for tracking the performance of parkinson's patients. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1664–1669.

[Kusmakar et al., 2016] Kusmakar, S., Muthuganapathy, R., Yan, B., O'Brien, T. J., and Palaniswami, M. (2016). Gaussian mixture model for the identification of psychogenic non-epileptic seizures using a wearable accelerometer sensor. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 1006–1009.

[Längkvist et al., 2014] Längkvist, M., Karlsson, L., and Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24.

[Laptev, 2005] Laptev, I. (2005). On space-time interest points. *International journal of computer vision*, 64(2-3):107–123.

[Le et al., 2017] Le, T. H. N., Quach, K. G., Zhu, C., Duong, C. N., Luu, K., Savvides, M., and Center, C. B. (2017). Robust hand detection and classification in vehicles and in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1203–1210.

[Le et al., 2016] Le, T. H. N., Zhu, C., Zheng, Y., Luu, K., and Savvides, M. (2016). Robust hand detection in vehicles. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 573–578.

[Le et al., 2012] Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. (2012). Interactive facial feature localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 679–692.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324.

[Lee et al., 1985] Lee, H.-J., Zen, C., et al. (1985). Determination of 3D human-body postures from a single view. *Computer Vision Graphics and Image Processing*, 30(2):148–168.

[Lee et al., 2017] Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., and Kim, N. (2017). Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584.

[Leung et al., 2008] Leung, H., Schindler, K., Clusmann, H., Bien, C. G., Pöpel, A., Schramm, J., Kwan, P., Wong, L. K., and Elger, C. E. (2008). Mesial frontal epilepsy and ictal body turning along the horizontal body axis. *Archives of neurology*, 65(1):71–77.

[Li et al., 2015] Li, H., Lin, Z., Shen, X., Brandt, J., and Hua, G. (2015). A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5325–5334.

[Li et al., 2017] Li, M. H., Mestre, T. A., Fox, S. H., and Taati, B. (2017). Automated vision-based analysis of levodopa-induced dyskinesia with deep learning. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 3377–3380.

[Li et al., 2018a] Li, M. H., Mestre, T. A., Fox, S. H., and Taati, B. (2018a). Automated assessment of levodopa-induced dyskinesia: Evaluating the responsiveness of video-based features. *Parkinsonism & related disorders*, 53:42–45.

[Li and Chan, 2014] Li, S. and Chan, A. B. (2014). 3D human pose estimation from monocular images with deep convolutional neural network. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 332–347.

[Li and Deng, 2018] Li, S. and Deng, W. (2018). Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*.

[Li et al., 2018b] Li, Y., Wang, X.-D., Luo, M.-L., Li, K., Yang, X.-F., and Guo, Q. (2018b). Epileptic seizure classification of EEGs using time–frequency analysis based multiscale radial basis functions. *IEEE journal of biomedical and health informatics*, 22(2):386–397.

[Li et al., 2002] Li, Z., da Silva, A. M., and Cunha, J. P. S. (2002). Movement quantification in epileptic seizures: a new approach to video-EEG analysis. *IEEE Transactions on Biomedical Engineering*, 49(6):565–573.

[Lin and Hung, 2018] Lin, C.-C. and Hung, Y. (2018). A prior-less method for multi-face tracking in unconstrained videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–547.

[Lin et al., 2016] Lin, Q., Ye, S.-q., Huang, X.-m., Li, S.-y., Zhang, M.-z., Xue, Y., and Chen, W.-S. (2016). Classification of epileptic EEG signals with stacked sparse autoencoder based on deep learning. In *Proceedings of the International Conference on Intelligent Computing (ICIC)*, pages 802–810.

[Lin et al., 2017] Lin, T.-Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., and Belongie, S. J. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755.

[Lipton et al., 2015] Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

[Litjens et al., 2017] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

[Liu et al., 2016] Liu, F., Zeng, D., Zhao, Q., and Liu, X. (2016). Joint face alignment and 3D face reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 545–560.

[Liu et al., 2013] Liu, M., Li, S., Shan, S., and Chen, X. (2013). Au-aware deep networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 1–6.

[Liu et al., 2014a] Liu, M., Li, S., Shan, S., Wang, R., and Chen, X. (2014a). Deeply learning deformable facial action parts model for dynamic expression analysis. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 143–157.

[Liu et al., 2014b] Liu, P., Han, S., Meng, Z., and Tong, Y. (2014b). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1812.

[Liu et al., 2017] Liu, Y., Jourabloo, A., Ren, W., and Liu, X. (2017). Dense face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1619–1628.

[Loddenkemper and Kotagal, 2005] Loddenkemper, T. and Kotagal, P. (2005). Lateralizing signs during seizures in focal epilepsy. *Epilepsy & Behavior*, 7(1):1–17.

[Lopes et al., 2017] Lopes, A. T., de Aguiar, E., De Souza, A. F., and Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628.

[Lu et al., 2013] Lu, H., Pan, Y., Mandal, B., Eng, H.-L., Guan, C., and Chan, D. W. (2013). Quantifying limb movements in epileptic seizures through color-based video analysis. *IEEE Transactions on Biomedical Engineering*, 60(2):461–469.

[Lucas and Kanade, 1981] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679.

[Lüders et al., 1998] Lüders, H., Acharya, J., Baumgartner, C., Benbadis, S., Bleasel, A., Burgess, R., Dinner, D., Ebner, A., Foldvary, N., Geller, E., et al. (1998). Semiological seizure classification. *Epilepsia*, 39(9):1006–1013.

[Lun and Zhao, 2015] Lun, R. and Zhao, W. (2015). A survey of applications and human motion recognition with microsoft kinect. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(05):1555008.

[Luo et al., 2018] Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., and Lin, L. (2018). LSTM pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Ma et al., 2015] Ma, L., Minett, J. W., Blu, T., and Wang, W. S. (2015). Resting state EEG-based biometrics for individual identification using convolutional neural networks. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 2848–2851.

[Maia et al., 2019] Maia, P., Hartl, E., Vollmar, C., Noachtar, S., and Cunha, J. P. S. (2019). Epileptic seizure classification using the neuromov database. In *Proceedings of the IEEE Portuguese Meeting on Bioengineering (ENBENG)*, pages 1–4. IEEE.

[Mandal et al., 2012] Mandal, B., Eng, H.-L., Lu, H., Chan, D. W., and Ng, Y.-L. (2012). Non-intrusive head movement analysis of videotaped seizures of epileptic origin. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 6060–6063.

[Mathias et al., 2014] Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L. (2014). Face detection without bells and whistles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–735.

[Maurel et al., 2008] Maurel, P., Mcgonigal, A., Keriven, R., and Chauvel, P. (2008). 3D model fitting for facial expression analysis under uncontrolled imaging conditions. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 1–4.

[McGonigal and Chauvel, 2004] McGonigal, A. and Chauvel, P. (2004). Frontal lobe epilepsy: seizure semiology and presurgical evaluation. *Practical Neurology*, 4(5):260–273.

[McKee et al., 2018] McKee, R., McKee, D., and Alexander, D. (2018). NZ sign language exercise. *Deaf Studies Department Deaf Studies Department at the School of Linguistics and Applied Language Studies of Victoria University of Wellington. Access on: https://www.victoria.ac.nz*.

[Mehta et al., 2017] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., and Theobalt, C. (2017). Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 506–516.

[Mehta et al., 2018] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., and Theobalt, C. (2018). Single-shot multi-person 3D pose estimation from monocular RGB. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 120–130.

[Meier et al., 2004] Meier, A., Cunha, J., Mauerer, C., Vollmar, C., Feddersen, B., and Noachtar, S. (2004). Quantified analysis of wrist and trunk movements differentiates between hypermotor and automotor seizures. *Klinische Neurophysiologie*, 35(03):178.

[Mierlo et al., 2013] Mierlo, P., Carrette, E., Hallez, H., Raedt, R., Meurs, A., Vandenberghe, S., Roost, D., Boon, P., Staelens, S., and Vonck, K. (2013). Ictal-onset localization through connectivity analysis of intracranial EEG signals in patients with refractory epilepsy. *Epilepsia*, 54(8):1409–1418.

[Mirzadjanova et al., 2010] Mirzadjanova, Z., Peters, A. S., Rémi, J., Bilgin, C., Silva Cunha, J. P., and Noachtar, S. (2010). Significance of lateralization of upper limb automatisms in temporal lobe epilepsy: a quantitative movement analysis. *Epilepsia*, 51(10):2140–2146.

[Mittal et al., 2011] Mittal, A., Zisserman, A., and Torr, P. H. (2011). Hand detection using multiple proposals. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11.

[Miyazaki et al., 2000] Miyazaki, S., Ishida, A., and Komatsuzaki, A. (2000). A clinically oriented video-based system for quantification of eyelid movements. *IEEE transactions on biomedical engineering*, 47(8):1088–1096.

[Moreno-Noguer, 2017] Moreno-Noguer, F. (2017). 3D human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1561–1570.

[Mueller et al., 2018] Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., and Theobalt, C. (2018). Ganerated hands for real-time 3D hand tracking from monocular RGB. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Mueller et al., 2017] Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., and Theobalt, C. (2017). Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *Proceedings of International Conference on Computer Vision (ICCV)*, volume 10.

[Naghavi and Wade, 2018] Naghavi, N. and Wade, E. (2018). Design of a paradigm to elicit gait-related symptoms of parkinson's disease: A case study. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*.

[Nasehi and Pourghassem, 2013] Nasehi, S. and Pourghassem, H. (2013). Patient-specific epileptic seizure onset detection algorithm based on spectral features and ipsonn classifier. In *Proceedings of the International Conference on Communication Systems and Network Technologies (CSNT)*, pages 186–190.

[Neverova et al., 2014] Neverova, N., Wolf, C., Taylor, G. W., and Nebout, F. (2014). Multi-scale deep learning for gesture detection and localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 474–490.

[Newell et al., 2016] Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 483–499.

[Nguyen et al., 2017] Nguyen, D., Nguyen, K., Sridharan, S., Ghasemi, A., Dean, D., and Fookes, C. (2017). Deep spatio-temporal features for multimodal emotion recognition. In *Proceeding of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1223.

[Nigam and Graupe, 2004] Nigam, V. P. and Graupe, D. (2004). A neural-network-based detection of epilepsy. *Neurological research*, 26(1):55–60.

[Noachtar, 2003] Noachtar, S. (2003). Video analysis for defining the symptomatogenic zone. *Handbook of Clinical Neurophysiology*, 3:187–200.

[Noachtar and Peters, 2009] Noachtar, S. and Peters, A. S. (2009). Semiology of epileptic seizures: a critical review. *Epilepsy & Behavior*, 15(1):2–9.

[Noachtar et al., 2003] Noachtar, S., Winkler, P. A., and Lüders, H. O. (2003). Surgical therapy of epilepsy. *Neurological disorders: course and treatment*, pages 235–244.

[Oberweger et al., 2015a] Oberweger, M., Wohlhart, P., and Lepetit, V. (2015a). Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*.

[Oberweger et al., 2015b] Oberweger, M., Wohlhart, P., and Lepetit, V. (2015b). Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3316–3324.

[O'Brien et al., 2008] O'Brien, T. J., Mosewich, R. K., Britton, J. W., Cascino, G. D., and So, E. L. (2008). History and seizure semiology in distinguishing frontal lobe seizures and temporal lobe seizures. *Epilepsy research*, 82(2-3):177–182.

[O'Dwyer et al., 2004] O'Dwyer, R., Cunha, J., Vollmar, C., Mauerer, C., Ebner, A., Feddersen, B., and Noachtar, S. (2004). Quantification of ipsilateral and contralateral head movements during seizures in patients with temporal lobe epilepsy. *Klinische Neurophysiologie*, 35(03):204.

[O'Dwyer et al., 2007] O'Dwyer, R., Silva Cunha, J. P., Vollmar, C., Mauerer, C., Feddersen, B., Burgess, R. C., Ebner, A., and Noachtar, S. (2007). Lateralizing significance of quantitative analysis of head movements before secondary generalization of seizures of patients with temporal lobe epilepsy. *Epilepsia*, 48(3):524–530.

[O'Muircheartaigh and Richardson, 2012] O'Muircheartaigh, J. and Richardson, M. P. (2012). Epilepsy and the frontal lobes. *Cortex*, 48(2):144–155.

[Orlandi et al., 2018] Orlandi, S., Raghuram, K., Smith, C., Mansueto, D., Church, P., Shah, V., Luther, M., and Chau, T. (2018). Detection of atypical and typical infant movements using computer-based video analysis. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*.

[Orosco et al., 2013] Orosco, L., Correa, A. G., and Laciar, E. (2013). A survey of performance and techniques for automatic epilepsy detection. *Journal of Medical and Biological Engineering*, 33(6):526–537.

[Oviatt, 2003] Oviatt, S. (2003). Advances in robust multimodal interface design. *IEEE computer graphics and applications*, 23(5):62–68.

[Owens and Efros, 2018] Owens, A. and Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Page et al., 2016] Page, A., Shea, C., and Mohsenin, T. (2016). Wearable seizure detection using convolutional neural networks with transfer learning. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1086–1089.

[Pasupa and Sunhem, 2016] Pasupa, K. and Sunhem, W. (2016). A comparison between shallow and deep architecture classifiers on small dataset. In *Proceedings of the International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–6.

[Pathak et al., 2017] Pathak, D., Girshick, R., Dollár, P., Darrell, T., and Hariharan, B. (2017). Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Pavlakos et al., 2017] Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272.

[Paysan et al., 2009] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *Proceedings of the IEEE International Conference on Advanced video and signal based surveillance (AVSS)*, pages 296–301.

[Pediaditis et al., 2011] Pediaditis, M., Tsiknakis, M., Bologna, V., and Vorgia, P. (2011). Model-free vision-based facial motion analysis in epilepsy. In *Proceedings of the International Workshop on Biomedical Engineering*, pages 1–4.

[Pediaditis et al., 2012a] Pediaditis, M., Tsiknakis, M., Koumakis, L., Karachaliou, M., Voutoufianakis, S., and Vorgia, P. (2012a). Vision-based absence seizure detection. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, pages 65–68.

[Pediaditis et al., 2012b] Pediaditis, M., Tsiknakis, M., and Leitgeb, N. (2012b). Vision-based motion detection, analysis and recognition of epileptic seizures—a systematic review. *Computer methods and programs in biomedicine*, 108(3):1133–1148.

[Pemasiri et al., 2019a] Pemasiri, A., Ahmedt-Aristizabal, D., Nguyen, K., Sridharan, S., Dionisio, S., and Fookes, C. (2019a). A hierarchical multi-modal system for motion analysis in epileptic patients. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*.

[Pemasiri et al., 2019b] Pemasiri, A., Nguyen, K., Sridharan, S., and Fookes, C. ([Submitted] 2019b). VX-Mask R-CNN: Multi-modal semantic segmentation of human body parts. *IEEE Journal of Biomedical and Health Informatics*.

[Pereira et al., 2018] Pereira, H. C., Rocha, A., Fernandes, J., Vollmar, C., Noachtar, S., and Silva, J. C. (2018). Neurokinect 3.0: Multi-bed 3Dvideo-EEG system for epilepsy clinical motion monitoring. *Studies in health technology and informatics*, 247:46–50.

[Pfänder et al., 2002] Pfänder, M., Arnold, S., Henkel, A., Weil, S., and Noachtar, S. (2002). findings differentiating mesial from neocortical temporal. *Epileptic Disorders*, 4(3):189–95.

[Pfister et al., 2015] Pfister, T., Charles, J., and Zisserman, A. (2015). Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1913–1921.

[Pishchulin et al., 2016] Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., and Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937.

[Plis et al., 2014] Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., Johnson, H. J., Paulsen, J. S., Turner, J. A., and Calhoun, V. D. (2014). Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229.

[Pons-Moll and Rosenhahn, 2011] Pons-Moll, G. and Rosenhahn, B. (2011). Model-based pose estimation. In *Visual analysis of humans*, pages 139–170.

[Poppe, 2007] Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer vision and image understanding*, 108(1):4–18.

[Qu et al., 2015] Qu, C., Monari, E., Schuchert, T., and Beyerer, J. (2015). Adaptive contour fitting for pose-invariant 3D face shape reconstruction. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[Qu and Gotman, 1997] Qu, H. and Gotman, J. (1997). A patient-specific algorithm for the detection of seizure onset in long-term EEG monitoring: possible use as a warning device. *IEEE transactions on biomedical engineering*, 44(2):115–122.

[Quintero-Rincón et al., 2016] Quintero-Rincón, A., Pereyra, M., D'Giano, C., Batatia, H., and Risk, M. (2016). A new algorithm for epilepsy seizure onset detection and spread estimation from EEG signals. *Journal of Physics: Conference Series*, 705(1):012032.

[Ramabhadran et al., 1999] Ramabhadran, B., Frost Jr, J. D., Glover, J. R., and Ktonas, P. Y. (1999). An automated system for epileptogenic focus localization in the electroencephalogram. *Journal of clinical neurophysiology*, 16(1):59–68.

[Ramakrishna et al., 2012] Ramakrishna, V., Kanade, T., and Sheikh, Y. (2012). Reconstructing 3D human pose from 2D image landmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 573–586.

[Ramakrishna et al., 2014] Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J. A., and Sheikh, Y. (2014). Pose machines: Articulated pose estimation via inference machines. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 33–47.

[Ramgopal et al., 2014] Ramgopal, S., Thome-Souza, S., Jackson, M., Kadish, N. E., Fernández, I. S., Klehm, J., Bosl, W., Reinsberger, C., Schachter, S., and Loddenkemper, T. (2014). Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. *Epilepsy & behavior*, 37:291–307.

[Ranjan et al., 2015] Ranjan, R., Patel, V. M., and Chellappa, R. (2015). A deep pyramid deformable part model for face detection. In *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, pages 1–8.

[Ranjan et al., 2019] Ranjan, R., Patel, V. M., and Chellappa, R. (2019). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135.

[Ranjan et al., 2017] Ranjan, R., Sankaranarayanan, S., Castillo, C. D., and Chellappa, R. (2017). An all-in-one convolutional neural network for face analysis. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 17–24.

[Razzak et al., 2018] Razzak, M. I., Naz, S., and Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. In *Classification in BioApps*, pages 323–350.

[Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

[Redmon and Farhadi, 2017] Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Rémi et al., 2011a] Rémi, J., Cunha, J. P. S., Vollmar, C., Topçuoğlu, Ö. B., Meier, A., Ulowetz, S., Beleza, P., and Noachtar, S. (2011a). Quantitative movement analysis differentiates focal seizures characterized by automatisms. *Epilepsy & Behavior*, 20(4):642–647.

[Rémi et al., 2011b] Rémi, J., Wagner, P., O'Dwyer, R., Silva Cunha, J. P., Vollmar, C., Krotofil, I., and Noachtar, S. (2011b). Ictal head turning in frontal and temporal lobe epilepsy. *Epilepsia*, 52(8):1447–1451.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Neural Information Processing Systems (NIPS)*.

[Ren et al., 2017] Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, (6):1137–1149.

[Ridley, 1994] Ridley, R. (1994). The psychology of perseverative and stereotyped behaviour. *Progress in neurobiology*, 44(2):221–231.

[Rodrigues et al., 2018] Rodrigues, J., Maia, P., Choupina, H. M. P., and Cunha, J. P. S. (2018). On the fly reporting of human body movement based on kinect V2. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*.

[Rodriguez et al., 2017] Rodriguez, P., Cucurull, G., Gonzàlez, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B., and Roca, F. X. (2017). Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics*, (99):1–11.

[Rogez and Schmid, 2016] Rogez, G. and Schmid, C. (2016). Mocap-guided data augmentation for 3D pose estimation in the wild. In *Advances in Neural Information Processing Systems*, pages 3108–3116.

[Rohrbach et al., 2012] Rohrbach, M., Amin, S., Andriluka, M., and Schiele, B. (2012). A database for fine grained activity detection of cooking activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1194–1201.

[Romdhani and Vetter, 2005] Romdhani, S. and Vetter, T. (2005). Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 986–993.

[Rosenow and Lüders, 2001] Rosenow, F. and Lüders, H. (2001). Presurgical evaluation of epilepsy. *Brain*, 124(9):1683–1700.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

[Ryvlin et al., 2006] Ryvlin, P., Minotti, L., Demarquay, G., Hirsch, E., Arzimanoglou, A., Hoffman, D., Guénot, M., Picard, F., Rheims, S., and Kahane, P. (2006). Nocturnal hypermotor seizures, suggesting frontal lobe epilepsy, can originate in the insula. *Epilepsia*, 47(4):755–765.

[Sagonas et al., 2016] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18.

[Sagonas et al., 2013a] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013a). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 397–403.

[Sagonas et al., 2013b] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013b). A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 896–903.

[Salam and Séguier, 2018] Salam, H. and Séguier, R. (2018). A survey on face modeling: building a bridge between face analysis and synthesis. *The Visual Computer*, 34(2):289–319.

[Sanzari et al., 2016] Sanzari, M., Ntouskos, V., and Pirri, F. (2016). Bayesian image based 3D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 566–582.

[Sapp and Taskar, 2013] Sapp, B. and Taskar, B. (2013). Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3681.

[Sapp et al., 2011] Sapp, B., Weiss, D., and Taskar, B. (2011). Parsing human motion with stretchable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1281–1288.

[Sarafianos et al., 2016] Sarafianos, N., Boteanu, B., Ionescu, B., and Kakadiaris, I. A. (2016). 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20.

[Satapathy et al., 2016] Satapathy, S. K., Dehuri, S., and Jagadev, A. K. (2016). An empirical analysis of different machine learning techniques for classification of EEG signal to detect epileptic seizure. *International Journal of Applied Engineering Research*, 11(1):120–129.

[Sathyanarayana et al., 2015] Sathyanarayana, S., Satzoda, R. K., Sathyanarayana, S., and Thambipillai, S. (2015). Identifying epileptic seizures based on a template-based eyeball detection technique. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 4689–4693.

[Sathyanarayana et al., 2018] Sathyanarayana, S., Satzoda, R. K., Sathyanarayana, S., and Thambipillai, S. (2018). Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journal of Ambient Intelligence and Humanized Computing*, 9(2):225–251.

[Savran et al., 2008] Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Bosphorus database for 3D face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56.

[Schirrmeister et al., 2017] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 38(11):5391–5420.

[Schmidhuber, 2015] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

[Sela et al., 2017] Sela, M., Richardson, E., and Kimmel, R. (2017). Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1585–1594.

[Serletis et al., 2014] Serletis, D., Bulacio, J., Bingaman, W., Najm, I., and González-Martínez, J. (2014). The stereotactic approach for mapping epileptic networks: a prospective study of 200 patients: Clinical article. *Journal of neurosurgery*, 121(5):1239–1246.

[Sevilla-Lara and Learned-Miller, 2012] Sevilla-Lara, L. and Learned-Miller, E. (2012). Distribution fields for tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1910–1917.

[Shao et al., 2016] Shao, Z., Ding, S., Zhao, Y., Zhang, Q., and Ma, L. (2016). Learning deep representation from coarse to fine for face alignment. *arXiv preprint arXiv:1608.00207*.

[Sharma et al., 2019] Sharma, V., Tapaswi, M., Sarfraz, M. S., and Stiefelhagen, R. (2019). Self-supervised learning of face representations for video face clustering. *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*.

[Sharp et al., 2015] Sharp, T., Keskin, C., Robertson, D., Taylor, J., Shotton, J., Kim, D., Rhemann, C., Leichter, I., Vinnikov, A., Wei, Y., et al. (2015). Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642.

[Shen et al., 2017] Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248.

[Shen et al., 2015] Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 50–58.

[Shi and Tomasi, 1994] Shi, J. and Tomasi, C. (1994). Good features to track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600.

[Shih et al., 2018] Shih, P., Nikpour, A., Bleasel, A., Herkes, G., Mitchell, R., Seah, R., Mumford, V., Braithwaite, J., Vagholkar, S., and Rapport, F. (2018). Leading up to saying" yes": A qualitative study on the experience of patients with refractory epilepsy regarding presurgical investigation for resective surgery. *Epilepsy & Behavior*, 83:36–43.

[Shuo Yang and Tang, 2015] Shuo Yang, Ping Luo, C. C. L. and Tang, X. (2015). From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3676–3684.

[Sigal et al., 2010] Sigal, L., Balan, A. O., and Black, M. J. (2010). Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4.

[Simon et al., 2017] Simon, T., Joo, H., Matthews, I. A., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 2.

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

[Sirven and Glosser, 1998] Sirven, J. I. and Glosser, D. S. (1998). Psychogenic nonepileptic seizures: theoretic and clinical considerations. *Neuropsychiatry, neuropsychology, and behavioral neurology*, 11(4):225–235.

[Sminchisescu, 2008] Sminchisescu, C. (2008). 3D human motion analysis in monocular video: techniques and challenges. In *Human Motion*, pages 185–211.

[So, 2006] So, E. L. (2006). Value and limitations of seizure semiology in localizing seizure onset. *Journal of clinical neurophysiology*, 23(4):353–357.

[Sønderby et al., 2015] Sønderby, S. K., Sønderby, C. K., Nielsen, H., and Winther, O. (2015). Convolutional LSTM networks for subcellular localization of proteins. In *Proceeding of the International Conference on Algorithms for Computational Biology (ALCoB)*, pages 68–80.

[Song et al., 2017] Song, J., Wang, L., Van Gool, L., and Hilliges, O. (2017). Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5563–5572.

[Sors et al., 2018] Sors, A., Bonnet, S., Mirek, S., Vercueil, L., and Payen, J.-F. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomedical Signal Processing and Control*, 42:107–114.

[Souirti et al., 2014] Souirti, Z., Landré, E., Mellerio, C., Devaux, B., and Chassoux, F. (2014). Neural network underlying ictal pouting ("chapeau de gendarme") in frontal lobe epilepsy. *Epilepsy & Behavior*, 37:249–257.

[Spencer and Huh, 2008] Spencer, S. and Huh, L. (2008). Outcomes of epilepsy surgery in adults and children. *The Lancet Neurology*, 7(6):525–537.

[Sridhar et al., 2016] Sridhar, S., Mueller, F., Zollhöfer, M., Casas, D., Oulasvirta, A., and Theobalt, C. (2016). Real-time joint tracking of a hand manipulating an object from RGB-D input. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 294–310.

[Stober et al., 2015] Stober, S., Sternin, A., Owen, A. M., and Grahn, J. A. (2015). Deep feature learning for EEG recordings. *arXiv preprint arXiv:1511.04306*, page abs/1511.04306.

[Stolojescu-CriŞan and Holban, 2013] Stolojescu-CriŞan, C. and Holban, Ş. (2013). A comparison of x-ray image segmentation techniques. *Advances in Electrical and Computer Engineering Engineering*, 13(3).

[Sun et al., 2018] Sun, X., Wu, P., and Hoi, S. C. (2018). Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing*, 299:42–50.

[Sun et al., 2013] Sun, Y., Wang, X., and Tang, X. (2013). Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483.

[Supratak et al., 2016] Supratak, A., Wu, C., Dong, H., Sun, K., and Guo, Y. (2016). Survey on feature extraction and applications of biosignals. *Machine Learning for Health Informatics*, pages 161–182.

[Syed et al., 2011] Syed, T. U., LaFrance, W. C., Kahriman, E. S., Hasan, S. N., Rajasekaran, V., Gulati, D., Borad, S., Shahid, A., Fernandez-Baca, G., Garcia, N., et al. (2011). Can semiology predict psychogenic nonepileptic seizures? a prospective study. *Annals of neurology*, 69(6):997–1004.

[Szaflarski et al., 2018] Szaflarski, J. P., Allendorfer, J. B., Nenert, R., LaFrance, W. C., Barkan, H. I., DeWolfe, J., Pati, S., Thomas, A. E., and Ver Hoef, L. (2018). Facial emotion processing in patients with seizure disorders. *Epilepsy & Behavior*, 79:193–204.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

[Taigman et al., 2014] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708.

[Tang et al., 2014] Tang, D., Jin Chang, H., Tejani, A., and Kim, T.-K. (2014). Latent regression forest: Structured estimation of 3D articulated hand posture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3786–3793.

[Tatum IV, 2012] Tatum IV, W. O. (2012). Mesial temporal lobe epilepsy. *Journal of Clinical Neurophysiology*, 29(5):356–365.

[Tekin et al., 2016] Tekin, B., Rozantsev, A., Lepetit, V., and Fua, P. (2016). Direct prediction of 3D body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 991–1000.

[Tewari et al., 2017] Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., and Theobalt, C. (2017). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 5.

[Thevenot et al., 2017] Thevenot, J., López, M. B., and Hadid, A. (2017). A survey on computer vision for assistive medical diagnosis from faces. *IEEE journal of biomedical and health informatics*.

[Thodoroff et al., 2016] Thodoroff, P., Pineau, J., and Lim, A. (2016). Learning robust features using deep learning for automatic seizure detection. In *Proceedings of the Machine learning for healthcare conference*, pages 178–190.

[Thornton et al., 2010] Thornton, R., Laufs, H., Rodionov, R., Cannadathu, S., Carmichael, D. W., Vulliemoz, S., Salek-Haddadi, A., McEvoy, A. W., Smith, S. M., Lhatoo, S., et al. (2010). EEG correlated functional MRI and postoperative outcome in focal epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 81(8):922–927.

[Tomasi and Kanade, 1991] Tomasi, C. and Kanade, T. (1991). Detection and tracking of point features. *School of Computer Science, Carnegie Mellon Univ. Pittsburgh*.

[Tome et al., 2017] Tome, D., Russell, C., and Agapito, L. (2017). Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Tompson et al., 2015] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656.

[Tompson et al., 2014] Tompson, J., Stein, M., Lecun, Y., and Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169.

[Toshev and Szegedy, 2014] Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660.

[Tran et al., 2017] Tran, A. T., Hassner, T., Masi, I., and Medioni, G. (2017). Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502.

[Tran et al., 2018] Tran, A. T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., and Medioni, G. (2018). Extreme 3D face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3935–3944.

[Tufenkjian and Lüders, 2012] Tufenkjian, K. and Lüders, H. O. (2012). Seizure semiology: its value and limitations in localizing the epileptogenic zone. *Journal of Clinical Neurology*, 8(4):243–250.

[Tzallas et al., 2006] Tzallas, A., Karvelis, P., Katsis, C., Fotiadis, D., Giannopoulos, S., and Konitsiotis, S. (2006). A method for classification of transient events in EEG recordings: application to epilepsy diagnosis. *Methods of Information in Medicine*, 45(6):610–621.

[Tzallas et al., 2009] Tzallas, A. T., Tsipouras, M. G., and Fotiadis, D. I. (2009). Epileptic seizure detection in EEGs using time–frequency analysis. *IEEE transactions on information technology in biomedicine*, 13(5):703–710.

[Tzimiropoulos and Pantic, 2014] Tzimiropoulos, G. and Pantic, M. (2014). Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858.

[Tzionas et al., 2016] Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., and Gall, J. (2016). Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193.

[Ulate-Campos et al., 2016] Ulate-Campos, A., Coughlin, F., Gainza-Lein, M., Fernández, I. S., Pearl, P., and Loddenkemper, T. (2016). Automated seizure detection systems and their effectiveness for each type of seizure. *Seizure*, 40:88–101.

[Ulowetz et al., 2005] Ulowetz, S., Cunha, J., Mauerer, C., Vollmar, C., Feddersen, B., and Noachtar, S. (2005). Quantitative movement analysis of extent of wrist movements identifies hypermotor seizures in a non-selected sample of focal epileptic motor seizures. *Aktuelle Neurologie*, 32(S 4):P529.

[Unzueta et al., 2014] Unzueta, L., Pimenta, W., Goenetxea, J., Santos, L. P., and Dornaika, F. (2014). Efficient generic face model fitting to images and videos. *Image and Vision Computing*, 32(5):321–334.

[Uřičář et al., 2016] Uřičář, M., Franc, V., Thomas, D., Sugimoto, A., and Hlaváč, V. (2016). Multi-view facial landmark detector learned by the structured output svm. *Image and Vision Computing*, 47:45–59.

[Valstar et al., 2017] Valstar, M. F., Sánchez-Lozano, E., Cohn, J. F., Jeni, L. A., Girard, J. M., Zhang, Z., Yin, L., and Pantic, M. (2017). Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 839–847.

[van Andel et al., 2016] van Andel, J., Thijs, R. D., de Weerd, A., Arends, J., and Leijten, F. (2016). Non-EEG based ambulatory seizure detection designed for home use: What is available and how will it influence epilepsy care. *Epilepsy & Behavior*, 57:82–89.

[Van de Vel et al., 2016] Van de Vel, A., Cuppens, K., Bonroy, B., Milosevic, M., Jansen, K., Van Huffel, S., Vanrumste, B., Cras, P., Lagae, L., and Ceulemans, B. (2016). Non-EEG seizure detection systems and potential SUDEP prevention: state of the art. *Seizure*, 41:141–153.

[Vidyaratne et al., 2016] Vidyaratne, L., Glandon, A., Alam, M., and Iftekharuddin, K. M. (2016). Deep recurrent neural network for seizure detection. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1202–1207.

[Vielzeuf et al., 2017] Vielzeuf, V., Pateux, S., and Jurie, F. (2017). Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the International Conference on Multimodal Interaction*, pages 569–576.

[Vijayan et al., 2011] Vijayan, V., Bowyer, K., and Flynn, P. (2011). 3D twins and expression challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2100–2105.

[Viola and Jones, 2001] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Vlachos et al., 2005] Vlachos, M., Yu, P., and Castelli, V. (2005). On periodicity detection and structural periodic similarity. In *Proceedings of the SIAM international conference on data mining*, pages 449–460.

[Wagner et al., 2004] Wagner, P., Cunha, J., Mauerer, C., Vollmar, C., Feddersen, B., and Noachtar, S. (2004). Comparison of quantified ipsilateral and contralateral head movements in patients with frontal and temporal lobe epilepsies. *Klinische Neurophysiologie*, 35(03):308.

[Wang et al., 2017a] Wang, H., Li, Z., Ji, X., and Wang, Y. (2017a). Face r-cnn. *arXiv preprint arXiv:1706.01061*.

[Wang et al., 2018a] Wang, N., Gao, X., Tao, D., Yang, H., and Li, X. (2018a). Facial feature point detection: A comprehensive survey. *Neurocomputing*, 275:50–65.

[Wang et al., 2018b] Wang, P., Li, W., Li, C., and Hou, Y. (2018b). Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158:43–53.

[Wang et al., 2017b] Wang, S., Wang, Y., Tang, J., Shu, K., Ranganath, S., and Liu, H. (2017b). What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *Proceedings of the International World Wide Web Conference*, pages 391–400.

[Wang et al., 2018c] Wang, S.-H., Phillips, P., Sui, Y., Liu, B., Yang, M., and Cheng, H. (2018c). Classification of alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *Journal of medical systems*, 42(5):85.

[Wang et al., 2017c] Wang, Y., Ji, X., Zhou, Z., Wang, H., and Li, Z. (2017c). Detecting faces using region-based fully convolutional networks. *arXiv preprint arXiv:1709.05256*.

[Wei et al., 2016] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732.

[Wendling et al., 2003] Wendling, F., Bartolomei, F., Bellanger, J.-J., Bourien, J., and Chauvel, P. (2003). Epileptic fast intracerebral EEG activity: evidence for spatial decorrelation at seizure onset. *Brain*, 126(6):1449–1459.

[Wendling et al., 2009] Wendling, F., Bartolomei, F., and Senhadji, L. (2009). Spatial analysis of intracerebral EEG in the time and frequency domain: identification of epileptogenic networks in partial epilepsy. *Philos T Roy Soc A*, 367:297–316.

[Wendling et al., 2016] Wendling, F., Benquet, P., Bartolomei, F., and Jirsa, V. (2016). Computational models of epileptiform activity. *Journal of neuroscience methods*, 260:233–251.

[Werbos, 1990] Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

[Wiebe et al., 2001] Wiebe, S., Blume, W. T., Girvin, J. P., and Eliasziw, M. (2001). A randomized, controlled trial of surgery for temporal-lobe epilepsy. *New England Journal of Medicine*, 345(5):311–318.

[Williamson et al., 1985] Williamson, P. D., Spencer, D. D., Spencer, S. S., Novelly, R. A., and Mattson, R. H. (1985). Complex partial seizures of frontal lobe origin. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 18(4):497–504.

[Wilson and Emerson, 2002] Wilson, S. B. and Emerson, R. (2002). Spike detection: a review and comparison of algorithms. *Clinical Neurophysiology*, 113(12):1873–1881.

[Wojke, 2018] Wojke, N. (2018). Simple online realtime tracking with a deep association metric. https://github.com/nwojke/deep_sort.

[Wojke et al., 2017] Wojke, N., Bewley, A., and Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 3645–3649.

[Wu et al., 2017] Wu, Y., Gou, C., and Ji, Q. (2017). Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3471–3480.

[Wu et al., 2018] Wu, Y., Shah, S. K., and Kakadiaris, I. A. (2018). Godp: Globally optimized dual pathway deep network architecture for facial landmark localization in-the-wild. *Image and Vision Computing*, 73:1–16.

[Xiao et al., 2016] Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., and Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 57–72.

[Xie et al., 2017] Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995.

[Xiong and De la Torre, 2013] Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539.

[Xu and Huang, 2012] Xu, G. and Huang, J. Z. (2012). Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *The Annals of Statistics*, 40(6):3003–3030.

[Yamamoto et al., 2018] Yamamoto, A., Nakamoto, H., Bessho, Y., Terada, T., and Ishikawa, A. (2018). Innovative larynx elevation counter during saliva swallowing using a series of flexible stretchable strain sensors. In *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*.

[Yan et al., 2014] Yan, J., Zhang, X., Lei, Z., and Li, S. Z. (2014). Face detection by structural models. *Image and Vision Computing*, 32(10):790–799.

[Yan et al., 2017] Yan, S., Xia, Y., Smith, J. S., Lu, W., and Zhang, B. (2017). Multiscale convolutional neural networks for hand detection. *Applied Computational Intelligence and Soft Computing*, 2017.

[Yang et al., 2014] Yang, B., Yan, J., Lei, Z., and Li, S. Z. (2014). Aggregate channel features for multi-view face detection. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8.

[Yang et al., 2015] Yang, H., Mou, W., Zhang, Y., Patras, I., Gunes, H., and Robinson, P. (2015). Face alignment assisted by head pose estimation. *arXiv preprint arXiv:1507.03148*.

[Yang et al., 2016] Yang, S., Luo, P., Loy, C.-C., and Tang, X. (2016). Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533.

[Yang et al., 2018a] Yang, S., Luo, P., Loy, C. C., and Tang, X. (2018a). Faceness-net: Face detection through deep facial part responses. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1845–1859.

[Yang et al., 2018b] Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., and Wang, X. (2018b). 3D human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1.

[Yasin et al., 2016] Yasin, H., Iqbal, U., Kruger, B., Weber, A., and Gall, J. (2016). A dual-source approach for 3D pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4948–4956.

[Yin et al., 2008] Yin, L., Chen, X., Sun, Y., Worm, T., and Reale, M. (2008). A high-resolution 3D dynamic facial expression database. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 1–6.

[Yin et al., 2006] Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006). A 3D facial expression database for facial behavior research. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 211–216.

[Yu et al., 2017] Yu, R., Saito, S., Li, H., Ceylan, D., and Li, H. (2017). Learning dense facial correspondences in unconstrained images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4723–4732.

[Zadeh et al., 2017] Zadeh, A., Lim, Y. C., Baltrusaitis, T., and Morency, L.-P. (2017). Convolutional experts constrained local model for 3D facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2519–2528.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833.

[Zhang et al., 2018] Zhang, D., Guo, G., Huang, D., and Han, J. (2018). Poseflow: A deep motion representation for understanding human behaviors in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6762–6770.

[Zhang et al., 2016a] Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., and Yang, Q. (2016a). 3D hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*.

[Zhang et al., 2014a] Zhang, J., Shan, S., Kan, M., and Chen, X. (2014a). Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–16.

[Zhang et al., 2017a] Zhang, K., Huang, Y., Du, Y., and Wang, L. (2017a). Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203.

[Zhang et al., 2016b] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016b). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

[Zhang et al., 2017b] Zhang, L., Zhu, G., Shen, P., Song, J., Shah, S. A., and Bennamoun, M. (2017b). Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3120–3128.

[Zhang et al., 2017c] Zhang, W., Liu, X., Zuo, L., Guo, Q., Wang, Y., et al. (2017c). Ipsiversive ictal eye deviation in inferioposterior temporal lobe epilepsy—two SEEG cases report. *BMC neurology*, 17(1):38.

[Zhang et al., 2013] Zhang, W., Zhu, M., and Derpanis, K. G. (2013). From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2248–2255.

[Zhang et al., 2019] Zhang, X., Li, Q., Zhang, W., and Zheng, W. (2019). End-to-end hand mesh recovery from a monocular RGB image. *arXiv preprint arXiv:1902.09305*.

[Zhang et al., 2014b] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., and Girard, J. M. (2014b). BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706.

[Zhang et al., 2014c] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014c). Facial landmark detection by deep multi-task learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 94–108.

[Zhang et al., 2016c] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2016c). Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930.

[Zhou et al., 2016a] Zhou, J., Hong, X., Su, F., and Zhao, G. (2016a). Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 84–92.

[Zhou et al., 2016b] Zhou, T., Pillai, P. J., and Yalla, V. G. (2016b). Hierarchical context-aware hand detection algorithm for naturalistic driving. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 1291–1297.

[Zhou et al., 2017] Zhou, X., Huang, Q., Sun, X., Xue, X., and Wei, Y. (2017). Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[Zhou et al., 2016c] Zhou, X., Wan, Q., Zhang, W., Xue, X., and Wei, Y. (2016c). Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*.

[Zhou et al., 2016d] Zhou, X., Zhu, M., Leonardos, S., Derpanis, K. G., and Daniilidis, K. (2016d). Sparseness meets deepness: 3D human pose estimation from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4966–4975.

[Zhu et al., 2015] Zhu, S., Li, C., Change Loy, C., and Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006.

[Zhu et al., 2016] Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2016). Face alignment across large poses: A 3D solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155.

[Zhu and Ramanan, 2012] Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. IEEE.

[Zimmermann and Brox, 2017] Zimmermann, C. and Brox, T. (2017). Learning to estimate 3D hand pose from single RGB images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 3.

Appendix A

Participant Information Sheet

A.1 Consent Form Study and Publication

Participant Information Sheet/Consent Form Publication

Title	Multi-modal analysis of Video-SEEG Monitoring for the automatic evaluation of epilepsy.
Short Title	Automatic contribution of Seizure semiology and Electrical activity.
Protocol Number	Version 2
Investigator(s)	Dr. Clinton Fookes[1], Dr. MD. Sasha Dionisio[2], David Ahmedt[1], Dr. Sridha Sridharan[1], Dr. Kien Nguyen[1] [1]Queensland University of Technology; [2]The Mater Hospital
Location	Mater Advanced Epilepsy Unit, Brisbane

Part 1 What does my participation involve?

1. Introduction

You were invited to take part in the research project called *Multimodal analysis of Video-SEEG monitoring for the automatic evaluation of epilepsy*, and you agreed to the researchers using videos and personal information of you as explained in the previous Participants Information Sheet and Consent Form.

This Participant Information Sheet for publication purpose explains what this stage involved helping you decide if you want to take part in this important research phase.

Please read this information carefully. Ask questions if you don't understand or want to know more about. Before deciding whether or not to take part, you might want to discuss this study with a relative, friend or the researchers. Please ensure that you have discussed any questions and are comfortable with the response before providing consent.

If you choose to participate in this research project phase, you will be asked to sign the consent section. By signing it you are telling us that you:

• Understand what you have read

• Consent to use information and pictures in high-impact internationally peer-reviewed publications.

You will be given a copy of this Participant Information and Consent Form for reference.

2. What is the purpose of this research phase?

The researchers would like to illustrate to the scientific community some of the benefits, impacts and challenges of performing video analytics of face expression and body movements for the epilepsy evaluation within peer reviewed conference/journal papers as well as presentations at conferences by using small portions of video or images recorded during observations. These

images will be used to highlight the findings and outcomes of the research and will contribute to improving the understanding of epilepsy. The captured photos will be used to illustrate how the automatic application works in real-world scenarios of epileptic patients.

3. What are the possible benefits of taking part?

You will not directly benefit from the publication; however, it will provide a better understanding of the benefits/challenges of the applications of video analytics for solving real-world scenarios.

4. What are the possible risks of taking part?

Your images will be included in peer-reviewed publications and scientific presentations. The researchers guarantee that data will be treated with respect and the publications will only be in high-quality journals, which provide respect and dignity to the patients that participated in the research. However, the researchers understand that once the information is published, it is not possible to control it because it is a public domain, and we are aware that the research community or others could access and use the images published for different purposes.

5. What if I withdraw from this research project phase?

You will have the opportunity to review the images/footage before our usage and can decide to withdraw at that stage. We will only use your images with your express permission. However, once the video or a publication is produced it will not be possible to withdraw.

6. What will happen to information about me?

We are requesting your consent to publish images captured of your facial expression and body posture in peer-reviewed academic papers, internationally respected scientific journals and slides used for presentations at conferences. This image will allow identification of your identity.

We would like to illustrate some of the benefits and impacts of performing video analytics in specific behaviours for the epilepsy evaluation within peer-reviewed papers and journals.

Researchers understand that video participants may not wish to be named in this video. As a result, the names of all video participants will be excluded from the corresponding videos. Investigators will only identify you in the video on the basis of your association with the researchers, i.e. participation in epilepsy monitoring. Each publication will use a small number of images to describe the results.

7. Further information and who to contact

If you want any further information concerning this project phase, you can contact the following person:

Name	Dr. Clinton Fookes
Position	Research Officer – Contact Person
Telephone	+617 3138 2458
Email	c.fookes@qut.edu.au

Consent Form – Publication

Title	Multi-modal analysis of Video-SEEG Monitoring for the automatic evaluation of epilepsy.
Short Title	Automatic contribution of Seizure semiology and Electrical activity.
Protocol Number	Version 2
Principal/ Associate Investigator(s)	Dr. Clinton Fookes[1], Dr. MD. Sasha Dionisio[2], David Ahmedt[1], Dr. Sridha Sridharan[1], Dr. Kien Nguyen[1] [1]Queensland University of Technology; [2]The Mater Hospital
Location	Mater Advanced Epilepsy Unit, Brisbane

Declaration by Participant

I have read the attached Participant Information Sheet for publication outlining the nature and purpose of the phase and the extent of my involvement, and have had these details explained to me. I have had the opportunity to ask further questions and am satisfied that I understand.

I agree with the research using, reproducing and disclosing photographic or video images of me as explained in this Participant Information Sheet and Consent Form for publication purposes.

I have been informed that no information regarding my medical history and personal information such as name and family will be divulged.

I agree that I will make no claim against QUT or Mater for any payment or fee for appearing in promotional material or advertisements and release researchers from any other claims arising out of the use of the images of me.

I understand that once a publication is produced it will not be possible to withdraw.

I understand that I will be given a signed copy of this document to keep.

Name of Participant (please print) _____

Signature _____ Date _____

Name of Witness* to Participant's Signature (please print) _____

Signature _____ Date _____

Declaration by Researcher†

I have given a verbal explanation of the research project, its procedures and risks and I believe that the participant has understood that explanation.

Name of Researcher (please print) _____

Signature _____ Date _____

* Witness is not to be the investigator, a member of the study team or their delegate. In the event that an interpreter is used, the interpreter may not act as a witness to the consent process. Witness must be 18 years or older.

Form for Withdrawal of Participation

Title	Multi-modal analysis of Video-SEEG Monitoring for the automatic evaluation of epilepsy.
Protocol Number	Version 2
Investigator(s)	Dr. Clinton Fookes[1], Dr. MD. Sasha Dionisio[2], David Ahmedt[1], Dr. Sridha Sridharan[1], Dr. Kien Nguyen[1] [1]Queensland University of Technology; [2]The Mater Hospital
Location	Mater Advanced Epilepsy Unit, Brisbane

Declaration by Participant

I wish to withdraw from participation in the above research project phase and understand that such withdrawal will not affect my routine care, or my relationships with the researchers, The Queensland University of Technology and the Mater Hospital.

Name of Participant (please print) _____

Signature _____ Date _____

In the event that the participant's decision to withdraw is communicated verbally, the Senior Researcher must provide a description of the circumstances below.

Declaration by Researcher†

I have given a verbal explanation of the implications of withdrawal from the research project and I believe that the participant has understood that explanation.

Name of Researcher (please print) _____

Signature _____ Date _____

† An appropriately qualified member of the research team must provide information concerning withdrawal from the research project.

Appendix B

Ethical Clearance Application

B.1 Mater Health Services Research Ethics Approval

Mater
Exceptional People. Exceptional Care.

4 January 2017

Professor Clinton Fookes
Queensland University of Technology
2 George St
School of Science and Engineering S-Block Gardens Point
Brisbane QLD 4000

Dear Professor Fookes

Re: HREC Reference number: HREC/16/MHS/95
Project title: Multimodal analysis of Video-SEEG monitoring for the automatic evaluation of epilepsy

Thank you for submitting the above research project for single ethical review. This project was considered by the Mater Health Services Human Research Ethics Committee (MHS HREC) **(EC00332)** at its meeting held on 22.11.16 and I further reviewed on 08.12.16, 20.12.16 and 03.01.17.

I am pleased to advise that the MHS Human Research Ethics Committee has granted ethical approval of this research project.

Condition of approval: *Please confirm for the HREC that Dr Dionisio will nominate an alternate neurologist to provide advice around recruitment if the participant should wish to seek independent expert input to their decision to participate or not.*

The nominated participating sites for this project are:

- Mater Misericordiae Ltd

This letter constitutes ethical approval only. This project cannot proceed at any site until separate research governance authorisation has been obtained from the CEO or Delegate of the institution under whose auspices the research will be conducted at that site. **Please liaise with your Research Governance office in regard to any additional requirements. At Mater Health Services please contact the Research Governance Office on 07 3163 3769.**

This HREC is constituted and operates in accordance with the National Health and Medical Research Council's (NHMRC) National Statement on Ethical Conduct in Human Research (2007), updated in 2015. The processes used by this HREC to review multi-centre research proposals have been certified by the National Health and Medical Research Council.

Mater Research HREC Office
Room 294 Level 2 Aubigny Place *Ph: 07 3163 1585 Fax: 07 3163 1588* *Email: research.ethics@mmri.mater.org.au*

Mater Misericordiae Health Services Brisbane Limited
ACN 096 708 922
Raymond Terrace,
South Brisbane,
Queensland 4101 Australia
Phone + 61 7 3163 8111
www.mater.org.au

F0621
11/07

The approved documents include:

Document	Version	Date
Cover Letter	1	02 November 2016
Application: Online Forms NEAF Submission Code AU/1/B9E9220	2.2 (2014)	07 November 2016
Protocol	2	05 December 2016
Participant Information Sheet/Consent Form Main Study	3	21 December 2016
Participant Information Sheet/Consent Form Publication	3	21 December 2016
Investigator CV: Dr Kien Nguyen Thanh		18 October 2016
Investigator CV: Emeritus Professor Sridha Sridharan	submitted 07.1.16, valid to 07.11.18	
Investigator CV: Dr Sasha Dionisio	submitted 07.1.16, valid to 07.11.18	
Investigator CV: Professor Clinton Fookes	submitted 07.11.16, valid to 07.11.18	
Investigator CV: David Esteban AhmedtAristizabal	submitted 07.11.16, valid to 07.11.18	
3D model fitting for facial expression analysis under uncontrolled imaging conditions	*(publication) 978-1-4244-2175-6/08/$25.00 ©2008 IEEE*	
Neural network underlying ictal pouting ("chapeau de gendarme") in frontal lobe epilepsy	*(publication) Epilepsy & Behavior 37 (2014) 249–257*	
Frontal lobe epilepsy	*(publication) Practical Neurology, 2004, 4, 260–273*	
Frontal lobe seizures: From clinical semiology to localization	*(publication) Epilepsia, 55(2):264–277, 2014*	
Response to Request for Further Information		7 December 2016

Approval of this project by the MHS HREC is valid from **03.01.17** to **03.01.20**, subject to the following conditions being met:

- The Principal Investigator will immediately report anything that might warrant review of ethical approval of the project.
- The Principal Investigator will notify the MHS HREC of any event that requires a modification to the protocol or other project documents and submit any required amendments.
- The Principal Investigator will submit any necessary reports related to the safety of research participants.
- In accordance with *Section 3.3.22(b)* of the National Statement the Principal Investigator will report to the MHS HREC annually, the first report is to be submitted by **03.01.18**. Template may be downloaded at: http://www.mater.org.au/Home/Research/Human-Research-Ethics-Committee/HREC-and-RGO-Resources
- The Principal Investigator will notify the MHS HREC if the project is discontinued before the expected completion date, with reasons provided.

- The Principal Investigator will notify the MHS HREC of any plan to extend the duration of the project past the approval period listed above and will submit any associated required documentation.
- A copy of this ethical approval letter together with completed Site Specific Assessment (SSA) and any other requirements must be submitted by all site Principal Investigators to the Research Governance Office at each participating institution in a timely manner to enable the institution to authorise the commencement of the project at its site/s.

Please confirm the commencement date with the Research Ethics Office.

Should you have any queries about the MHS HREC's consideration of your project, please contact the HREC Coordinator on (07) 3163 1585. The MHS HREC Terms of Reference, membership and standard forms are available at *http://www.mater.org.au/Home/Research/Human-Research-Ethics-Committee/Human-Research-Ethics/HREC-Resources)*

The MHS HREC wishes you every success in your research.

Yours sincerely

Dr Conor Brophy MBBS; MD; MBioethics; FRCP; AFRACMA
Chairperson
Mater Health Services Human Research Ethics Committee

B.2 Site Specific Assessment Approval

Mater
Exceptional People. Exceptional Care.

Mater Research

Mater Human Research Governance – Site Specific Assessment Authorisation

3 April 2017

Dr Sasha Dionisio
Neurologist
Mater Misericordiae Limited

Dear Dr Dionisio

Re: Project Title: Multi-Modal analysis of Video SEEG monitoring for the automatic evaluation of epilepsy

Mater Research Governance Reference Number: RG-17-008
Human Research Ethics Committee (HREC) Reference Number: HREC/16/MHS/95
Mater Research Hub Reference Number: MR-2017-4

Thank you for submitting an application for authorisation of this project. I am pleased to inform you that authorisation has been granted for this study to take place at the following site(s):

Mater Hospital Brisbane
Mater Centre for Neuroscience

Documents reviewed and authorised by Mater Research Governance are as per those listed on HREC approval letter dated 4 January 2017.

The following conditions apply to this research project. These are additional to those conditions imposed by the HREC that granted ethical approval.

1. The Mater Research Governance Office must be informed of any problems that arise during the course of the study which may affect conduct of the study at Mater, including serious or unexpected adverse events occurring at the site.
2. Proposed amendments to the research protocol, study documentation, or conduct of the research must be submitted to both the reviewing HREC and the Mater Research Governance Office. Written HREC approval and Mater Research Governance authorisation are both required before an amendment may be implemented at Mater.
3. Proposed amendments to the conduct of the research which affect the ongoing acceptability of the project at Mater (including changes to any of the following: funding and budget, existing research agreements, site investigator team or the Mater contact person) are to be submitted to the Mater Research Governance Office.
4. The Mater Research Governance Office must be notified of any students involved with the study, regardless of their role in the study or the stage of the study at which they are added to the investigator team.
5. The Mater Research Governance Office is responsible for monitoring research authorised for conduct at Mater, and for determining the method of monitoring which is appropriate to each project:

(a) For all projects: Annual Progress Reports must be provided to the Mater Research Governance Office on the anniversary of the original HREC Approval. Mater Research Governance accepts progress reports on whichever template is required by the reviewing HREC.
(b) A proportion of all human research projects authorised for conduct at Mater will be eligible for site monitoring each year and the research team will be informed prior to the monitoring visit if this project has been selected.

We wish you every success in undertaking this research.

Yours sincerely

Dominique Williams
Research Governance Officer
Room 270, Lvl 2, Aubigny Place
Raymond Terrace
South Brisbane Qld 4101

B.3 National Ethics Application Form

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

Online Forms
National Ethics Application Form

Within which Jurisdictions will your research application be submitted to: *(tick all that apply)*

- [ ] New South Wales
- [x] Queensland
- [ ] South Australia
- [ ] Victoria

HREC Application Reference Number:

1. TITLE AND SUMMARY OF PROJECT

1. Title

What is the formal title of this research proposal?
Multimodal analysis of Video-SEEG monitoring for the automatic evaluation of epilepsy.

What is the short title / acronym of this research proposal (if applicable)?
Automatic contribution of Seizure semiology and Electrical activity.

2. Description of the project in plain language

Give a concise and simple description (not more than 400 words), in plain language, of the aims of this project, the proposal research design and the methods to be used to achieve those aims.

About 30% of the patients with partial epilepsy are resistant to medical therapy (WHO, 2016), where epilepsy surgery is widely accepted as an effective therapeutic option (Wiebe et al., 2001). Seizure freedom and improvement of seizure control are the desired and most commonly reported outcomes, where the final goal of the pre-surgical evaluation is to delineate the epileptogenic network (Rosenow 2001). There are significant contributions of pre-surgery assessment from non-invasive techniques based on neuroimaging, electrophysiology, and neuropsychological testing; however, certain clinical settings call for intracranial recordings such as stereo-encephalography (sEEG), which accurately maps the eloquent networks involve during a seizure event. It is widely accepted that semiologic and electrical patterns are difficult to characterise and liable to be misleading in predicting the localisation of the epileptogenic network (Chauvel & McGonigal, 2014). Most of the contribution are focused on semi-automatic techniques, where surgery diagnosis still relies on the experts' experience and time-consuming subjective interpretation. Since misdiagnosis currently reaches a rate of 30%, there is an evident keen interest in improving the diagnostic precision using computer-based methodologies that in the last years have shown near-human performance.

The overall aim is to develop a methodology based on synchronous multi-modal analysis to perform a novel epilepsy assessment procedure that can jointly learn across visually observed semiology patterns of behaviour and brain electrical activity recorded in the sEEG signals. The main expected outcome is a learning system that predicts the linked anatomical structures that constitute the epileptogenic network from the electro-clinical correlation. This research is capable of attributing and localising subsets of brain networks related to the semiology production, leading to a better understanding of the distinctive types of motor manifestation and increasing the diagnostic precision in epilepsy surgery for the achievement of seizure freedom.

The research addresses the challenge of modelling and extracting features of the human motion analysis (through videos captured by a fixed camera in the hospital room) during the semiologic production of upper limbs movements and facial expression, and their correlation with the brain electrical activity. Different computer methodologies based on deep learning methods such as convolutional neural networks and unsupervised learning techniques will be investigated.

This research will take place at the Mater Advanced Epilepsy Unit in Brisbane. The methodology includes the processing and analysis of both Videos and sEEG signals, where this can be accomplished at no risk to the

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

patients because the automatic application will not be in direct contact with the patients or the Epilepsy Monitoring Units of Mater.

2. RESEARCHERS / INVESTIGATORS

2. Principal researcher(s) / investigator(s)

Principal researcher / investigator 1

Title:	Forename/Initials:	Surname:
Professor	Clinton	Fookes

Mailing Address:	Queensland University of Technology
	2 George St
	School of Science and Engineering, S-Block Gardens Point
Suburb/Town:	Brisbane
State:	QLD
Postcode:	4000
Country:	Australia
Organisation:	Queensland University of Technology
Department*:	Science and Engineering Faculty
Position:	Discipline Leader – Vision & Signal Processing QUT
E-mail:	c.fookes@qut.edu.au
Phone (BH):	
Phone (AH)*:	
Mobile*:	+617 3138 2458
Pager*:	
Fax:	

Is this person the contact person for this application?
○ Yes ◉ No

Summary of qualifications and relevant expertise
Dr. Clinton Fookes is a Professor in Vision & Signal Processing and the Speech, Audio, Image and Video Technologies group within the Science and Engineering Faculty at QUT. He holds a BEng (Aerospace/Avionics), a MBA with a focus on technology innovation/management, and a PhD in the field of computer vision. He actively researches in the fields of computer vision and pattern recognition and he has published over 140 internationally peer-reviewed articles.

Please declare any general competing interests
Nil

Name the site(s) for which this principal researcher / investigator is responsible.
Mater Hospital
Queensland University of Technology

Describe the role of the principal researcher / investigator in this project.
To oversee the conduct of the research, data analysis and interpretation. Methodological advisor. Dissemination of results.

Is the principal researcher a student? ○ Yes ◉ No

Principal researcher / investigator 2

Title:	Forename/Initials:	Surname:
Dr	Sasha	Dionisio

Mailing Address:	Mater Advanced Epilepsy Unit
	Salmon Building, 551 Stanley Street
Suburb/Town:	South Brisbane
State:	QLD
Postcode:	4101
Country:	Australia

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Organisation: Mater Hospital
Department*: Mater Advanced Epilepsy Unit
Position: Neurologist, Consultant Epileptologist and Head of Mater Advanced Epilepsy Unit
E-mail: Sasha.Dionisio@mater.org.au
Phone (BH):
Phone (AH)*:
Mobile*:
Pager*:
Fax:

Is this person the contact person for this application?
○ Yes ◉ No

Summary of qualifications and relevant expertise
Dr Sasha Dionisio is head of the advanced epilepsy unit at Mater Centre for Neurosciences. Dr Dionisio attended medical school in Ireland, prior to moving to Australia to continue his training in Neurology. Dr Dionisio was a clinical fellow at the prestigious Cleveland Clinic Epilepsy Centre (Ohio, USA) where he was trained in all aspects of advanced epilepsy care, including surgical epilepsy assessment and understanding of the Stereo-EEG methodology and interpretation, and time frequency analysis of cerebral signals.

Please declare any general competing interests
Nil

Name the site(s) for which this principal researcher / investigator is responsible.
Mater Hospital

Describe the role of the principal researcher / investigator in this project.
Oversee the neurological care of patient. Aid recruitment and data collection. Assessment and dissemination of results.

Is the principal researcher a student? ○ Yes ◉ No

Principal researcher / investigator 3

Title:	Forename/Initials:	Surname:
Mr	David	Ahmedt

Mailing Address: Queensland University of Technology
2 George St
School of Science and Engineering, S-Block Gardens Point
Suburb/Town: Brisbane
State: QLD
Postcode: 4059
Country: Australia
Organisation: Queensland University of Technology
Department*: Science and Engineering Faculty
Position: PhD Student QUT
E-mail: david.aristizabal@hdr.qut.edu.au
Phone (BH):
Phone (AH)*:
Mobile*: 0478779075
Pager*:
Fax:

Is this person the contact person for this application?
○ Yes ◉ No

Summary of qualifications and relevant expertise
Bachelor of Science in Engineering (Mechatronics).
Master of Science in Automation and Industrial Control
Research interest include pattern recognition, biometrics and medical applications.

Please declare any general competing interests
Nil

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

Name the site(s) for which this principal researcher / investigator is responsible.
Mater Hospital
Queensland University of Technology

Describe the role of the principal researcher / investigator in this project.
The proposed research will be performed entirely by the PhD student under the supervision of his supervisory team.

Is the principal researcher a student? (•) Yes ( ) No

What is the educational organisation, faculty and degree course of the student?

Organisation:	Queensland University of Technology
Faculty:	Science and Engineering Faculty
Degree course:	IF49 Doctor of Philosophy

Is this research project part of the assessment of the student? (•) Yes ( ) No

Is the student's involvement in this project elective or compulsory? (•) Elective ( ) Compulsory

What training or experience does the student have in the relevant research methodology?	The Phd candidate has experience in signal analysis and is taking courses of computer vision and deep learning to apply the methodology in the data.
What training has the student received in the ethics of research?	The PhD candidate has attended several seminars from QUT including Human Research Ethics Application process and Risk Management Training sessions. In addition, the Phd candidate has a certificate of completion of Research Ethics, Integrity and Safety course from QUT.
Describe the supervision to be provided to the student.	The PhD candidate has a supervisory team of 4 members (3 from QUT and 1 from Mater). The normal expectation for supervision is weekly meetings of a normal duration of 1 hour. Meeting will be held with at least two members of the supervisory team based on the particular needs of the project. The supervision is focused on methodology and technical issues.

How many supervisors does the student have?

Supervisor 1
Provide the name, qualifications, and expertise, relevant to this research, of the students' supervisor.

Title:	Professor
First Name:	Clinton
Surname:	Fookes

Summary of qualifications and relevant expertise
Principal Supervisor is the Principal researcher 1.
Associate Supervisor are Principal researcher 2, 4 and 5.

Dr. Clinton Fookes is a Professor in Vision & Signal Processing and the Speech, Audio, Image and Video Technologies group within the Science and Engineering Faculty at QUT. He holds a BEng (Aerospace/Avionics), a MBA with a focus on technology innovation/management, and a PhD in the field of computer vision. He actively researches in the fields of computer vision and pattern recognition and he has published over 140 internationally peer-reviewed articles.

Principal researcher / investigator 4

Title:	Forename/Initials:	Surname:
Professor	Sridha	Sridharan

Mailing Address:	Queensland University of Technology
	2 George St
	School of Science and Engineering, S-Block Gardens Point
Suburb/Town:	Brisbane
State:	QLD
Postcode:	4059
Country:	Australia
Organisation:	Queensland University of Technology

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Department*:	Science and Engineering Faculty
Position:	Emeritus Professor/Adjunct Professor QUT
E-mail:	s.sridharan@qut.edu.au
Phone (BH):	
Phone (AH)*:	
Mobile*:	0415 164 698
Pager*:	
Fax:	

Is this person the contact person for this application?
(•) Yes ( ) No

Summary of qualifications and relevant expertise
Dr. Sridha Sridharan is the leader of the SAIVT group and holds a BSc (Electrical Engineering), an MSc (Communication Engineering) and a PhD in the field of Signal Processing. He has supervised 72 PhD students at QUT (as a Principal or active Associate Supervisor) in the area of Image and Speech Technologies and he has published over 400 papers in the areas of Speech and Image technologies.

Please declare any general competing interests
Nil

Name the site(s) for which this principal researcher / investigator is responsible.
Mater Hospital
Queensland University of Technology

Describe the role of the principal researcher / investigator in this project.
Analysis and dissemination of results.
Methodological advisor.

Is the principal researcher a student? ( ) Yes (•) No

Principal researcher / investigator 5

Title:	Forename/Initials:	Surname:
Mr	Kien	Nguyen

Mailing Address:	Queensland University of Technology
	2 George St
	School of Science and Engineering, S-Block Gardens Point
Suburb/Town:	Brisbane
State:	QLD
Postcode:	4059
Country:	Australia
Organisation:	Queensland University of Technology
Department*:	Science and Engineering Faculty
Position:	Research Fellow
E-mail:	k.nguyenthanh@qut.edu.au
Phone (BH):	
Phone (AH)*:	
Mobile*:	0402580882
Pager*:	
Fax:	

Is this person the contact person for this application?
( ) Yes (•) No

Summary of qualifications and relevant expertise
Dr. Kien Nguyen Thanh is a research Fellow at Queensland University of Technology and has spent last 6 years conducting research on computer vision techniques, including object detection, segmentation, registration, object recognition, optical flow and super-resolution for biometrics. Recently, is investigating Graphical Models and Deep Learning with application in satellite imagery and scene understanding.

Please declare any general competing interests
Nil

Name the site(s) for which this principal researcher / investigator is responsible.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

Mater Hospital
Queensland University of Technology

Describe the role of the principal researcher / investigator in this project.
Analysis and dissemination of results.
Technical advisor

Is the principal researcher a student? ○ Yes ◉ No

3. Associate Researcher(s) / investigator(s)

How many known associate researchers are there? (You will be asked to give contact details for these associate researchers / investigators) 0

Do you intend to employ other associate researchers / investigators? ○ Yes ◉ No

5. Other personnel relevant to the research project

5a. How many known other people will play a specified role in the conduct of this research project?

1

5b. Describe the role, and expertise where relevant (e.g. counsellor), of these other personnel.

Technicians to collect Video and EEG/sEEG monitoring

5c. Is it intended that other people, not yet known, will play a specified role in the conduct of this research project?

○ Yes ◉ No

6. Certification of researchers / investigators

6a. Are there any relevant certification, accreditation or credentialing requirements relevant to the conduct of this research?

○ Yes ◉ No

7. Training of researchers

7a. Do the researchers / investigators or others involved in any aspect of this research project require any additional training in order to undertake this research?

○ Yes ◉ No

3. RESOURCES

Project Funding / Support

1. Indicate how the project will be funded?

Type of funding.

[Please note that all fields in any selected funding detail column (with the exception of the code) will need to be completed.]

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

Funding	Confirmed or Sought?
External Competitive Grant	○ Confirmed ○ Sought ◉ Not Sought
Internal Competitive Grant	○ Confirmed ○ Sought ◉ Not Sought
Sponsor	○ Confirmed ○ Sought ◉ Not Sought
By Researchers Department or Organisation	○ Confirmed ○ Sought ◉ Not Sought

2. How will you manage a funding shortfall (if any)?

Funding will be sought from External and Internal Competitive Grants.

3. Will the project be supported in other ways eg. in-kind support/equipment by an external party eg. sponsor?

○ Yes ◉ No

4. Is this a study where capitation payments are to be made, and will participants be made aware of these payments to clinicians or researchers / investigators?

No caption payments will be made

Duality of Interest

5. Describe any commercialisation or intellectual property implications of the funding/support arrangement.

Nil IP implications

6. Does the funding/support provider(s) have a financial interest in the outcome of the research?

○ Yes ◉ No

7. Does any member of the research team have any affiliation with the provider(s) of funding/support, or a financial interest in the outcome of the research?

○ Yes ◉ No

8. Does any other individual or organisation have an interest in the outcome of this research?

○ Yes ◉ No

9. Are there any restrictions on the publication of results from this research?

○ Yes ◉ No

4. PRIOR REVIEWS

Ethical Review

Some HRECs may require researchers to provide information additional to that contained in a NEAF proposal. For this reason, it is prudent to check whether the HRECs to whom you propose to submit this proposal require additional information.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

Duration and location

1. In how many Australian sites, or site types, will the research be conducted?

2

2. In how many overseas sites, or site types, will the research be conducted?

0

3. Provide the following information for each site or site type (Australian and overseas, if applicable) at which the research is to be conducted

1

Site / Site Type Name: Queensland University of Technology

Site / Site Type Location: 2 George St
School of Science and Engineering,
S-Block Gardens Point, Brisbane, QLD 4000

2

Site / Site Type Name: Mater Health Service

Site / Site Type Location: Mater Advanced Epilepsy Unit, Mater Centre for Neuroscience
Salmon Building, 551 Stanley Street, South Brisbane, QLD 4101

4. Provide the start and finish dates for the whole of the study including data analysis

Anticipated start date: 01/12/2016 *(dd/mm/yyyy)*

Anticipated finish date: 01/12/2019 *(dd/mm/yyyy)*

5. Are there any time-critical aspects of the research project of which an HREC should be aware?

○ Yes ◉ No

6. To how many Australian HRECs (representing site organisations or the researcher's / investigator's organisation) is it intended that this research proposal be submitted?

2

A list of NHMRC registered Human Research Ethics Committees (HRECs), along with their institutional affiliations and contact details is available on the NHMRC website at the following web address: http://www.nhmrc.gov.au/health_ethics/hrecs/overview.htm#d.

7. HRECs

HREC 1

Name of HREC:

Queensland University of Technology University Human Research Ethics Committee (EC00171)

Provide the start and finish dates for the research for which this HREC is providing ethical review:

Anticipated start date or date range: 01/12/2016 *(dd/mm/yyyy)*

Anticipated finish date or date range: 01/12/2019 *(dd/mm/yyyy)*

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

For how many sites at which the research is to be conducted will this HREC provide ethical review?
1

Site 1

Name of Site: Queensland University of Technology

Principal Researcher 1

Principal Researcher Name:
Professor Clinton Fookes

Principal Researcher 2

Principal Researcher Name:
Mr David Ahmedt

Principal Researcher 3

Principal Researcher Name:
Professor Sridha Sridharan

Principal Researcher 4

Principal Researcher Name:
Mr Kien Nguyen

HREC 2

Name of HREC:
Mater Health Services HREC (EC00332)

Provide the start and finish dates for the research for which this HREC is providing ethical review:
Anticipated start date or date range: 01/12/2016 *(dd/mm/yyyy)*
Anticipated finish date or date range: 01/12/2019 *(dd/mm/yyyy)*

For how many sites at which the research is to be conducted will this HREC provide ethical review?
1

Site 1

Name of Site: Mater Health Service

Principal Researcher 1

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Principal Researcher Name:
Dr Sasha Dionisio

Principal Researcher 2

Principal Researcher Name:
Mr David Ahmedt

Principal Researcher 3

Principal Researcher Name:
Mr Kien Nguyen

Principal Researcher 4

Principal Researcher Name:
Professor Clinton Fookes

Associate Researcher 1

Associate Researcher Name:

8. Have you previously submitted an application, whether in NEAF of otherwise, for ethical review of this research project to any other HRECs?

○ Yes ◉ No

9. HRECs

Research conducted overseas

Peer review

11. Has the research proposal, including design, methodology and evaluation undergone, or will it undergo, a peer review process?

○ Yes ◉ No

Explain why the research proposal will not undergo a peer review process.
The methodology of anatomo-electro-clinical analysis has been clinical certified (Bonini et. al, 2014; Chauvel & McGonigal, 2014). However, quantitative methods that describe motion patterns in epilepsies combined with progress in the electrophysiological analysis is still under development.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

This project has been validated by peer review in the research community (Vilas-Boas & Cunha, 2016). The SAIVT research group from QUT, is an expert in computer vision and signals analysis, specifically in deep learning techniques proposed in the methodology. The methodology is a novel epilepsy assessment procedure that can jointly learn across visually observed semiology patterns of behaviour and brain electrical activity recorded in the sEEG signals.

5. PROJECT

1. Type of Research

Tick as many of the following 'types of research' as apply to this project. Your answers will assist HRECs in considering your proposal. A tick in some of these boxes will generate additional questions relevant to your proposal (mainly because the National Statement requires additional ethical matters to be considered), which will appear in Section 9 of NEAF.

The project involves:

- ☐ Research using qualitative methods
- ☐ Research using quantitative methods, population level data or databanks, e.g survey research, epidemiological research
- ☐ Clinical research
- ☐ Research involving the collection and / or use of human biospecimens
- ☐ Genetic testing/research
- ☐ A cellular therapy
- ☐ Research on workplace practices or possibly impacting on workplace relationships
- ☐ Research conducted overseas involving participants
- ☐ Research involving ionising radiation
- ☐ Research involving gametes or use or creation of embryos
- ☑ None of the above

Does the research involve limited disclosure to participants?

○ Yes ◉ No

Does the research involve:

- ☐ Opt out approach
- ☐ Waiver
- ☐ None of the above

Research plan

2. Describe the theoretical, empirical and/or conceptual basis, and background evidence, for the research proposal, eg. previous studies, anecdotal evidence, review of literature, prior observation, laboratory or animal studies.

Although semiology is an important component of epilepsy evaluation, the understanding of the neural basis of semiological expression has advanced surprisingly insufficient (Chauvel & McGonigal 2014). Seizure semiology is still widely analysed by visual inspection and is prone to considerable inter-observer variability. It is not infrequent to see conflicting lateralising or localising signs occurring in a single seizure. It is complex to derive a general model describing which motions indicate a seizure or related behaviour (Tufenkjian & Luders, 2012). Recent contributions in the medical field have been mainly focused on the manual anatomo-electro-clinical analysis of sEEG signal and video analysis (Bonini et. al., 2014). It can be imagined that future progress in semiological analysis, using quantitative methods, coupled with the progress in imaging and electrophysiological analysis could lead to significant shifts in current understanding (McGonigal, 2015).

The fundamental concept of human motion analysis is to detect complex motor patterns by automatic interpretation of the patient's clinical video data. Pediaditis et al. 2012, evaluated different methodologies and techniques using marker-based and maker-free systems, where this study revealed that the quantification of motion patterns in epilepsies is still under development. Equally, Vilas-Boas & Cunha, 2016 identified that clinical qualitative scales are considered reliable for medical practices but still incomplete without effective quantitative motion capture solutions.

Deep learning has become the mainstream in computer vision last several years, surpassing human in image recognition, time-series analysis, and biological applications (Langkvist et al., 2015; Sonderby et. al, 2015).

Human Motion Analysis is a challenging process because of the uncoordinated movement in the patients' body and the marked variation that exists between patients. A good pose estimation system must be robust to heavy occlusion, severe deformation and invariant to changes in the environment. Deep learning architectures have modelled long-range dependencies between variables in structured prediction tasks such as articulated pose estimation (Wei et. al, 2016; Newell et. al, 2016). Similarly, these techniques have applied in the understanding of facial motions, which has the potential to provide an intelligent facial expression system as well as a unique encoding of the dynamic of facial actions (Burkert et. al, 2016; Taigman et. al, 2014). Achilles et al. 2016, developed a novel seizure detection method based on convolutional neural networks, using a combined depth and infrared (IR) sensor, which detect seizure-related static and slow patient motions.

Regards the brain electrical activity analysis, applying deep learning techniques to time-series data is gaining increasing attention (Ma et al., 2015; Stober et al., 2014). Deep learning methodologies address the challenges of the state-of-the-art techniques based on morphology, temporal and spatial context, wavelet transforms or energy analysis. They could be robust enough to analyse high-dimensional with a poor signal-to-noise ratio data and considerable variability between individuals subjects and recording sessions (Stober et al., 2015). Johansen et al., 2016, applied convolutional neural networks (CNN) to learn the discriminative features of spikes, which are related to the location of the epileptogenic network.

In our research, deep learning techniques are the promising candidates to analyse semiology. This methodology could manage challenges of traditional techniques such as occlusion of the body, vulnerability to image noise, sensitive to motion discontinuities and the plane that incidence the camera. Furthermore, deep learning may advance the prognostic value of EEG significantly, in part as it does not depend on 'hand-made' features, as it can learn a feature hierarchical representation from raw data automatically.

3. State the aims of the research and the research question and/or hypotheses, where appropriate.

Movement detection and motion features constitute a reliable source of information to assess seizure semiology (clinical information), where the seizure symptoms could represent the habitual description of the patient. This information correlated with the sEEG analysis (electrical activity) in medically refractory partial epilepsy, could predict the anatomic localisation of the epileptogenic network (anatomical). The robustness and quantitative of the assessment can be improved by using automated techniques that in the last years have shown near-human performance.

Research Question:
How could automatic human motion analysis assess the seizure semiology and the correlation with the automated sEEG analysis for the epileptogenic network location in patients with medically refractory partial epilepsy?

The overall aim is to develop a methodology based on synchronous multi-modal analysis to seek a novel epilepsy assessment procedure that can jointly learn across visually observed semiology patterns of behaviour and brain electrical activity recorded in the sEEG signals
This research will be articulate in four objectives:
(1)Analyse and identify the contribution of the human motion analysis in assessing the semiologic features.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

(2)Analyse and identify the contribution of the electrical brain analysis (sEEG) in assessing the anatomical localisation of the epileptogenic network.
(3)Develop the synchronous analysis techniques to perform correlation between clinical semiology and electrical activity.
(4)Analyse meaningful categorization of semiologic features to predict the likely subset of anatomical areas for pre-surgical localisation.

4. Has this project been undertaken previously?

○ Yes ◉ No

Benefits/Risks

In answering the following questions (Q 5 – 11) please ensure that you address all issues relevant to the type of participants that will be involved in your research project. Refer for guidance to relevant chapters of the National Statement.

5. Does the research involve a practice or intervention which is an alternative to a standard practice or intervention?

○ Yes ◉ No

7. What expected benefits (if any) will this research have for the wider community?

This research can greatly benefit patients and neurologists, as it could attribute subsets of brain networks to the semiology production. This will lead to better understanding of the distinctive types of motor manifestation and increase the diagnostic precision in epilepsy surgery for the achievement of seizure freedom.

It is expected that this outcome would help to validate the opinion of the clinicians in case of an agreement between the diagnostic test and the expert experience or to compel the clinicians to re-examine their choices and look into more detail in the case. In addition, it will be the basis for future developments to perform pre-surgical assessment accurately without the use of invasive monitoring equipment.

8. What expected benefits (if any) will this research have for participants?

There is no direct benefit to the participants at this stage because the results of the research will not affect or change their medical care. However, the project will increase scientific knowledge about how to understand the seizure events and contribute to better treatment outcomes for future patients.

9. Are there any risks to participants as a result of participation in this research project?

◉ Yes ○ No

10. Explain how the likely benefit of the research justifies the risks of harm or discomfort to participants.

The current study is an observational study only. It does not have any foreseeable risk, harm or inconvenience to the people involved. There are no clinical risks beyond normal day-to-day activity on the ward associated with the participation in this project. The participation in this project will not involve a new activity different to the regular monitoring of the patients. The brain electrical activity and the movements recorded will be used to understand the characteristics of epilepsy. All clinical data and patient records obtained would have been collected as a part of their existing treatment plan. The project will increase scientific knowledge about how to understand seizure events and the subsets of brain networks involved for a better diagnostic precision and treatment outcomes for future patients.

Providing photographs for publication in internationally recognised scientific journals will greatly benefit the research community in the study of epilepsy evaluation. By doing this, the research could provide highlights of the findings and outcomes related to the body pose and facial expression evaluation and contribute to improving the diagnostic precision in epilepsy surgery for the achievement of seizure freedom. The captured photos/videos will be used to illustrate how the automatic application works in real-world scenarios of epileptic patients.

The project has the potential to enhance the effectiveness of existing pre-surgical evaluation, and lead to the

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

development of new methodologies. The test bed will also provide an ideal environment to develop and demonstrate prototype systems, which lead to further projects.

11. Are there any other risks involved in this research? eg. to the research team, the organisation, others

○ Yes ◉ No

12. Is it anticipated that the research will lead to commercial benefit for the investigator(s) and or the research sponsor (s)?

○ Yes ◉ No

16. Is there a risk that the dissemination of results could cause harm of any kind to individual participants - whether their physical, psychological, spiritual, emotional, social or financial well-being, or to their employability or professional relationships - or to their communities?

◉ Yes ○ No

Describe the risk and explain how it will be managed.
The minimal risk with the patient participation mainly is that the patient's images could be included in peer-reviewed publications. The patient has the option to provide his/her photographic images by signing a specific consent for publication purpose. These images are sufficiently clear to identify the identity. The captured photos/videos will be used to illustrate how the automatic application works in real-world scenarios of epileptic patients.

The patient's image will be included in peer-reviewed publications and scientific presentations. The researchers guarantee that data will be treated with respect and the publications will only be in high-quality journals, which provide respect and dignity to the patients that participate in the research. However, the researchers understand that once the information is published, it is not possible to control it because it is a public domain, and we are aware that the research community or others could use the images published for different purposes or even committing faults of copyright.

This research ensures that personal information is used and disclosed only in ways which are consistent with privacy principles and will otherwise comply with QUT's privacy obligations under statute. To comply with the ethics policies, the research team will not use the data for other purposes that this research. The researchers guarantee that data will be treated with respect and the publications will only be in high-quality journals, which provide respect and dignity to the patients that participated in the research. In the publication process, private information such as the name of the participant will not be associated with the images published and will not be redistributed.

Monitoring

17. What mechanisms do the researchers / investigators intend to implement to monitor the conduct and progress of the research project?

The researchers will monitor the conduct and progress of the research project through regular team meetings and annual reports to the ethics committees. If any adverse event occurs during the project, the report will be reviewed to ensure that the event does not occur again.

6. PARTICIPANTS

1. Research participants

The National Statement identifies the need to pay additional attention to ethical issues associated with research involving certain specific populations.

This question aims to assist you and the HREC to identify and address ethical issues that are likely to arise in your research, if its design will include one or more of these populations. Further, the National Statement recognizes the cultural diversity of Australia's population and the importance of respect for that diversity in the recruitment and involvement of participants. Your answer to this question will guide you to additional questions (if any) relevant to the participants in your study.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Tick as many of the following 'types of research participants' who will be included because of the project design, or their inclusion is possible, given the diversity of Australia's population. If none apply, please indicate this below.

If you select column (a) or (b), column (c) will not apply.

The participants who may be involved in this research are: *If you select column (a) or (b), column (c) will not apply.*	**a) Primary intent of research**	**b) Probable coincidental recruitment**	**c) Design specifically excludes**
People whose primary language is other than English (LOTE)	☐	☑	☐
Women who are pregnant and the human fetus	☐	☐	☑
Children and/or young people (ie. <18 years)	☐	☐	☑
People in existing dependent or unequal relationships	☐	☑	☐
People highly dependent on medical care	☑	☐	
People with a cognitive impairment, an intellectual disability or a mental illness	☐	☐	☑
Aboriginal and/or Torres Strait Islander peoples	☐	☑	☐
People who may be involved in illegal activity	☐	☑	☐
None apply	☐		

You have indicated that it is probable that
- People whose primary language is other than English (LOTE)
- People in existing dependent or unequal relationships
- Aboriginal and/or Torres Strait Islander peoples
- People who may be involved in illegal activity

may be coincidentally recruited into this project. The National Statement identifies specific ethical considerations for these groups(s).

Please explain how you will address these considerations in your proposed research.
1. The research focus on the behaviour (semiologic production) and brain electrical activity analysis. The project does not involve language assessments. The consent will be explained clearly understanding that English is his/her second language.

2. The project will not seek to deliberately include or exclude Aboriginal or Torres Strait Islander peoples; however, it is possible that some Aboriginal or Torres Strait Islander peoples may be coincidentally recruited. If Aboriginal or Torres Strait Islander peoples participate in the study, the researchers will ensure that the cultural and language diversity of the participants are respected at all times. As per the 6 core values; reciprocity, respect, responsibility, equality, survival and protection, spirit and integrity, as outlined in the Values and Ethic: Guidelines for Ethical Conduct in Aboriginal and Torres Strait Islander Health Research position paper.

3. People who are involved in illegal activity have been listed as "probable coincidental recruitment"; as participants' involvement in illegal activity may not be known to the researcher. (National Statement, S1.4b).

Participant description

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

2. How many participant groups are involved in this research project?

1

3. What is the expected total number of participants in this project at all sites?

20

4. Groups

Group 1

Group name for participants in this group:	Semiology and SEEG Group
Expected number of participants in this group:	4-20
Age range:	18-65

Other relevant characteristics of this participant group:
Participants that are under epilepsy monitoring, eligible for sEEG surgery or already have undergone sEEG surgery and are to be monitored by the Advanced Epilepsy Unit at Mater.
Why are these characteristics relevant to the aims of the project?
The primary aim of this research is to analyse the behaviour of patients that have refractory focal epilepsies (drug resistant) and are eligible for sEEG and epilepsy surgery.

Your response to question 1 at Section 6 - "Research Participants" indicates that the following participant groups are excluded from your research. If this is not correct please return to question 1 at Section 6 to amend your answer.

- Women who are pregnant and the human fetus
- Children and/or young people (ie. <18 years)
- People with an intellectual or mental impairment

5. Have any particular potential participants or groups of participants been excluded from this research? In answering this question you need to consider if it would be unjust to exclude these potential participants.

- Pregnant women and the human foetus will be excluded because these individuals are not considered suitable for neurosurgery (all of the participants is the study will be listed for neurosurgery).

- Children will be excluded because the research will not consider developing brains.

- People with mental impairment will be excluded for consent purpose because they need to understand clearly the consequences of being included in peer-reviewed publications.

Participant experience

6. Provide a concise detailed description, in not more than 200 words, in terms which are easily understood by the lay reader of what the participation will involve.

The participation in this project will not involve a new activity different from the regular clinical monitoring, which includes the record of the brain activity and movements during seizures events. This information is the data that the research will use. Non-Mater researchers will not be in contact with the patients.

Relationship of researchers / investigators to participants

7. Specify the nature of any existing relationship or one likely to rise during the research, between the potential participants and any member of the research team or an organisation involved in the research.

It is possible that some of the participants may be patients of the neurologist and/or neuropsychologist involved in the research project.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

9. Describe what steps, if any, will be taken to ensure that the relationship does not impair participants' free and voluntary consent and participation in the project.

To ensure that the relationship does not impair participants' free and voluntary consent, potential participants will be informed prior to commencing the study that their participation is voluntary and that they are free to withdraw at any time without negative consequences on the care that they are receiving. Potential participants will also be informed that they can choose not to participate in the study and that this will not negatively impact upon their current or future care.

10. Describe what steps, if any, will be taken to ensure that decisions about participation in the research do not impair any existing or foreseeable future relationship between participants and researcher / investigator or organisations.

The decision to participate or to not participated will in no way impact upon the current or future relationship between the patient and Mater or with QUT.

11. Will the research impact upon, or change, an existing relationship between participants and researcher / investigator or organisations?

○ Yes ◉ No

Recruitment

13. What processes will be used to identify potential participants?

To be recruited from the Mater Advanced Epilepsy Unit in Brisbane. The research team will provide the neurology team with a synopsis of the study objectives, i.e. what it entails and the participant inclusion/exclusion criteria. The neurologist will then identify patients who could be candidates for the specific semiologic production and electrical brain activity analysis according to their experience.

14. Is it proposed to 'screen' or assess the suitability of the potential participants for the study?

◉ Yes ○ No

How will this be done?
This will be done by the neurologist in the process of evaluating the patient's eligibility according to the semiologic production.

15. Describe how initial contact will be made with potential participants.

Non-Mater researchers will not be the first point of contact for potential participants. They will be approached first by the neurology team at the Mater Advanced Epilepsy Unit to see if they have an expressed interest in the research.

QUT researchers will be responsible for consenting patients after they have given their verbal consent to be contacted by the neurology team. Those patients who express an interest in the study will be provided with the study information sheet and consent forms.

Potential participants will be informed that participation in the study is entirely voluntary and that they can choose not to participate in the study or withdraw at any time. Potential participants will also be informed that if they choose not to participate in the study, their medical care will not be compromised in any way

16. Do you intend to include both males and females in this study?

◉ Yes ○ No

What is the expected ratio of males to females that will be recruited into this study and does this ratio accurately reflect the distribution of the disease, issue or condition within the general community?
Gender is not a factor of interest and the research team will not have a control in the ratio selection. It is expected a proportion of 50:50.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

17. Is an advertisement, e-mail, website, letter or telephone call proposed as the form of initial contact with potential participants?

○ Yes ◉ No

18. If it became known that a person was recruited to, participated in, or was excluded from the research, would that knowledge expose the person to any disadvantage or risk?

○ Yes ◉ No

Consent process

Do you propose to obtain consent from individual participants for your use of their stored data/samples for this research project?

◉ Yes ○ No

7. Participants Specific

8. CONFIDENTIALITY/PRIVACY

Answers to the questions in section 8.1 will establish whether an HREC will need to apply guidelines under federal or State/territory privacy legislation in reviewing your application. Answers to questions in the remaining parts of section 8 will show how confidentiality of participants is to be protected in your research.

1. Do privacy guidelines need to be applied in the ethical review of this proposal?

Indicate whether the source of the information about participants which will be used in this research project will involve:

- ☑ collection directly from the participant
- ☑ collection from another person about the participant
- ☐ use or disclosure of information by an agency, authority or organisation other than your organisation
- ☐ use of information which you or your organisation collected previously for a purpose other than this research project

Information which will be collected for this research project directly from the participant

Describe the information that will be collected directly from participants. Be specific where appropriate.
Behavioural data from Video monitoring recorded.
Brain electrical activity from EEG and sEEG monitoring recorded.

The information collected by the research team about participants will be in the following form(s). Tick more than one box if applicable.

- ☑ individually identifiable
- ☑ re-identifiable
- ☐ non-identifiable

Give reasons why it is necessary to collect information in individually identifiable or re-identifiable form
The data will be collected in an identifiable and re-identifiable format. Participation is in person, therefore, the researchers know the identity of the participant. The photographic images (including the full video recordings) are sufficiently clear to identify the patient.

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

The video and brain signals data will potentially be re-identified using the hospital code. Each participant will be assigned a number. This number will be used to label in experiments so that the participants are not identified by name.

Information which will be collected for this research project from another person about the participant

Describe the information which will be collected from another person about participants. Be specific where appropriate The data will be collected by technicians or scientist of Mater. This information includes the presurgical evaluation of each patient.
- scalp EEG and video monitoring.
- sEEG surgery history, localisation of the contact and electrodes.
- seizure history (epilepsy type, the number of seizures, duration of seizures - electrical onset and clinical onset).

1c. Will the information be used in medical research?

(•) Yes ( ) No

1d. Does this application include an attachment relevant to state/territory privacy legislation?

( ) Yes (•) No

1e. Is the information health information?

(•) Yes ( ) No

Using information from participants

2. Describe how information collected about participants will be used in this project.

Video-EEG/sEEG monitoring will be used to automatically analyse the anatomo-electro-clinical correlation.
(1) Video monitoring will be used in the quantification of motion patterns or specific motor sign during semiologic production.
(2) EEG/SEEG monitoring will be used to assess the anatomical localisation of the epileptogenic network and the correlation with the semiologic production.

3. Will any of the information be used by the research team be in identified or re-identifiable (coded) form?

(•) Yes ( ) No

Indicate whichever of the following applies to this project:

- [x] Information collected for, used in, or generated by, this project will not be used for any other purpose.
- [ ] Information collected for, used in, or generated by, this project will/may be used for another purpose by the researcher for which ethical approval will be sought.
- [ ] Information collected for, used in, or generated by, this project is intended to be used for establishing a database/data collection/register for future use by the researcher for which ethical approval will be sought.
- [ ] Information collected for, used in, or generated by, this project will/may be made available to a third party for a subsequent use for which ethical approval will be sought.

4. List ALL research personnel and others who, for the purposes of this research, will have authority to use or have access to the information and describe the nature of the use or access. Examples of others are: student supervisors, research monitors, pharmaceutical company monitors.

All personnel listed at the beginning of this application will have access to the data. Collaborators will have access to the deidentified data for the purposes of dissemination/publication once the participant agrees to use his/her

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

image in the consent form.

Storage of information about participants during and after completion of the project

5. In what formats will the information be stored during and after the research project? (eg. paper copy, computer file on floppy disk or CD, audio tape, videotape, film)

All raw images are stored securely with restricted access on QUT HPC facilities and at QUT Research Data Storage, which are password protected. The data will be automatically backup nightly in two physical locations. No permanent data will be stored in cloud services or portable USBs, thus avoiding legal and privacy issues arising from third parties gaining access to the information.

The clinical history, video monitoring, brain electrical activity and this consent form will be stored in separate locations to protect the participant identity. All paper-based information and scan prints such as the Participation Information Sheet/Consent Form will be stored in a locked filing cabinet at the Mater Hospital and at QUT (office of the leader of the SAIVT research group).

At the commencement of the study, all participants will be assigned a number. The number will be used to label any test forms so that the participants are not identified by name.

6. Specify the measures to be taken to ensure the security of information from misuse, loss, or unauthorised access while stored during and after the research project? (eg. will identifiers be removed and at what stage? Will the information be physically stored in a locked cabinet?)

The data will be stored in a password protected places: QUT Research Data Storage and High-Performance Computing facilities (HPC accounts). All paper-based information will be stored in a locked filing cabinet at the Mater Hospital and at QUT (office of the leader of the SAIVT research group).

9. The information which will be stored at the completion of this project is of the following type(s). Tick more than one box if applicable.

- [x] individually identifiable
- [x] re-identifiable
- [ ] non-identifiable

Give reasons why it is necessary to store information in individually identifiable or re-identifiable form.
Understanding that the purpose of this research is to analyse the semiologic production related to movements of interest, the researchers request to store the information in which the patients are sufficiently clear to be identified. The videos captured allow the team to better analyse the semiologic production related to body pose (especially upper limbs) and facial expressions, and better understand their relationship to the brain electrical activity. The personal information and the video data will be stored in different locations.

If the data can be re-identified using a code, specify the security arrangements and access for the code.
All participants will be assigned a number. The number will be used to label any experiment so that the participants are not identified by name. The code description will be held on a password controlled computer.

10. For how long will the information be stored after the completion of the project and why has this period been chosen?

As per NHMRC guidelines, the data will be stored for 7 years at the completion of the study.

11. What arrangements are in place with regard to the storage of the information collected for, used in, or generated by this project in the event that the principal researcher / investigator ceases to be engaged at the current organisation?

If the principal researcher ceases to be employed by the organisation, information pertaining to the project will remain at QUT and Mater Hospital.

Ownership of the information collected during the research project and resulting from the research project

13. Who is understood to own the information resulting from the research, eg. the final report or published form of the results?

The information resulting from the research will be owned by Queensland University of Technology and Mater Hospital.

14. Does the owner of the information or any other party have any right to impose limitations or conditions on the publication of the results of this project?

○ Yes ◉ No

Disposal of the information

15. Will the information collected for, used in, or generated by this project be disposed of at some stage?

◉ Yes ○ No

At what stage will the information be disposed?
After 7 years the data pertaining to this study will be destroyed.
How will information, in all forms, be disposed?
The paper-based material will be shredded, while any information stored digitally that was not published will be erased.

Reporting individual results to participants and others

16. Is it intended that results of the research that relate to a specific participant be reported to that participant?

○ Yes ◉ No

Explain/justify why results will not be reported to participants:
The results of the research will not report to the participants. However, if the participant chose to take part in high-impact internationally peer-reviewed publication, he/she will have the opportunity to view the image as we plan to use to illustrate the automatic applications.

17. Is the research likely to produce information of personal significance to individual participants?

○ Yes ◉ No

18. Will individual participant's results be recorded with their personal records?

◉ Yes ○ No

19. Is it intended that results that relate to a specific participant be reported to anyone other than that participant?

◉ Yes ○ No

To whom will the results be reported other than the participant?
Only if the participant consents to take part in high-impact internationally peer-reviewed publications, the results and images captured of the facial expression and body poses of the patient will be published in high-quality journals and conferences of engineering and medicine.
Explain why the results will be reported to a person other than the participant?
We would like to illustrate some of the benefits and impacts of performing video analytics in specific behaviours for the epilepsy evaluation within peer-reviewed conferences and journals. These images will be used to highlight the findings and outcomes of the research, and will contribute to improving understanding of the epilepsy. The captured photos/videos will be used to illustrate how the automatic application works in real-world scenarios of epileptic

patients.
Will the participant be told that their results will be reported to another person?
(•) Yes ( ) No

20. Is the research likely to reveal a significant risk to the health or well being of persons other than the participant, eg family members, colleagues

( ) Yes (•) No

21. Is there a risk that the dissemination of results could cause harm of any kind to individual participants - whether their physical, psychological, spiritual, emotional, social or financial well-being, or to their employability or professional relationships - or to their communities?

(•) Yes ( ) No

Describe the risk and explain how it will be managed:
The risk with the patient participation mainly is that the researchers know the identity of the participant, and in the publications of the results no personal information will be declared. The research will guarantee that any identifiable information will remain confidential and only the investigators of this study who the participants have given consent to will be able to access the personal information.

If the patient consents to take part in peer-reviewed publications, where he/she agrees to the research to use photographic or video images of his/her identity in academic papers and presentation, the researchers guarantee that data will be treated with respect and the publications will only be in high-quality journals, which provide respect and dignity to the patients that participated in the research.

This research ensures that personal information is used and disclosed only in ways which are consistent with privacy principles and will otherwise comply with QUT's privacy obligations under statute. To comply with the ethics policies, the research team will not use the data for other purposes that this research.

22. How is it intended to disseminate the results of the research? eg report, publication, thesis

It is intended that the experiments and results of the research will be disseminated in the form of publication in peer reviewed academic papers, internationally respected scientific journals of biomedicine and engineering, conferences and thesis document.

23. Will the confidentiality of participants and their data be protected in the dissemination of research results?

(•) Yes ( ) No

Explain how confidentiality of participants and their data will be protected in the dissemination of research results:
The captured photos/videos will be used to illustrate how the automatic application works in real-world scenarios of epileptic patients. Researchers understand that video participants may not wish to be named in this video. As a result, the names of participants will be excluded from their videos or images. Each publication will use a small number of images to describe the results.

The researchers will guarantee that data will be treated with respect in high-quality journals; however, the researchers understand that once the information is published, it is not possible to control it because it is a public domain, and we are aware that the research community or others could use the images published for different purposes. The technology has advanced to the point that from a published image, a search engine could correlated it with public information such as social networks.

Examples of research publications in the state-of-the-art of semiology analysis are Pierre et. al, 2008; Souirti et. al, 2014 and Bonini et. al., 2014.

10. Declarations And Signatures

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

Applicant / Principal Researchers (including students where permitted)

Project Title (in full):	Multimodal analysis of Video-SEEG monitoring for the automatic evaluation of epilepsy.
HREC to which this application is made:	
HREC Reference number:	

I/we certify that:

- All information is truthful and as complete as possible.
- I/we have had access to and read the National Statement on Ethical Conduct in Research InvolvingHumans.
- The research will be conducted in accordance with the National Statement.
- The research will be conducted in accordance with the ethical and research arrangements of the organisations involved.
- The research will be conducted in accordance with the ethical and research arrangements of the organisations involved.
- I/we have consulted any relevant legislation and regulations, and the research will be conducted in accordance with these.
- I/we will immediately report to the HREC anything which might warrant review of the ethical approval of the proposal (NS 2.37), including:
- serious or unexpected adverse effects on participants;
- proposed changes in the protocol; and
- unforseen events that might affect continued ethical acceptability of the project.
- I/we will inform the HREC, giving reasons, if the research project is discontinued before the expected date of completion (NS 2.38);
- I/we will not continue the research if ethical approval is withdrawn and will comply with any special conditions required by the HREC (NS. 2.45);
- I/we will adhere to the conditions of approval stipulated by the HREC and will cooperate with HREC monitoring requirements. At a minimum annual progress reports and a final report will be provided to the HREC.

Applicant / Chief Researcher(s) / Principal Researcher(s)

Principal Researcher section was signed electronically by Professor Sridha Sridharan on 03/11/2016 11:00

Job Title/Post:	Professor
Organisation:	QUT
Email:	s.sridharan@qut.edu.au
Decision/Comments:	

Principal Researcher section was signed electronically by Mr David Ahmedt on 02/11/2016 16:17

Job Title/Post:	PhD Student QUT
Organisation:	Queensland University of Technology
Email:	david.aristizabal@hdr.qut.edu.au
Decision/Comments:	

Principal Researcher section was signed electronically by Dr Kien Nguyen Thanh on 02/11/2016 17:33

Job Title/Post:	Research Fellow
Organisation:	QUT

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Email: k.nguyenthanh@qut.edu.au

Decision/Comments:

Principal Researcher section was signed electronically by Dr Sasha Dionisio on 02/11/2016 16:33

Job Title/Post: Head of Epilepsy Unit

Organisation: Mater

Email: sasha.dionisio@mater.org.au

Decision/Comments: Accepted with thanks

Principal Researcher section was signed electronically by Prof Clinton Fookes on 04/11/2016 01:25

Job Title/Post: Professor

Organisation: QUT

Email: c.fookes@qut.edu.au

Decision/Comments:

	..... Signature	...../ ...../ ..... Date
Professor Clinton Fookes Queensland University of Technology	..... Signature	...../ ...../ ..... Date
Dr Sasha Dionisio Mater Hospital	..... Signature	...../ ...../ ..... Date
Mr David Ahmedt Queensland University of Technology	..... Signature	...../ ...../ ..... Date
Professor Sridha Sridharan Queensland University of Technology	..... Signature	...../ ...../ ..... Date
Mr Kien Nguyen Queensland University of Technology	..... Signature	...../ ...../ ..... Date

Associate Researchers

Supervisor(s) of student(s)

Project Title (in full): Multimodal analysis of Video-SEEG monitoring for the automatic evaluation of epilepsy.

HREC to which this application is made:

HREC Reference number:

Australian National Ethics Application Form (c) 2006 Commonwealth of Australia

Submission Code Date: 07/11/2016 10:54:48 Reference: Online Form

Student Supervisor section was signed electronically by Prof Clinton Fookes on 04/11/2016 01:27

Job Title/Post: Professor
Organisation: QUT
Email: c.fookes@qut.edu.au
Decision/Comments:

I/we certify that:

- I/we will provide appropriate supervision to the student to ensure that the project is undertaken in accordance with the undertakings above;
- I/we will ensure that training is provided necessary to enable the project to be undertaken skilfully and ethically.

Professor Clinton Fookes

..... Signature

...../ ...../ ..... Date

Heads of departments/schools/research organisation

Project Title (in full): Multimodal analysis of Video-SEEG monitoring for the automatic evaluation of epilepsy.

HREC to which this application is made:

HREC Reference number:

I/we certify that:

- I/we are familiar with this project and endorse its undertaking;
- the resources required to undertake this project are available;
- the researchers have the skill and expertise to undertake this project appropriately or will undergo appropriate training as specified in this application.

..... Title ..... First Name ..... Surname

..... Position ..... Organisation Name

..... Signature ...../ ...../ ..... Date

Australian National Ethics Application Form (c) 2006 Commonwealth of Australia

11. Attachments

List of Attachments

Core Attachments	Attachments which may be required/appropriate
Recruitment/invitation	Copy of advertisement, letter of invitation etc
Participant Information	Copy or script for participant Copy or script for parent, legal guardian or person responsible as appropriate
Consent Form	Copy for participant For parent, legal guardian or person responsible as appropriate For, optional components of the project eg. genetic sub study
Peer review	Copy of peer review report or grant submission outcome
HREC approvals	Copy of outcome of other HREC reviews

Attachments specific to project or participant group	Attachments which may be required/appropriate
People whose primary language is other than English (LOTE)	English translation of participant information/consent forms
People highly dependent on medical care	Information/consent form for legal guardian or person responsible
Aboriginal and/or Torres Strait Islander peoples	Evidence of support / permission of elders and/or other appropriate bodies

Participant information elements

Core Elements
Provision of information to participants about the following topics should be considered for all research projects.

Core Elements	Issues to consider in participant information
About the project	Full title and / or short title of the project Plain language description of the project Purpose / aim of the project and research methods as appropriate Demands, risks, inconveniences, discomforts of participation in the project Outcomes and benefits of the project Project start, finish, duration
About the investigators / organisation	Researchers conducting the project (including whether student researchers are involved) Organisations which are involved / responsible Organistions which have given approvals Relationship between researchers and particpants and organisations
Participant description	How and why participants are chosen How participants are recruited How many participants are to be recruited
Participant experience	What will happen to the particant, what will they have to do, what will they

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

	experience? Benefits to individual, community, and contribution to knowledge Risks to individual, community Consequences of participation
Participant options	Alternatives to participation Whether participation may be for part of project or only for whole of project Whether any of the following will be provided: counselling, post research follow-up, or post research access to services, equipment or goods
Participants rights and responsibilities	That participation is voluntary That participants can withdraw, how to withdraw and what consequences may follow Expectations on participants, consequences of non-compliance with the protocol How to seek more information How to raise a concern or make a complaint
Handling of information	How information will be accessed, collected, used, stored, and to whom data will be disclosed Can participants withdraw their information, how, when Confidentiality of information Ownership of information Subsequent use of information Storage and disposal of information
Unlawful conduct	Whether researcher has any obligations to report unlawful conduct of participant
Financial issues	How the project is funded Declaration of any duality of interests Compensation entitlements Costs to participants Payments, reimbursements to participants Commercial application of results
Results	What will particpants be told, when and by whom Will individual results be provided What are the consequences of being told or not being told the results of research How will results be reported / published Ownership of intellectual property and commercial benefits
Cessation	Circumstances under which the participation of an individual might cease Circumstances under which the project might be terminated

Research Specific Elements

Provision of information to participants about the following topics should be considered as may be relevant to the research project.

Specific to project or participant group	Additional issues to consider in participant information
Aboriginal and/or Torres Strait Islander peoples	Describe consultation process to date and involvement of leaderswhether ATSI status will be recorded

Australian National Ethics Application Form (c) 2006
Commonwealth of Australia

Appendix C

Collaborative research

C.1 Facial analysis for psychophysiological research

Abstract: Thermal Imaging (Infrared-Imaging-IRI) is a promising new technique for psychophysiological research and application. Unlike traditional physiological measures (like skin conductance and heart rate), it is uniquely contact-free, substantially enhancing its ecological validity. Investigating facial regions and subsequent reliable signal extraction from IRI data is challenging due to head motion artefacts. Exploiting its potential thus depends on advances in analytical methods. Here, we developed a novel semi-automated thermal signal extraction method employing deep learning algorithms for facial landmark identification. We applied this method to physiological responses elicited by a sudden auditory stimulus, to determine if facial temperature changes induced by a stimulus of a loud sound can be detected. We compared thermal responses with psycho-physiological sensor-based tools of galvanic skin response (GSR) and electrocardiography (ECG). We found that the temperatures of selected facial regions, particularly the nose tip, significantly decreased after the auditory stimulus. Additionally, this response was quite rapid at around 4–5 seconds, starting less than 2 seconds following the GSR changes. These results demonstrate that our methodology offers a sensitive and robust tool to capture facial physiological changes with minimal manual intervention and manual pre-processing of signals. Newer methodological developments for reliable temperature extraction promise to boost IRI use as an ecologically-valid technique in social and affective neuroscience.

- S. Sonkusare, **D. Ahmedt-Aristizabal**, M. Aburn, V. Nguyen, T. Pang, S. Frydman, S. Denman, C. Fookes, M. Breakspear, C. Guo, Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking, *Scientific Reports*, 9 (2019) 4729.

C.2 Body motion analysis for breathing disorders research

Abstract: Recent breakthroughs in computer vision offer an exciting avenue to develop new remote, and non-intrusive patient monitoring techniques. A very challenging topic to address is the automated recognition of breathing disorders during sleep. Due to its complexity, this task has rarely been explored in the literature on real patients using such marker-free approaches. Here, we propose an approach based on deep learning architectures capable of classifying breathing disorders. The classification is performed on depth maps recorded with 3D cameras from 76 patients referred to a sleep laboratory that present a range of breathing disorders. Our system is capable of classifying individual breathing events as normal or abnormal with an accuracy of 61.8%, hence our results show that computer vision and deep learning are viable tools for assessing locally or remotely breathing quality during sleep.

- M. Martinez, **D. Ahmedt-Aristizabal**, T. Väth, C. Fookes, A. Benz, R. Stiefelhagen, A Vision-based System for Breathing Condition Identification: A Deep Learning Perspective, *Proceedings of the IEEE International Conference of Engineering in Medicine and Biology Society (EMBC)*, 2019.