Deep Facial Analysis: A new phase I epilepsy evaluation using computer vision

David Ahmedt-Aristizabal, Clinton Fookes, Kien Nguyen, Simon Denman, Sridha Sridharan and Sasha Dionisio

Abstract

Semiology observation and characterization play a major role in the pre-surgical evaluation of epilepsy. However, the interpretation of patient movements has subjective and intrinsic challenges. In this paper, we develop approaches to attempt to automatically extract and classify semiological patterns from facial expressions. We address limitations of existing computer-based analytical approaches of epilepsy monitoring, where facial movements have largely been ignored. This is an area that has seen limited advances in the literature. Inspired by recent advances in deep learning, we propose two deep learning models, landmark-based and region-based, to quantitatively identify changes in facial semiology in patients with Mesial Temporal Lobe Epilepsy (MTLE) from spontaneous expressions during phase 1 monitoring. A dataset has been collected from the Mater Advanced Epilepsy Unit, (Brisbane, Australia) and is used to evaluate our proposed approach. Our experiments show that a landmark-based approach achieves promising results in analysing facial semiology, where movements can be effectively marked and tracked when there is a frontal face on visualization. However, the region-based counterpart with spatio-temporal features achieves more accurate results when confronted with extreme head positions. A multi-fold cross-validation of the region based approach exhibited an average test accuracy of 95.19% and an average AUC of 0.98 of the ROC curve. Conversely, a leave-one-subject-out cross-validation scheme for the same approach reveals a reduction in accuracy for the model, as it is affected by data limitations and achieves an average test accuracy of 50.85%. Overall, the proposed deep learning models have shown promise in quantifying ictal facial movements in patients with MTLE. In turn, this may serve to enhance the automated pre-surgical epilepsy evaluation by allowing for standardization, mitigating bias and assessing key features. The computer-aided diagnosis may help to support clinical decision making and prevent erroneous localization and surgery.

Keywords: Epilepsy Evaluation, Facial Semiology, Deep Learning, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM).

1. INTRODUCTION

Epilepsy is among the most common of the neurological conditions. Mesial Temporal Lobe Epilepsy (MTLE) often with hippocampal sclerosis, is one of the most common causes of drug-resistant epilepsy [1]. Epilepsy surgery has been accepted as an effective treatment for patients with medically refractory epilepsy or who are non-responsive to medication [2]. The complete resection of the epileptogenic zone (*i.e.* the region of the brain that generates epileptic seizures) is the primary goal. Epileptic patients exhibit different clinical manifestations, based on the underlying networks activated. Semiology has played a pivotal role to provide localizing and lateralizing information in order to allow for successful surgery in addition to neurophysiological and imaging data [3, 4]. While semiology is important, a single sign in isolation is not helpful, but rather the progression of events that

David Ahmedt-Aristizabal, Prof. Clinton Fookes, Dr. Kien Nguyen, Dr. Simon Denman and Prof. Sridha Sridharan are with Speech, Audio, Image and Video Technologies (SAIVT) research program, School of Electrical Engineering & Computer Science, Queensland University of Technology, Brisbane, Australia.
Dr. Sasha Dionisio is with the Department of Mater Advanced Epilepsy Unit, Mater Centre for Neurosciences, Brisbane, Australia.
Address correspondence to David Ahmedt-Aristizabal, Science and Engineering Faculty, School of Electrical Engineering and Computer Science, Queensland University of Technology, S Block, 612, GP, GPO Box 2434, Brisbane, QLD 4001, Australia. E-mail: david.aristizabal@hdr.qut.edu.au.

Fig. 1. Selected example of facial semiology from mouth motion in MTLE patient. (Best in color)

underlie the integration of various neuronal networks. In MTLE for example, certain facial modifications are more commonly exhibited, (although not exclusive) including unilateral blinking, eye deviation, chewing automatisms, fear expression, disgust, unilateral mouth deviation and post ictal nose wiping [5-8].

Epilepsy Monitoring relies on video analysis to assist with the diagnosis of seizures. However, this evaluation is subjective, dependent on observer experience and may lead to misdiagnosis [3]. Automated analysis of semiological patterns, *i.e* detection, quantification and recognition of body movement patterns, could help increase diagnostic precision [9] by standardising the assessment evaluation among evaluators and identifying features that are unambiguous. However, the automated analysis of semiology has made little progress over recent years [10, 11]. The majority of existing automated systems are limited to the ictal analysis of limb and head movements [10, 12]. While some attempts at automating the semiology of facial expressions have been made [13-16], the field is still largely unexplored. One reason for this includes the immense complexity in detecting and tracking key facial regions, especially in the clinical environment, where the face may often be obscured from view with electrodes, bedding, inadequate camera capture and positioning, poor illumination and movements during seizures [10].

Deep learning (DL) has entered the mainstream in computer vision and machine learning in the last several years, achieving near-human and super-human performance in many tasks such as object detection and sequence learning [17]. Using DL, researchers have also demonstrated state-of-the-art performance in analysing videos, outperforming traditional techniques in emotion recognition and facial expression analysis [18-20]. DL is now becoming widely-employed in biological and medical applications; however, despite its advantages, there has not been an application of this technology for the purpose of monitoring facial changes in seizures, particularly in the pre-surgical evaluation. While automated detection of EEG signals based on DL exists to help identify seizures [21, 22], apart from video recordings, there are no devices which detect ictal changes in semiology [10]. DL is a promising field for analysing video data due to its advantages in automatically learning key features extracted from raw data. DL is a technique that can be adapted to new problems because of its ability to perform transfer learning, *i.e.* learning on one dataset and applying the trained model to another. Thus this results in a richer representation and greater learning capability [17, 23].

Given that epilepsy surgery requires considerable accuracy to help the patient [24], this research has been developed to improve the diagnostic precision using quantitative methods from clinical data. In this paper, we endeavour to develop quantitative methods that characterise motion semiology, using facial expression. We have concentrated our techniques to distinguish between ictal and non-ictal/random facial expressions in patients with MTLE. Localization of MTLE was confirmed in these patients with a combination of Stereo-EEG assessment or seizure freedom for over two years in the setting of a lesion, *i.e* Hippocampal sclerosis. The techniques and equipment employed in this study are a combination of video detection systems and advanced computer vision techniques. Deep learning architectures characterize the various semiological patterns from quantitative motion detection and training. The learnt patterns from ictal facial modification are extracted to identify between ictal and non-ictal pattern. We conduct experiments using our own dataset jointly developed by the Queensland University of Technology, Australia (QUT) and the Mater Advanced Epilepsy Unit, Brisbane, Australia. The remainder of this paper is organised as follows: Section 2 describes our dataset, the methodology and experimental plan; Section 3 presents the results and Section 4 discusses the main findings and the significance of the results. Finally, Section 5 draws the paper's concluding remarks.

2. Materials and Methods

2.1 Video Monitoring Dataset

The video recordings were captured as a part of the routine long-term Video-EEG monitoring protocol at the Mater Hospital in Brisbane, Australia with the epilepsy patients who were undergoing phase 1 workup for their drug-resistant epilepsy. The patients were monitored over a time period ranging 2-7 days. The patients were selected if clinical evaluation suggested a possible surgical procedure may be suitable or if invasive studies were necessary to analyse the onset. A random sample of 16 patients with MTLE was retrospectively selected from the overall dataset. Localization of MTLE was confirmed either from a Stereo-EEG evaluation or if a temporal lobectomy had been performed in the setting of hippocampal sclerosis, with seizure freedom of no less than 2 years.

All seizures recorded from MTLE patients 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 illustrated by a selected example in Fig. 1. 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. Instances of seizures in the video recorded for the dataset were selected from the first epileptic discharge until the full expression of semiology prior to version and convulsion if it was experienced. All digitised recorded images from each video clip, recorded at a frame rate of 25 frames/second, were in the PNG format with an image dimension of 1280x720 pixels.

Following this preliminary study, we developed our dataset to perform the quantitative identification between facial expressions during ictal (Class 1) and non-ictal events (Class 2). Table I illustrates the demographic statistics of the 8 patients nominated as Class 1 with the most common ictal pattern, while for Class 2 we randomly selected 8 epileptic patients with video clips recording with non-ictal/random facial expressions such as answering questions from the doctors, eating, watching television and speaking with family members. To avoid incorrect instances of natural behaviour in Class 2, video recordings of interictal periods were not considered. A total of 55 videos clips from the day and night monitoring were recorded, representing 24 videos for **Class 1** and 31 videos for **Class 2**.

TABLE I
MESIAL TEMPORAL LOBE EPILEPSY (**MTLE**) PATIENTS DEMOGRAPHICS DURING ICTAL ACTIVITY

Test Subject	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**	**24425**	

Seizure events are uncommon and public data of semiology is absent. Inadequate training data hinders the validation of algorithms that could quantify semiology. For this reason, in order to exploit the ability of transfer learning of deep learning architectures in the epilepsy task, public datasets traditionally used in the facial analysis under unconstrained conditions were considered to train models used in the proposed approach, and will be described in Sections 2.2.1 and 2.2.2.

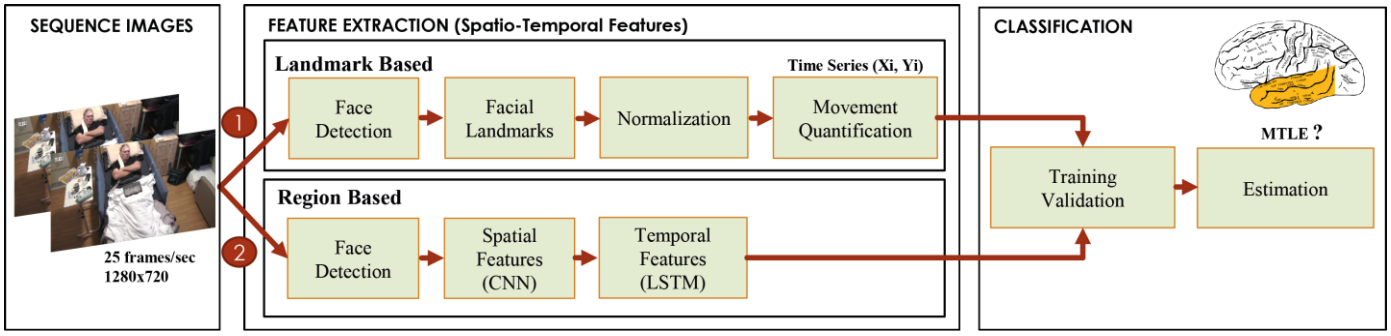

Fig. 2. Two approaches proposed for the automatic analysis of semiology in facial expressions.

2.2 The Proposed System

Given a sequence of colour images of a patient, our method 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 aim of our methodology is to capture such dynamic variation of facial physical structure from consecutive frames. In order to validate the research hypothesis which states that similar semiological patterns are sufficient to categorise patients with MTLE, an experimental design, displayed in Fig. 2, was proposed to assess semiology from facial movements. To analyse the facial expressions, two methods have been considered: landmark-based and region-based. A landmark-based method (geometry information) is based on a detector of anatomical points of reference in the face for the measurement and quantification of facial motions over time. In a 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 convolutional neural network (CNN) and recurrent neural network (RNN). Spatial features provides information in the facial expressions of a single video frame. On the other hand, temporal features exhibits the relationship between facial expressions revealed in consecutive video frames. Both approaches were selected to evaluate and verify which method excels at classifying MTLE patients in the real conditions of clinical monitoring.

Our supervised methodology is divided into two main phases: feature extraction and classification. Feature extraction can be viewed as finding a set of measured data which effectively represent the information content of an observation. The classification phase concerned to which of a set of categories or class a new observation belongs, on the basis of a training and validation set whose class membership is known. In order to extract features from facial expressions, we used well-established models pre-trained with public datasets that have been used with success in facial research different from clinical applications. We analyse a broad range of available techniques of face detectors and facial landmarks estimators with the purpose to train our system to recognise a human 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 our dataset of epileptic patients and conducting a process of fine-tuning. Once the spatio-temporal features are extracted from the sequence of the video clips recorded, training and validation are performed with the aim of training the system to classify facial semiology of MTLE from natural facial expressions.

The **landmark-based method**, which is an important representation of the facial expression, was 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 effective to capture the dynamic variation of the facial physical structure. Pre-existing algorithms based on DL capable of achieving a high accuracy for the facial landmark estimation and alignment were used in the process of landmarks estimation for all frames. The movement detection is represented by the two-dimensional movements, X axis and Y axis, of each landmark during the facial expression. The resulting (X, Y) time series of each

marker are used to perform the movement quantification using metrics based on 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 and a Long Short-Term Memory (LSTM) architecture [25], which is a special type of RNN. 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.

2.2.1 Face Detection

The initial step to estimate 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 monitoring including scale and pose changes, occlusions, and illumination variations. Traditional techniques used for face detection such as cascade-based and deformable part models can be further improved by deep learning [26]. For instance, the widely-used Viola and Jones algorithm [27], provides real-time face detection, but only performs well on frontal and well-lit face images [14, 15]. However, more recent deep learning models better capture non-linear mappings between intrinsic facial features and facial muscle motions.

We have evaluated representative state-of-the-art facial detectors [28-31] based on DL models and compared their performance on our epilepsy dataset using the Average Precision (AP) metric. The face detector proposed in [28] 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 labeled as belonging to the positive class. The face detector is based on the Faster R-CNN architecture [32] which has achieved state-of-the-art object detection accuracy with a reduced running time. The Faster R-CNN consists of two modules: a Region Proposal Network (RPN) which generates a set of object proposals; and an Object Detection Network, based on the Fast R-CNN detector [33] which refines the proposal location. In the RPN, the CNN architecture considered was the VGG-16 model [34]. The RPN can be trained in an end-to-end manner using backpropagation and stochastic gradient descent (SGD) [35]. The face detector [28] was trained on a large scale public dataset, WIDER Face [36], and used the pre-trained ImageNet model, VGG-16 [34], 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 facial datasets FDDB [37] and IJB-A [38]. With the purpose to avoid false face detections of the patient because of the presence of clinical staff and family members inside the patients' room, a preceding phase of human detection is also developed with the unified network Faster R-CNN.

2.2.2 Landmark-Based Method

Once the face detection is completed facial landmark estimation can be performed. Since the pioneering method of [39], Deep Convolutional Neural Networks (DCNN) have been successfully used in facial landmark localization, overcoming limitations of traditional techniques based on generative and discriminative methods. DL 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 any facial landmark initialization [40].

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 [41] including the spatial rotations yaw, pitch and roll. We have assessed a number of benchmark methods for facial keypoint detection based on DL [41-45], analysing metrics such as the mean error and failure rate. In this study, we use the framework known as Tasks-Constrained

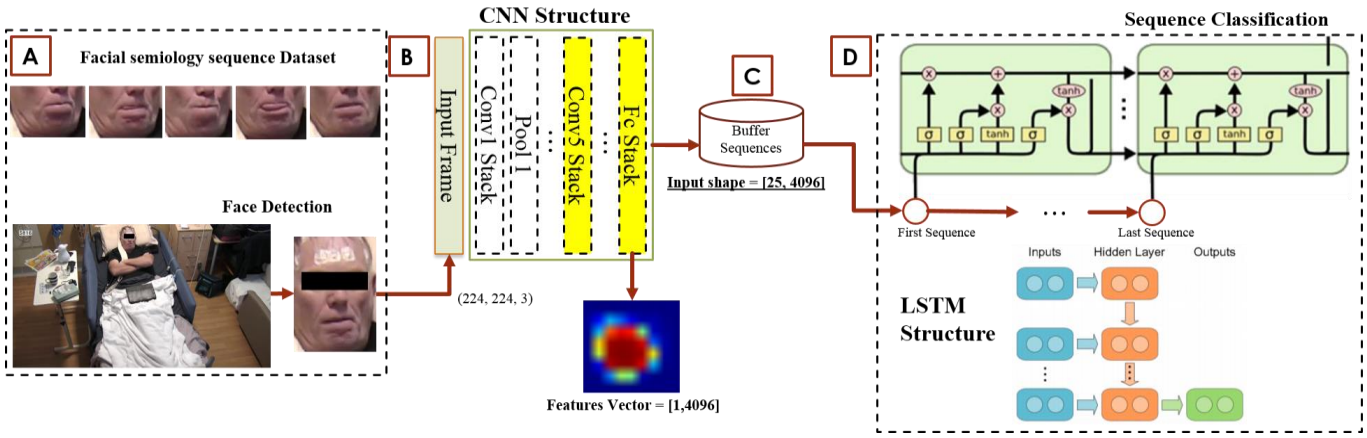

Fig. 3. The proposed framework of the region-based methodology with the CNN-LSTM architecture to classify sequences of facial semiology. **A.** A new dataset of facial semiology is created using face detection, obtaining over 20,000 images for both classes. **B.** With the CNN structure based on the Face Faster R-CNN architecture, the spatial features of the face image (224, 224, 3) 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 feed to a Long-Short-Term-Memory (LSTM) to exploit the temporal relation between video frames, to train and to predict the class of the sequence. (Best in color)

Deep Convolutional Network (TCDCN) [41], which is a near state-of-the-art facial landmark estimator system and returns precise landmark estimations for the faces in the epilepsy dataset. TCDCN incorporates auxiliary information into the fitting process such as head pose estimation or facial attribute inference. This architecture represents a method of transferring the representation from a network pre-trained 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 4 convolutional layers, 3 pooling layers and 1 fully connected layer [41]. The TCDCN model was trained and tested with the MAFL[41], 300-W (IBUG)[46], Helen[47], COFW[48] and AFLW [49] datasets.

Once the landmark estimation is complete, and X-axis and Y-axis trajectories of the facial expressions are extracted 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. Each video sequence has a feature vector with a dimensionality of [1,680], which corresponds 10 features for each of the 68 landmarks. A Support Vector Machine (SVM) is proposed to classify facial semiology from patients with MTLE.

2.2.3 Region-Based Method

The region-based method is based on well-known approaches for visual recognition and description [50], and works by extracting spatio-temporal features from video sequences to predict classes through an end-to-end deep learning model. With this approach, we are classifying video sequences of facial expressions from patients with MTLE. Compared to the method based on facial key points, the performance can be enhanced by feeding the raw frames to deep learning models, allowing the model to learn optimal features from the entire facial area.

The proposed approach uses the CNN fine-tuned for face detection to learn the spatial features of the face; then, these features are linked to an LSTM to exploit the temporal 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 Fig. 3. 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 4096 units.

Once the sequential features are extracted, they are fed to an LSTM network. The number of LSTM layers is one significant hyper-parameter to consider in the LSTM network. We adopt a many-to-one model where multiple stack LSTMs infer one output. We obtained the best performance with a network configured 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.

2.3 Experimental Approach

We fine-tuned the face detector [28], to improve the performance of images recorded during patient night monitoring using a infrared camera which is present in our epilepsy dataset. The fine-tuning was performed only on the last fully connected layer of the VGG-16 architecture to preserve the earlier learnt filters. This process was performed following the instructions in [32, 51]. The framework is implemented in Python and uses the Caffe [51] and OpenCV libraries. Similarly, the dataset of epileptic patients is used to fine-tune the trained model from [41] to conduct the facial landmark estimation.

In the case of the region-based method, 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 optimizing the binary cross entropy loss function. The LSTM was optimised with the ADAM optimizer [52] with a learning factor of 10^{-3}, and decay rate of first and second moments as 0.9 and 0.999 respectively. ADAM has been shown to achieve competitively fast convergence rates when used for multi-layer neural networks. It was found that the stochastic gradient descent optimizer (SGD) yielded worse performance. Dropout [53] with a probability of 0.35 and batch size set to 4 are also used as they are considered to be an effective method for reducing overfitting in deep neural networks when dealing with a huge number of parameters and a small training dataset. We balance the training data at the sequence level using the class weight parameters as in [54]. With an imbalanced dataset, it is probable that without class weights a model will get biased toward the prediction of the no-MTLE patients as this is the dominant case in the dataset. We perform the model training using 30 epochs and use the default initialization parameters from Keras package [54] for initializing the weights of LSTM hidden units. The LSTM is implemented in Python using Keras [54] with a Theano backend [55].

In the experiment, we adopt two approaches for cross-validation to evaluate 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 to that used for training the models. The *k*-fold cross-validation [56] allows us to confirm the reliability of the model by evaluating the approach for facial semiology detection for MTLE on data that has not been seen during training. The leave-one-subject-out cross-validation [57] aims to validate the ability of the trained model to capture subject invariant features such that it can predict whether facial expressions are indicative of MTLE on subjects not seen in the training set. This is the expected clinical scenario when analysing seizures recorded for a new patient, and outputs the probability that the patient has MTLE. For the *k*-fold cross-validation, 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. In a leave-one-subject-out cross-validation scheme, one patient with MTLE is left out as the test subject 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). During the training (without including the test patient), 70% of the data is used for training and 30% for validation.

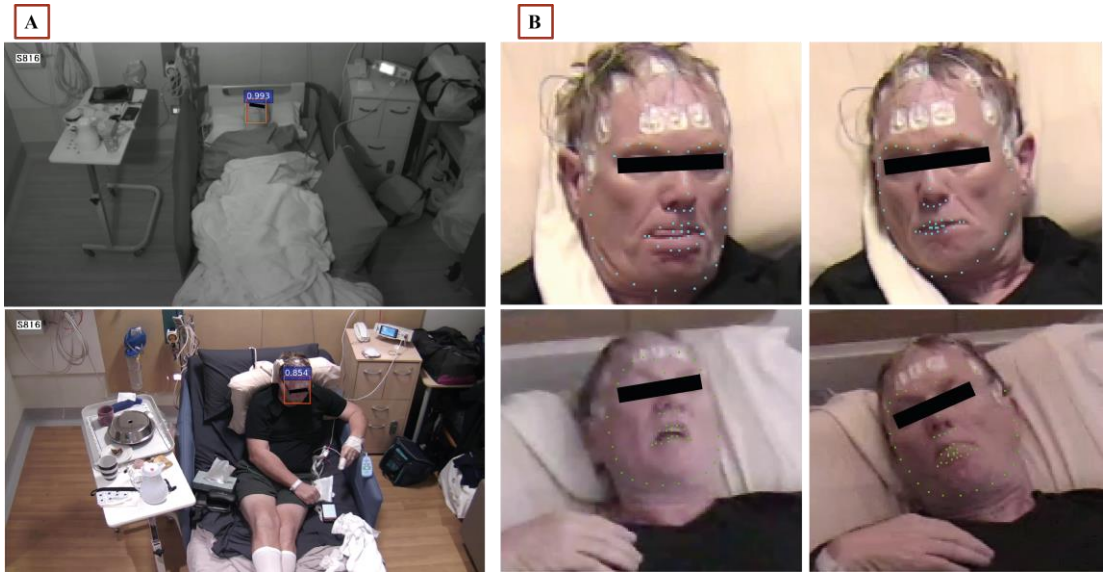

Fig. 4. **A.** Selected examples of face detection using deep learning. Automated face detection results are shown in the yellow bounding boxes. **B.** Selected examples of facial landmark estimation. The landmarks in the eyes are covered to protect the identity of the patient. (Best in color)

3. RESULTS

3.1 Face Detection

The intersection-over-union (IoU) is used to quantitatively evaluate the face detection in the epilepsy dataset. The fine-tuned face detector [28] reached an average accuracy of 0.920 in the IoU, in selected videos manually annotated from the data. Fig. 4A shows the qualitative performance in different clinical scenarios: day and night monitoring.

3.2 Landmark-based Classification

The outcome of the landmarks estimation is the position (x, y) for each facial key-point and the head rotation vector represented by the yaw, pitch and roll angles in degrees. The model has shown efficient qualitative results in faces with changes in illumination and with head rotation across the yaw axis that did not exceed the range of the training datasets, which is $(-35°, +35°)$. Fig. 4B depicts the qualitative performance of the facial landmark estimation in different scenarios in our dataset. 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 algorithm implemented is based on the architecture from [41], and reached an average accuracy of 92% of facial point detection across all the images annotated, which indicates an accurate level performance with semi-frontal faces. 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. As a result, the facial landmark estimation for the 55 videos clips of our dataset was restricted by the performance of the detector. 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%, which suggests that there are not sufficiently discriminative features to classify facial semiology from patients with MTLE. This experiment reveals that even the state-of-the-art landmark detection algorithms in the literature suffer when presented with these extreme pose cases and yield poor detection results.

3.3 Region-based Classification

3.3.1 Multi-fold Cross-validation (K-fold cross-validation)

The LSTM model was capable of achieving an average of 95.19% accuracy on the test set. Table II displays the comparison between the validation accuracy and test accuracy set for each fold. We compared the performance of each model from the multi-fold cross-validation computing the Receiver Operating Characteristic (ROC) curve and calculating the Area Under Curve (AUC). The area measures discrimination, *i.e*. the ability of the test to correctly classify those with and without MTLE epilepsy. For the five different models, the average area under the curve reached a value of 0.9926.

TABLE II
MULTI-FOLD CROSS-VALIDATION PERFORMANCE
MESIAL TEMPORAL LOBE EPILEPSY (**MTLE**) PATIENTS

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**

3.3.2 Leave-one-subject-out cross-validation

Table III illustrates the results of the performance of the validation during training and testing for each subject. The proportion of the subject in the total data indicates the number of video sequences for each patient with MTLE in the dataset for class 1, where subject 8 has more sequences of seizure recorded than any other. The deep learning framework reached an average validation accuracy of 97.69% and an average test accuracy of 50.85%.

TABLE III
LEAVE-ONE-SUBJECT-OUT CROSS-VALIDATION PERFORMANCE
MESIAL TEMPORAL LOBE EPILEPSY (**MTLE**) PATIENTS

Test Subject	Proportion 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. DISCUSSION

Seizure 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 differentiates facial semiology from MTLE patients 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 machine learning in practice. The results of the face detector and facial landmark estimator illustrate the ability of deep transfer learning to adapt models trained on out-domain data 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 method to accumulate and maintain temporal detail when processing sequence data, such as the evolution of a facial expression.

We have shown that the landmark-based method is considerably limited by the landmark estimation performance, and the estimator is appropriate only in certain monitoring situations of our dataset. Although this method is more intuitive because the features extracted are visually related to the amplitude and frequency of the landmark motions and will likely yield good results if the landmarks are consistently detected across challenging scenarios, the performance is diminished in extreme cases of head pose which is a routine reality in the clinical monitoring environment. In the long term, the landmark-based method will be revisited when new, more accurate and robust facial landmark detection algorithms to assess facial expressions are introduced. In contrast, the region-based method has demonstrated robust performance within the challenging, unconstrained (in-the-wild) conditions encountered in the clinical environment including changes in head pose and illumination. This method allows the extraction of sufficient amount of features to conduct a process of categorisation of semiological behaviour.

The high performance of the multi-fold cross-validation method compared with the leave-one-subject-out cross-validation technique for the region-based method (see Tables II and III) has verified the robustness to model variations in the data but at the same time highlights the disadvantage of small datasets when classifying semiology. This is evident from the performance of the leave-one-subject-out cross-validation and the differences between the average validation and test accuracies. In the particular case of the subject 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 patient 1 and 5, good classification results could be achieved due to their semiological patterns being exhibited by other patients in the dataset. This suggests that while performance at present is limited, a larger training dataset that better captures the variety of semilogical patterns that can occur will result in significantly improved performance. The *k*-fold cross-validation method classifies random sequences of any subject with a model that was trained using observations from all subjects, while the leave-one-subject-out cross-validation method evaluates the complete video corpus for one specific subject who is not seen at any moment during the training. This results in the leave one subject out evaluation being more sensitive to the small dataset. Although the region-based method has been affected by the available semiological 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 epileptic patients as the landmark-based approach struggles in the real world conditions encountered in a hospital setting.

The most significant limitation when studying semiology automatically is the underfitting of the models due to inadequate training data. As such, we are currently developing a larger seizure semiology database that may be used to mitigate this problem in the future. It remains to be seen how our model will scale when applied to a much larger dataset to explore facial expressions only during seizure events from different epilepsy types; however, early results are promising.

5. Conclusion

The main objective of this proposal is to provide quantitative motion information for supporting the assessment of epilepsy. In this paper, we have investigated facial-semiology for epilepsy evaluation from the deep learning perspective, overcoming a number of limitations of traditional techniques to quantify facial semiology including the extraction of robust facial features in a clinical monitoring environment, using a 2D video database.

We have proposed two approaches, landmark-based and region-based, using modern deep learning models including CNN and LSTM for enabling the automatic assessment of facial movements during seizures. The validation of the deep network system presented here has confirmed that our method reveals quantified motion patterns from facial expressions that can differentiate ictal semiology from natural expressions in patients with MTLE. This information has value to clinicians in seizure classification and

may aid in the localization of the epileptogenic network by supporting the analysis performed by the epileptologist, particularly in the setting of the pre-surgical evaluation.

A new version of the system will be validated to quantitatively classify Mesial Temporal (mTLE) and Extra-Temporal (ETLE) lobe epilepsies, relying on facial expressions and pose dynamics such as head and upper limb movements. Future work will explore features pertaining to facial expressions, hand and body movements which can be jointly extracted and combined to further enhance patient evaluation. From the clinical point of view, we expect that the quantified movement analysis could extract information that is imperceptible to visual inspection, provide additional evidence in the classification of epileptic seizures and contribute toward a more reliable diagnosis. The automatic computer-aided diagnosis of semiology could also be potentially useful for motion analysis in the evaluation of broader neurological diseases that experience movement disorders.

Acknowledgment

This research was partly supported by Mater Hospital under the Authorisation RG-17-008 and the Australian Research Council Discovery Grant DP140100793. The authors would like to thank QUT High-Performance Computing (HPC) for providing the computational resources for this research.

Ethical approval

We confirm that we have read the Journal's position on issues involved in ethical publication and this report is consistent with those guidelines. Consent was obtained for experimentation with human subjects.

Conflict of interest statement

The authors report no conflicts of interest.

REFERENCES

[1] Tatum IV WO. Mesial temporal lobe epilepsy. J. Clin. Neurophysiol. 2012;29: 356-365.
[2] Wiebe S, Blume WT, Girvin JP, Eliasziw M. A randomized, controlled trial of surgery for temporal-lobe epilepsy. N. Engl. J. Med. 2001;345: 311-318.
[3] Chauvel P, McGonigal A. Emergence of semiology in epileptic seizures. Epilepsy Behav. 2014;38: 94-103.
[4] Tufenkjian K, Lüders HO. Seizure semiology: its value and limitations in localizing the epileptogenic zone. Journal of Clinical Neurology 2012;8: 243-250.
[5] Noachtar S, Peters AS. Semiology of epileptic seizures: a critical review. Epilepsy Behav. 2009;15: 2-9.
[6] Kotagal P, Lüders HO, Williams G, Nichols TR, McPherson J. Psychomotor seizures of temporal lobe onset: analysis of symptom clusters and sequences. Epilepsy Res. 1995;20: 49-67.
[7] Ataoğlu EE, Yıldırım İ, Bilir E. An evaluation of lateralizing signs in patients with temporal lobe epilepsy. Epilepsy Behav. 2015;47: 115-119.
[8] Fogarasi A, Tuxhorn I, Janszky J, Janszky I, Rásonyi G, Kelemen A, Halász P. Age-dependent seizure semiology in temporal lobe epilepsy. Epilepsia 2007;48: 1697-1702.
[9] Cunha JPS, Rémi J, Vollmar C, Fernandes JM, Gonzalez-Victores JA, Noachtar S. Upper limb automatisms differ quantitatively in temporal and frontal lobe epilepsies. Epilepsy Behav. 2013;27: 404-408.
[10] Ahmedt-Aristizabal D, Fookes C, Dionisio S, Nguyen K, Cunha JPS, Sridharan S. Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey. Epilepsia 2017.
[11] Ulate-Campos A, Coughlin F, Gaínza-Lein M, Fernández IS, Pearl P, Loddenkemper T. Automated seizure detection systems and their effectiveness for each type of seizure. Seizure 2016;40: 88-101.
[12] do Carmo Vilas-Boas M, Cunha JPS. Movement Quantification in Neurological Diseases: Methods and Applications. IEEE reviews in biomedical engineering 2016;9: 15-31.
[13] Maurel P, Mcgonigal A, Keriven R, Chauvel P. 3D model fitting for facial expression analysis under uncontrolled imaging conditions. ICPR International Conference on Pattern Recognition 2008: 1-4.
[14] Pediaditis M, Tsiknakis M, Bologna V, Vorgia P. Model-free vision-based facial motion analysis in epilepsy. BioEng 10th International Workshop on Biomedical Engineering 2011: 1-4.
[15] Pediaditis M, Tsiknakis M, Koumakis L, Karachaliou M, Voutoufianakis S, Vorgia P. Vision-based absence seizure detection. EMBC Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2012: 65-68.
[16] Sathyanarayana S, Satzoda RK, Sathyanarayana S, Thambipillai S. Identifying epileptic seizures based on a template-based eyeball detection technique. ICIP IEEE International Conference on Image Processing 2015: 4689-4693.
[17] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521: 436-444.
[18] Rodriguez P, Cucurull G, Gonzàlez J, Gonfaus JM, Nasrollahi K, Moeslund TB, Roca FX. Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification. IEEE Transactions on Cybernetics 2017.
[19] Ghasemi A, Denman S, Sridharan S, Fookes C. Discovery of facial motions using deep machine perception. WACV IEEE Winter Conference on Applications of Computer Vision 2016: 1-7.

[20] Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognition 2017;61: 610-628.
[21] Thodoroff P, Pineau J, Lim A. Learning robust features using deep learning for automatic seizure detection. In: Machine Learning for Healthcare Conference; 2016. p. 178-190.
[22] Ahmedt-Aristizabal D, Fookes C, Nguyen K, Sridharan S. Deep Classification of Epileptic Signals. arXiv preprint arXiv:1801.03610 2018.
[23] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2015;61: 85-117.
[24] Spencer S, Huh L. Outcomes of epilepsy surgery in adults and children. The Lancet Neurology 2008;7: 525-537.
[25] Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems 2017.
[26] Yang S, Luo P, Loy CC, Tang X. Faceness-Net: Face Detection through Deep Facial Part Responses. IEEE Transactions on Pattern Analysis and Machine Intelligence 2017.
[27] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: CVPR Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition: IEEE; 2001. p. I-I.
[28] Jiang H, Learned-Miller E. Face detection with the faster R-CNN. FG IEEE International Conference on Automatic Face & Gesture Recognition 2017: 650-657.
[29] Zhang K, Zhang Z, Li Z, Qiao Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters 2016;23: 1499-1503.
[30] Chen D, Hua G, Wen F, Sun J. Supervised transformer network for efficient face detection. ECCV European Conference on Computer Vision 2016: 122-138.
[31] Li H, Lin Z, Shen X, Brandt J, Hua G. A convolutional neural network cascade for face detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015: 5325-5334.
[32] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Adv. Neural Inf. Process. Syst.; 2015. p. 91-99.
[33] Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 1440-1448.
[34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014.
[35] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989;1: 541-551.
[36] Yang S, Luo P, Loy C-C, Tang X. Wider face: A face detection benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 5525-5533.
[37] Jain V, Learned-Miller EG. Fddb: A benchmark for face detection in unconstrained settings. UMass Amherst Technical Report 2010.
[38] Klare BF, Klein B, Taborsky E, Blanton A, Cheney J, Allen K, Grother P, Mah A, Jain AK. 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; 2015. p. 1931-1939.
[39] Sun Y, Wang X, Tang X. Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. p. 3476-3483.
[40] Burkert P, Trier F, Afzal MZ, Dengel A, Liwicki M. Dexpression: Deep convolutional neural network for expression recognition. arXiv preprint arXiv:1509.05371 2015.
[41] Zhang Z, Luo P, Loy CC, Tang X. Learning deep representation for face alignment with auxiliary attributes. IEEE transactions on pattern analysis and machine intelligence 2016;38: 918-930.
[42] King D. Dlib c++ library. Access on: http://dlib. net 2012.
[43] Deng Z, Li K, Zhao Q, Zhang Y, Chen H. Effective face landmark localization via single deep network. arXiv preprint arXiv:1702.02719 2017.
[44] Shao Z, Ding S, Zhao Y, Zhang Q, Ma L. Learning deep representation from coarse to fine for face alignment. ICME IEEE International Conference on Multimedia and Expo 2016.
[45] Baltrušaitis T, Robinson P, Morency L-P. Openface: an open source facial behavior analysis toolkit. In: WACV IEEE Winter Conference on Applications of Computer Vision: IEEE; 2016. p. 1-10.
[46] Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 faces in-the-wild challenge: Database and results. Image and Vision Computing 2016;47: 3-18.
[47] Le V, Brandt J, Lin Z, Bourdev L, Huang T. Interactive facial feature localization. ECCV European Conference on Computer Vision 2012: 679-692.
[48] Burgos-Artizzu XP, Perona P, Dollár P. Robust face landmark estimation under occlusion. In: Proceedings of the IEEE International Conference on Computer Vision; 2013. p. 1513-1520.
[49] Koestinger M, Wohlhart P, Roth PM, Bischof H. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: ICCV Workshops IEEE International Conference on Computer Vision Workshops: IEEE; 2011. p. 2144-2151.
[50] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 2625-2634.
[51] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia: ACM; 2014. p. 675-678.
[52] Kingma D, Ba J. Adam: A method for stochastic optimization. ICLR International Conference for Learning Representations 2014.
[53] Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 2012.
[54] Chollet F. Keras. In; 2015.
[55] Team TTD, Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A. Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:1605.02688 2016.
[56] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Int. J. Conf. Artif. Intell: Stanford, CA; 1995. p. 1137-1145.
[57] Xu G, Huang JZ. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. The Annals of Statistics 2012;40: 3003-3030.