A Spatio-temporal Approach for Apathy Classification

Abhijit Das, *Member, IEEE*, Xuesong Niu, *Student Member, IEEE*, Antitza Dantcheva, *Member, IEEE*, S L Happy, Hu Han, *Member, IEEE*, Radia Zeghari, Philippe Robert, Shiguang Shan, *Senior Member, IEEE*, Francois Bremond, *Senior Member, IEEE* and Xilin Chen, *Fellow, IEEE*

Abstract—Apathy is characterized by symptoms such as reduced emotional response, lack of motivation, and limited social interaction. Current methods for apathy diagnosis require the patient's presence in a clinic and time consuming clinical interviews, which are costly and inconvenient for both, patients and clinical staff, hindering among other large-scale diagnostics. In this work, we propose a novel spatio-temporal framework for apathy classification, which is streamlined to analyze facial dynamics and emotion in videos. Specifically, we divide the videos into smaller clips, and proceed to extract associated facial dynamics and emotion-based features. Statistical representations/descriptors based on each feature and clip serve as input of the proposed Gated Recurrent Unit (GRU)-architecture. Temporal representations of individual features at the lower level of the proposed architecture are combined at deeper layers of the proposed GRU architecture, in order to obtain the final feature-set for apathy classification. Based on extensive experiments, we show that fusion of characteristics such as emotion and facial dynamics in proposed deep-bi-directional GRU obtains an accuracy of 95.34% in apathy classification.

Index Terms—Apathy, Gated recurrent units, Alzheimer, Spatio-temporal classification, behavioral, cognitive, emotion.

1 Introduction

APATHY is defined as the quantitative reduction of goal-directed activity either in behavioural, cognitive, emotional or social dimensions [1]. Within the cognitive dimension loss of interest is a central feature. Given that it is the same in depression, it is not surprising that apathy and depression often co-occur in several psychiatric, neurological and neurodegenerative conditions. It is therefore pertinent to improve the detection of possible differences. This is the case within the emotional dimension characterized in apathy by a *limited emotional response to positive and negative events*, whereas in depression the emotional response is always present but with emotional expression limits to negative emotion such as sadness.

Apathy is a pervasive neuropsychiatric symptom related to the majority of neurocognitive, neurodegenerative, and psychiatric disorders such as Alzheimer's disease (AD) [2], Parkinson's disease [3], and mild cognitive impairment [4] with nearly 65% of dementia patients exhibiting apathy [5].

While experts suggest that early indication of apathy could improve the intervention effects and decrease the global burden of the disease [6], apathy has been highly underdiagnosed. Its diagnosis is based on interviews with patients and their caregivers through a series of questionnaire sessions. Consequently, such interviews require the patient's presence in a clinic and are time-consuming examination involving clinical personnel. Thus, apathy-diagnosis is costly and logistically inconvenient for patients and clinical staff, which hinders large-scale diagnostics.

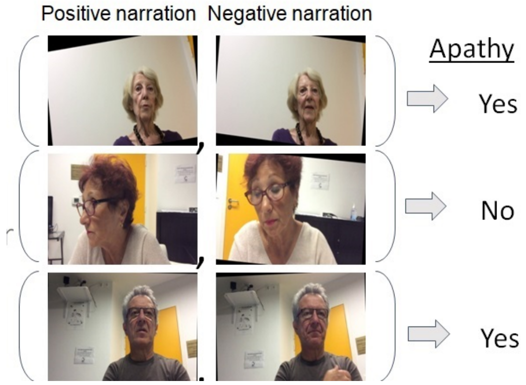


Fig. 1. Example of controlled and apathetic individuals in conversation.

Towards *assisting such subjective assessment*, an automated analysis carries the promise to *enable early apathy diagnosis*, leading to improved intervention effects, potentially increasing the performance of apathy detection in a non-invasive and efficient manner. In addition, such assessment carries the premise to relieve national health-care systems from the excessive workload and allow for large scale early and remote diagnostics.

Motivated by the above, in this article we introduce an automated apathy classification framework based on facial behaviour analysis. Three dimensions of apathy were identified in a recent medical paper by Robert et al. [1], namely (a) behaviour/cognition,

A. Das, A. Dantcheva, SL Happy and F. Bremomd are with STARS team, INRIA Sophia Antipolis, France, E-mail: abhijit.das, antitza.dantcheva, s-l.happy, francois.bremond,{@inria.fr}, X, Niu, H. Han, S. Shan and X. Chen are with Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, E-mail: xuesong.niu@vipl.ict.ac.cn, hanhu, sgshan, xlchen{@ict.ac.cn}, R.Zeghari and P. Robert are with CoBTeK/Memory center CHU, University Cote d'Azur, France, E-mail: radia.zeghari@gmail.com, philippe.robert@univ-cotedazur.fr

Manuscript received ; revised. (Corresponding author: Abhijit Das and Hu Han)

(b) emotion, and (c) social interaction.

We here aim at recognizing (a) the behaviour, as well as (b) the emotional dimension of apathy, characterized by exhibited limited spontaneous expressions, limited emotional responses to positive or negative events, diminished empathy, and reduced verbal or physical reactions to own emotional states. In addition, we explore specific attributes from other dimensions of which is gleaned from facial dynamics/movements. Patients with apathy are in particular less persistent in maintaining a conversation and withdraw often from verbal interaction (social interaction). Thus, we analyze the facial movements and use them as the observation cue of conversation attributed to (c). To validate the reduced emotional response of apathetic subjects, the spontaneous expressions are elicited by asking all subjects to briefly *narrate* past *positive* and *negative experiences*, see Figure 1. The clinical diagnosis of subjects was carried out by psychiatrists in interviews, in addition to the recording of facial videos during positive and negative narrations. We explore the video data for apathy classification using multiple emotion dimensions such as *expression*, *valence-arousal* and *action unit (AU)*, along with other facial behaviour/dynamics such as *eye gaze*, face *pose* and face *movement*. In a nutshell, our proposed approach analyzes patterns of facial expressions and dynamics in elderly subjects towards inferring their apathy state.

While the emotional aspect of apathy has been predominantly explored in previous works, we here study the two dimension of apathy, namely emotion and human facial analysis. We note that determining the state of apathy solely from appearance is highly challenging. We hence place emphasis on analyzing spatio-temporal features in this work. From the limited recent work on automated apathy diagnosis [7]–[9] we conclude that it is highly challenging to characterize the temporal dimension. In turn, we also find that the persistence of facial behaviour is an important factor for the problem in hand. Therefore, we here propose a new architecture that builds on a Gated Recurrent Units (GRUs) [10], which takes statistical features/descriptors as input that in turn describe facial behaviour as extracted from video sequences. Specifically, in seeking to find temporal patterns in facial behavior, we firstly divide the videos into shorter clips, from which statistical descriptors pertaining to face behaviour of each frame are extracted and classified temporally in our proposed GRU-architecture. The statistical features from each clip are fed individually in proposed GRU-architecture, in order to obtain temporal representations of individual features at lower levels of the proposed architecture, which are then combined with deeper layers, in order to obtain the final representation employed for apathy classification. Hence, our hypothesis is that a temporal level facial behaviour assessment can be instrumental in apathy classification, as apathy is among others characterized by *low conversation persistence*, as well as by *withdrawal from verbal interaction* [1].

The contributions of this paper are as follows.

- We are among the first to investigate statistical features from facial behaviour describing emotion and face dynamics towards automatic apathy detection.
- Our proposed GRU-based architecture is targeted to explore temporal synergy, as well as temporal persistence in such features, and exploit them for apathy detection.
- We show that proposed feature fusion and associated bi-directional temporal representation improve classification accuracy.

This work extends our initial studies on apathy detection [7], [8]. In our first work [7] we proposed to use histograms of statistics pertained to emotion and face movements, which we classified based on Support Vector Machine for apathy classification. Following that, in [8] we enhanced emotion and motion features by employing joint learning between apathy features and estimated clinical markers via CNN based multi-task learning. While these two works constituted image-based algorithms, we here explore spatial, as well as temporal features in examining temporal persistence in a wide range of features related to emotion and face dynamics for robust apathy classification.

This work is organized as follows. Section 2 revisits existing work on apathy diagnostics, as well as facial and face dynamics-based motion analysis. Section 3 describes our proposed model, which incorporates face dynamics, as well as emotion feature extraction and associated classification. We present experiments in Section 4 and the related results in Section 4.5 and finally conclude in Section 5.

2 Related Work

Apathy detection. To date, apathy has been determined in clinical interviews, by exposing patients to questionnaires [3]. Such practice is time-consuming and requires the physical presence of a patient at a clinic. The outcome of such diagnosis constitute bio-markers for apathy, as discussed by Hampel et al. [6]. *Neuroimaging* [2], [11] was pertinent in understanding apathy, where structural and functional alteration of frontal-subcortical networks were employed as cues in apathy patients through single-photon emission computed tomography, positron emission tomography, and diffusion tensor imaging [11]. Literature reports the use of the neuroimaging modalities in apathy diagnosis [2], [11]. The correlation of apathy to AD was studied by Aguera-Ortiz et al. [2] using magnetic resonance image analysis.

To mitigate challenges of classical apathy detection, automated apathy detection has been proposed as a novel research area in computer vision, with high impact and interest.

Computer vision-based study of face and gesture has been employed in a set of neurodegenerative disorders [12]–[15]. Apathy classification is a new area of research in the field of computer vision. Happy et al. [7] explored mid-level features from facial behaviour-based motion and emotion analysis to estimate apathy. In addition, they appended the regression of clinical attributes such as *mini mental state examination* (MMSE) and *neuropsychiatric apathy inventory* (NPI-apathy). Subsequently, Happy et al. [8] exploited task relatedness between apathy classification and estimated clinical markers via multi-task learning to obtain robust apathy classification. As opposed to that, Chung et al. [9] introduced an approach based on *visual scanning* behaviour, where sequences of fixations and saccades within and between regions of interest on visual stimuli were analyzed. In particular, a recurrent neural network (RNN) was proposed to learn group difference and individual difference in visual scanning process towards the emotional and non-emotional stimuli. Emotion and motion features were found to be significant in automated apathy classification.

Facial expression recognition and cognitive aspect in health diagnosis. The emotional health of an individual, and hence the ability of an individual to express and identify emotions, plays a vital part in cognitive behavioural therapy [12], [16], [17]. Facial expression recognition acts as an indicator for the internal emotional state, which has been widely explored in literature [18]. Long short-term memory (LSTM) and recurrent neural network (RNN) have been widely used to process sequential data, e.g., temporal analysis in expression and action recognition, as well as scene analysis. Examples for the former include continuous emotion recognition [19], in wild [20] using bi-directorial LSTM, emotion recognition in combination with ECG employing two stream LSTM [21] and LSTM with deep attention [22]. We note that classes in such settings are generally speaking highly distinct from human perspective. This deviates from our setting of interest for apathy classification.

Montenegro et al. [13] analyzed emotion recognition from facial video and electroencephalograph signals for early detection of autobiographical memory deficits in AD. Similarly, mood disorders, such as major depressive disorder and bipolar disorder were investigated employing facial expression analysis [23]. In addition, Montenegro et al. [24] studied emotion from facial depth imagery to investigate cognitive and emotional behaviour. Coco et al. [25] proposed a computational approach for diagnosis and assessment of autism spectrum disorders using facial analysis. Following that, Samad et al. [26] proposed automatic detection of the autism spectrum disorder using spontaneous expression analysis. They concluded that uncontrolled manifestation of smile without proper visual engagement was a fundamental indicator for impairment in social communication. Similarly, reduction in facial expressions or hypomimia was found to be a major cue for estimating the stage severity of Parkinson's disease [27]. Moreover, facial expression features (facial appearance and dynamics) were employed in estimating clinical depression scores [14].

Face dynamics. According to Hammal and Cohn [28], head motion and face dynamics plays an important role in gauging cognitive health. Further, both facial expression and head motion were studied by Adams and Robinson [29] in classifying complex categorical emotions. Following this line of research, Hammal et al. [30] explored the dynamics of head and face movements as cues for positive and negative behaviour. Pitch, toll, and yaw were analyzed based on 49 facial landmarks and related movement. Such cues were found to have a strong correlation with positive and negative behaviour. Similarly, in anxiety detection, head movement, lip deformation, and eyebrow movements were found to be major facial cues [31]. The results indicated that head, as well as eyes and mouth movements were distinct indicators for anxiety and stress. The work of Dibeklioglu et al. [32] extracted face, head movements, as well as speech features in order to detect the severity of depression, encoded in behaviour patterns. Anis et al. [15] analyzed 3D head motion by tracking facial landmarks, extracted associated histograms of velocity and acceleration intensities in estimating three levels of chronic depression severity. Specifically GMM, Fisher vectors and Support Vector Machine (SVM) were tested in classifying the 3 levels of chronic depression severity. Conclusions related to the fact that velocity and acceleration of facial movement were able to strongly map onto depression severity symptoms, which was found consistent with clinical data and theory.

3 Proposed method

Deviating from the above, we here propose a novel approach for apathy detection from video that employs face and behavioural dynamics-based statistical features/descriptor and their temporal relation. In what follows, we describe the basic steps of our proposed method, namely feature extraction and classification. The overall diagram is illustrated in Figure 2.

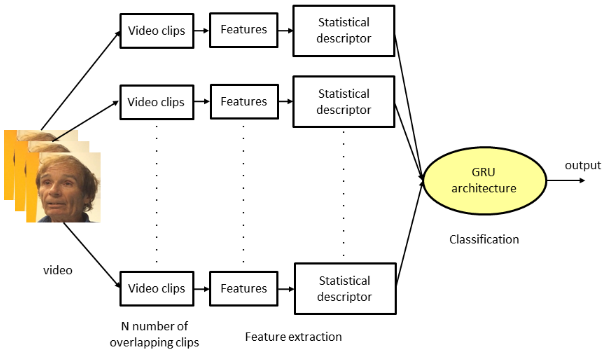


Fig. 2. Overall proposed framework for apathy classification incorporating video level pre-processing, feature extraction, statistical descriptors and classification.

In our setting, classifying apathy based on facial appearance and behavior entails subtle categories, which are challenging to distinguish, even for clinical experts. Motivated by such subtle categories, as well as by the limited duration of videos we have access to, we propose to employ statistical features from video segments/clips to represent videos in our dataset. We note that temporal patterns associated with the collected face videos are weak for classification and related persistence is of higher significance. The representation is fed separately to the *proposed GRU-based architecture*, which is based on bi-directional deep temporal classifier and their temporally related features are combined in deeper levels, used for final classification.

3.1 Feature Extraction

For extracting features, we firstly divide each face video into N number of overlapping segments/clips. Hence from each video segment with F number of frames, we firstly calculate the per-frame features (described below) and obtain temporal information by calculating statistical inference/descriptor for each clip. In this context, standard deviation, mean and max variation pertained to feature intensities represent the statistical descriptor. Heuristically, we choose a sliding window size of 150 frames. Non-uniform length of videos is tackled by employing dynamic overlapping of sliding windows. Such dense sampling is necessitated, as we seek to find temporal statistical synergy among clips. While overlapping clips contain naturally redundant information, at a local look, i.e., features pertained to eye gaze and AUs are significantly different and hence the statistical representation is not repetitive.

In the following, we proceed to explain employed features related to (a) face dynamics, as well as (b) emotion.

3.1.1 Eye gaze

We postulate that eye gaze is a pertinent cue in apathy detection, as it represents an indicator for human interaction and emotion. Towards exploiting eye gaze, we firstly seek to obtain frame-wise gaze prediction. Moreover, in assessing gaze change, we monitor gaze shift w.r.t. an arbitrary target. We note that gaze constitutes more than eye movement. A gaze shift can include movement of the head and torso, all of which must be considered, so that the overall movement is natural, and the emotional display is coherent.

Hence, in this paper, we consider the eye gaze direction vectors and gaze direction in radians in world coordinates predicted for both, left and right eye. In order to generally analyze the eye state of subjects, we further calculate the temporal trajectory of the mean locations of all the eye 2D and 3D landmarks.

3.1.2 Head pose

Similar to eye gaze, head pose constitutes a pertinent cue in assessing social interaction and hence we include related features in our framework. Specifically, head location and head pose direction vector of each frame represent the pose feature. In addition, 2D and 3D 68 facial landmarks are calculated and mean positions of all landmarks render the head pose state.

3.1.3 Local and global motion

We estimate rigid head movements by tracking the facial landmarks. Hereby, nose and inner eye corner landmarks are employed for computing rigid head movement, referred to as global head motion. In contrast, the non-rigid facial landmark movements, associated to lips, eyes, eyebrows, and chin characterize interaction or emotion expression. Specifically, the average movement of facial landmarks around these regions in successive frames is computed as the local motion feature. In addition to mean, and variance, we also utilize in the statistical descriptors minimum, maximum, mean, median, skewness and kurtosis as motion representation using motion information (separately for global and local motion). Further, we append the b-bin histograms of motion values with the intention of preserving motion intensity distribution information in the motion representation, thereby creating a vector of b+7 dimensions. An example of extracted head pose, AUs, eye gaze and detected landmarks is presented in Figure 3.

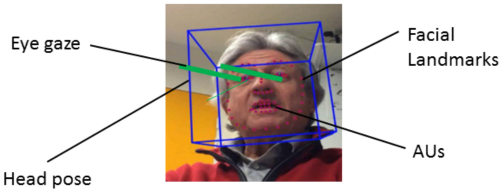


Fig. 3. Face dynamics extracted by our approach: gaze, AUs, landmarks and head pose.

3.1.4 Action Units

Given that an individual with apathy usually exhibits blunted emotion, we expect that action units (AUs) will be an additional indicator for apathy, which we proceed to exploit. Specifically, we adapt the information from 18 AUs in our framework, namely AU 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45, as introduced by Ghayoumi and Bansal [33] for expression analysis. Consequently, we consider individual AUs and their respective frequency as features.

We note that gaze, AUs and pose features are computed based on the OpenFace 2.0 toolkit[1].

3.1.5 Valence-arousal (VA)

W.r.t. emotion, we extract VA in our framework. *Valence* in terms of emotion may relate to the intrinsic attractiveness of goodness i.e., positive valence or averseness i.e., badness (negative valence) in an event or situation. In addition, the specific term describes the tone of feelings, affect, certain behaviours (for example, approach and avoidance), goal attainment or non-attainment. The term also characterizes specific emotions. Theoretically, valence can measure stress as negative valence or oppositely pleasure or happiness as a positive valence. Therefore, here valence is beneficial in indicating exhibited emotions i.e., negativity or passivity, as reflected by the subject.

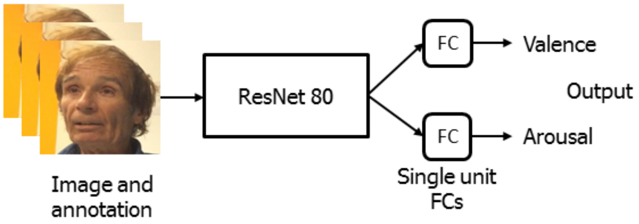


Fig. 4. Architecture used for VA feature extraction.

Arousal is defined as the physiological and psychological state of awoken or sense stimulated in terms of emotion. Therefore, arousal is instrumental in measuring degree consciousness for an expressed emotion.

Intuitively VA are powerful measures for characterizing mental states and in our work. States of arousal can be positive and negative, therefore can be considered as an important tool for our use case. Hence, it can be concluded from the above discussion that the VA will be a highly effective feature for our problem in hand.

Consequently, we here extract a single value for valence and a single value for arousal in each frame. Details are presented in Fig. 4. In particular, we use a ResNet-80 architecture, which at the end incorporates two single-unit fully connected layers, for valence and for arousal, respectively. Moreover, similar to described features, we extract the statistical information pertained to VA for each video clip, forming a descriptor. Again, related statistics include mean, standard deviation, max variation of the feature intensities, minimum, maximum, median, standard deviation, skewness, and kurtosis from VA-intensities are employed to obtain the statistical descriptor. We train our VA-model via AffectNet dataset [34] and then proceed to extract VA in our dataset.

3.1.6 Emotion Features

Classically, emotion recognition has been studied based on the six-expression model [18]. However, based on discussion with clinicians, we found that in our context this model is not suitable and simplified the model to a three category-model of expressions, including *positive* (compounding *happy* and *surprise*), *negative* (compounding *angry, disgust, fear* and *sadness*), and *neutral*, as introduced by Happy et al. [7]. This reduced choice of emotion categories stemmed from the rather flat expressions exhibited by participants.

Therefore, we train a convolutional neural network (CNN) model for *expression classification* with these *three categories*. In particular, a pre-trained VGG-Face [35] is utilized for transfer learning, as it has shown to be robust in facial feature extraction against variations such as pose and illumination [36], [37]. In order to fine-tune the last few layers of the network we introduce a set of skip-connections to the architecture. The details of the CNN architecture [7] used in our experiments are illustrated in Figure 5. The log probabilities of the Softmax layer i.e., the emotion intensities corresponding to each category are represented as a

[1] https://github.com/TadasBaltrusaitis/OpenFace

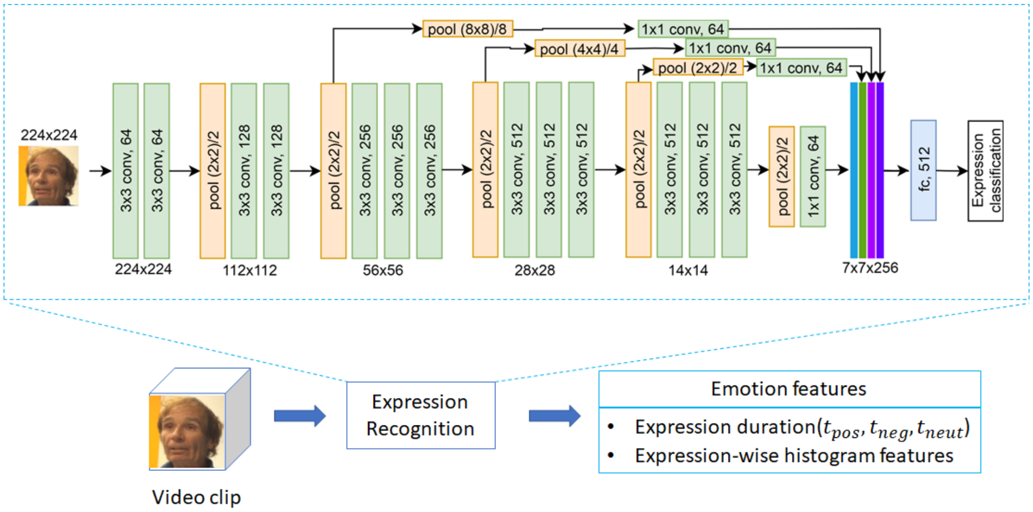


Fig. 5. Facial expression recognition framework. (conv: convolutional layer, fc: fully convolutional layer, pool: pooling layer)

TABLE 1
Summary of features used for apathy detection and related extracted statistical measures.

Feature	Cues	Dimension	Statistical measures
Gaze	gaze direction vectors	12	standard deviation, mean and the max variation (SDMMV)
	gaze direction in radians	4	
	2D & 3D eye landmarks	10	
Pose	head location	6	
	head pose vector	6	
	2D & 3D facial landmarks	10	
AU	AU presence frequency	18	
	AU intensity	51	
Emotion	emotion and duration	45	SDMMV+ histogram bin
VA	VA intensities	18	SDMMV+ min max, median, skewness, and kurtosis
Local and global motion	landmark movement	40	SDMMV+ min max, median, skewness, and kurtosis histogram bin

histogram vector. In addition to standard deviation, a mean and max variation of each clip, a histogram vector for each expression and its duration are considered here. For each clip b bins of a histogram for each expression are extracted, which are further combined to jointly obtain $3 \times b$ dimensional feature vector for 3 classes as a representation of expression intensities for each clip. The duration of a dominant expression is calculated as the e-th expression that is dominant for n_e number of frames out of total N number of video frames. Then we formulate $t_e = \frac{n_e}{N}$ as the expression duration of e-th expression. The expression duration $(t_{pos}, t_{neg}, t_{neut})$ is appended to the expression representation, resulting in a $3 \times (b+1)$ dimensional feature vector.

A detailed summary of the features used for apathy detection is enlisted in Table 1.

3.2 Feature Classification Framework

As mentioned above, apathy entails three dimensions, namely behavioural/cognitive, emotion, as well as social interaction. Therefore, we consider both, local and global information, proposing a dynamic temporal modelling employing small video segments, which extracts the high level statistical descriptor (their persistence) analyzing face behavior and such features are fed to the proposed GRU structure to classify a video.

GRU is a popular gating technique employed in recurrent neural networks [38], which have been successfully employed in natural language processing [39] and speech signal modelling [40]. GRU is composed of a cell, a reset gate and an update gate as illustrated by the following equation.

$$z_t = \sigma_t(W_z x_t + U_z h_{t-1} + b_z) \quad (1)$$

$$r_t = \sigma_t(W_r x_t + U_r h_{t-1} + b_r), \quad (2)$$

where x_t denotes an input vector; h_t represents an output vector; W, U and b are weight parameters and bias vector, respectively; σ_t denotes a sigmoid activation function; z_t is the update gate and r_t is the reset gate. The candidate activation vector is defined as follows.

$$\hat{h}_t = \phi_t(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h), \quad (3)$$

where the operator $\odot$ denotes the Hadamard product. Hence, the output vector is defined as

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t. \quad (4)$$

As illustrated in Fig. 6, we streamline the GRU-structure to model the temporal relation between adjacent video segments and obtain a temporal relation of the features extracted and then concatenate them by timestamps per feature for final classification. Next, we proceed to explain the overall GRU architecture.

First layer. The input feature of this layer is a different feature generated for each segment, and the output of the GRU cell is the refined feature vector as a fully connected layer. The fully connected network containing 128 hidden layers is targeted to obtain temporal features for each segment for each feature. Hence in our context the reset gate, update gate and candidate activation vector is defined as

$$z_n = \sigma_n(W_z x_{nf} + U_z h_{n-1} + b_z), \tag{5}$$

$$r_n = \sigma_n(W_r x_{nf} + U_r h_{n-1} + b_r), \tag{6}$$

$$\hat{h_n} = \phi_n(W_h x_{nf} + U_h(r_n \odot h_{n-1}) + b_h), \tag{7}$$

where x_{nf} represents the input vector for the $n-th$ clip from the $f-th$ feature. The output of the GRU is mapped to a fully connected layer (FC). The output of the 1st layer of the GRU architecture for fth feature is

$$O_{1f} = FC_{1f}, FC_{2f}, ..., FC_{nf}. \tag{8}$$

Second or deep layer. At this layer of the GRU network we concatenate the temporal representation of each feature per clip, as obtained from the previous layer. In other words, fully connected features obtained from the first clip for all features are concatenated, and which act as input of the GRU unit of this layer. Similarly, all features from each clip are concatenated and fed to a GRU unit. Hence in our context the reset gate, update gate and candidate activation vector are defined as

$$z_n = \sigma_n(W_z(FC_{11} \bigoplus FC_{12} \bigoplus ... \bigoplus FC_{1f}) + U_z h_{n-1} + b_z) \tag{9}$$

$$r_n = \sigma_n(W_r(FC_{11} \bigoplus FC_{12} \bigoplus ... \bigoplus FC_{1f}) + U_r h_{n-1} + b_r) \tag{10}$$

$$\hat{h}_n = \phi_n(W_h(FC_{11} \bigoplus FC_{12} \bigoplus ..._{1f}) + U_h(r_{n-1}) + b_h), \tag{11}$$

where FC_{11} signifies the fully connected feature from the first layer of the GRU from the first clip, first feature and $\bigoplus$ denotes the concatenation operation.

Then, the output of the GRU is passed to another set of 128 unit hidden layer fully connected network containing one hidden layer and concatenated to a single-cell FC output layer representing classification. The classification label of the whole video is computed as a combination of all FCs from the second set of the fully connected layer. Specifically, we employ a binary cross-entropy loss for the classification and hence the single-cell value acts as output. Both, first and second layered GRU units in our GRU architecture are bi-directional for related bi-directional gated unit. The output feature dimension of our GRU is set to 512, and the hidden layer of the fully connected layer is 128. Rectified linear unit (ReLU) is used as the activation function, and a dropout layer with the dropout rate of 0.5 is applied to avoid overfitting.

4 Experiments and Discussion

In this section we proceed to describe the employed dataset, preprocessing techniques employed both at video and image level, implementation details, performance measures, evaluation strategy and results obtained, as well as insight and analysis of the rigorous experiments performed in this work.

4.1 Dataset Description

The dataset was recorded at the Nice Memory Research Center located at the Institute Claude Pompidou in the Nice University Hospital. Patients suffering from subjective memory complaint to severe cognitive impairment were included in the study. Demographics and clinical details pertained to the subjects are provided in Table 2. Among apathy and control subjects, the number of female patients were 38% and 62%, respectively.

The patient-clinician interview involved (i) the collection of demographic details, (ii) a standardized neuropsychological assessment, and (iii) a short positive and negative experience narration. The one-on-one interview included the (ii) completion of a battery of cognitive tests including apathy and general behavioural scales (Apathy Inventory, Neuropsychiatric Inventory (NPI-apathy) and classical cognitive tests (Mini-Mental State Examination (MMSE), Clinical Dementia Rate Scale) To elicit spontaneous facial expressions, the participants were asked to narrate some positive and negative events or experiences from their past ("tell me a positive/negative event of your life in one minute"). Videos were acquired at a frame rate of 30fps by tablets (IPADs of fifth and sixth generation) controlled by the clinician. The tablet and environment were kept uniform. We note that the acquired videos include natural pose variations, facial occlusions (e.g., through hands), as well as relatively subtle expressions, see for examples Figure 1. We had to rerecord few videos, where high motion from the subject was exhibited.

4.2 Pre-processing

Prior to the main framework, we detect faces from the dataset-videos using MTCNN proposed in [41], followed by face alignment by positioning both eyes at a fixed distance parallel to the horizontal axis. The aligned faces are re-sized to 224 × 224 resolution, constituting the input for all further feature extraction. To pre-process videos, for each long video, we firstly use an overlapping sliding window to get n segments. For each segment, we calculate features and use the features for further computation. A small value of n will lead to missing local information, whereas a large value of n would cause the model to fail to obtain the general information. In this paper, we choose n=150 as the appropriate number of segments.

4.3 Implementation details

For *emotion*, the CNN model is trained to classify the face into three expression classes, namely positive, negative and neutral. We use in-the-wild (AffectNet [34]) dataset to train the CNN model. The Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.0001 is used for training.

The facial landmarks are detected using DLIB [42] library, in order to compute the *motion* features. In motion and emotion feature extraction, we consider the histograms with 10 bins ($b =$

TABLE 2
Demographic data of patients used in experiments. The mean values are reported with corresponding standard deviations in parenthesis.

	Number of Patients	Age	MMSE	NPI-Apathy
Apathy	18	73.5 (7.7)	22.6 (3.1)	6.2 (2.6)
Control	27	71.7 (8.8)	25.4 (3.6)	0.4 (0.8)

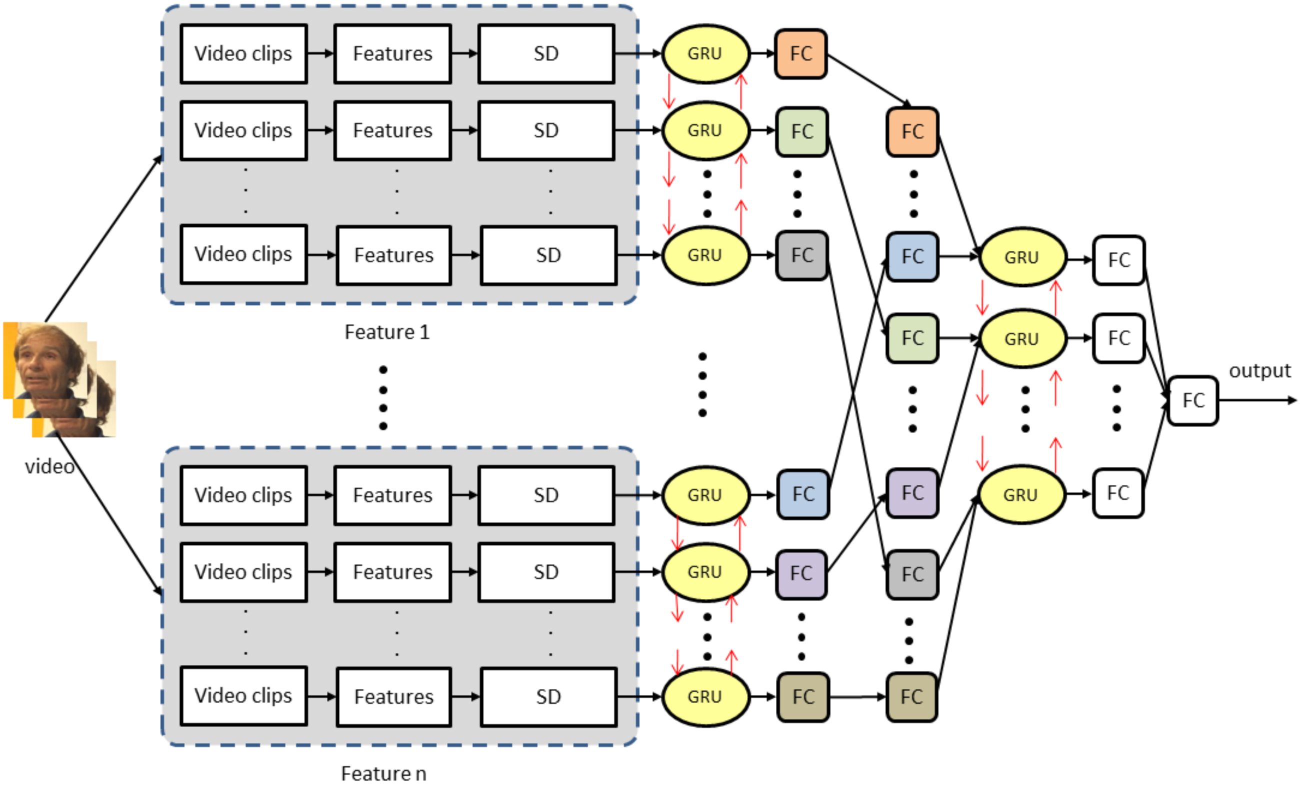


Fig. 6. The proposed GRU-based architecture for apathy classification. Videos are provided as input, which are divided into clips. Features are extracted from each clip-frame, based on which a statistical descriptor is formed, which serves as input to the proposed GRU architecture.

10) for both motion and emotion feature extraction. The extracted features are further normalized to zero mean and unit variance before feeding into the classifier.

From the OpenFace 2.0 toolkit *Gaze*, *Pose* and *AU* features are extracted. For the *VA* model, L1 loss with Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.001 is used.

For the GRU-based architecture for classification binary cross-entropy loss function is used and with Adam optimizer with an initial learning rate of 0.001. The max iteration is set to 20000. The model performing best on the validation set is chosen for testing.

4.4 Evaluation Strategy and Performance Measure

Given the limited number of samples in the dataset, we evaluate the performance using leave-one-subject-out (LOSO) [32], in which samples from one patient constitute the testing set, while the remaining data is used to train the model. All classification models in our experiments are trained with LOSO. For validation, the positive and negative videos from one person are kept apart from training. Therefore, while performing the experiment fusing the positive and negative narration videos from 43 subjects, 86 videos are used for training, the remaining positive and negative narration from one subject each is used for validation and testing. Without fusion, we have 43 videos for training, 1 each for validation and testing. When in the non-fused experiment, negative and positive narration are not classified in the same category, the subject is considered as misclassified. We report experimental results w.r.t. average of accuracy, F1-score, and area under the curve (AUC).

4.5 Experimental Results

We here proceed to describe the extensive experimental study of various feature performances. Firstly, we present a performance comparison of various features without fusing the features of positive and negative narrations with different architectures of GRU (single layer, single layer bi-directional and deep layer) along with our proposed architecture (deep-bi) in Table 3. Here, the performance without fusion is obtained using features from individual videos for classification. The best results are found employing VA (90.69%) followed by emotion and AU. We note that performance of motion features are less accurate. Among motion features, gaze and local features produce the best results followed by pose and global features. The reason behind such performance is that emotional blunting is more prominent in apathetic patients. Moreover, for most features and for most performance measures, the results improve employing deeper GRU and further by the bi-directional counterpart. Improvement by employing bi-directional GRU solicits that the direction of the temporal synergy of this feature is effective. As shown in Table 3, all performance metrics improve in most cases, when features are fused from both narrations. We also observe the improvement in F1-score and AUC both with emotion and face dynamics-based features.

As per Table 4, the combination of features improves the performance significantly. In single layers and its bi-directional version feature fusion is done by stream approach, i.e. the output of the GRU for each features are mapped to an FC layer which acts as the final feature. For the sake of brevity, we report only the combination that produces best improvement. The combination of gaze, pose, AU, VA and emotion naturally induces best result, with feature fusion. Specifically, the feature combination achieves an

TABLE 3
Apathy classification accuracy of different features with and without the fusion of features from positive and negative narration.

		Without fusion of positive and negative narration			After fusion of positive and negative narration		
Features used	GRU type	Accuracy	F1-score	AUC	Accuracy	F1-score	AUC
Local Motion	Single layered	69.76	0.568	0.697	72.09	0.601	0.718
	Single-bi	69.76	0.580	0.694	69.76	0.628	0.708
	Deep layered	72.09	0.599	0.701	72.09	0.641	0.721
	Proposed	74.41	0.615	0.727	74.41	0.670	0.739
Global Motion	Single layered	62.79	0.578	0.627	69.76	0.639	0.698
	Single-bi	65.11	0.590	0.651	72.09	0.641	0.711
	Deep layered	67.44	0.599	0.671	72.09	0.650	0.712
	Proposed	69.76	0.619	0.690	74.41	0.661	0.710
Emotion	Single layered	81.39	0.809	0.812	83.72	0.831	0.831
	Single-bi	83.72	0.821	0.829	83.72	0.831	0.837
	Deep layered	86.04	0.851	0.857	86.04	0.858	0.861
	Proposed	88.37	0.881	0.882	88.37	0.880	0.881
Gaze	Single layered	69.76	0.599	0.699	72.09	0.641	0.720
	Single-bi	72.09	0.605	0.711	74.41	0.655	0.731
	Deep layered	72.09	0.634	0.721	76.74	0.673	0.755
	Proposed	74.41	0.655	0.737	76.74	0.670	0.761
Pose	Single layered	62.79	0.556	0.623	69.76	0.597	0.690
	Single-bi	69.76	0.561	0.657	72.09	0.621	0.703
	Deep layered	69.76	0.575	0.671	74.41	0.645	0.733
	Proposed	72.09	0.599	0.701	0.768	0.695	0.750
AU	Single layered	79.06	0.771	0.793	81.39	0.810	0.811
	Single-bi	81.39	0.790	0.800	83.72	0.830	0.832
	Deep layered	83.72	0.821	0.822	86.04	0.859	0.859
	Proposed	86.04	0.855	0.854	88.37	0.889	0.879
VA	Single layered	83.72	0.821	0.823	86.04	0.850	0.857
	Single-bi	86.04	0.855	0.856	88.37	0.876	0.878
	Deep layered	88.37	0.880	0.881	88.37	0.881	0.883
	Proposed	90.69	0.891	0.897	90.69	0.895	0.908

accuracy of 95.34%, 0.945 of F1 score and AUC of 0.949, which is almost 5% accurate, 0.07 in F1 measure and 0.04 in AUC higher when employed individually. This shows the complementary nature included in face dynamics and emotion features. However, the performance is reduced, when both global and local motions features are taken into account with the above mentioned feature combination. This might be due to the redundant information that is encoded in local and global motion features. Another pertinent observation is that the use of emotion features or a combination with face dynamics always outperforms the motion features. Similar to the previous scenario, feature fusion for most combinations and for most performance measures bring to the fore improved accuracy employing deeper GRU and its bi-directional version. In addition, the performance metrics improve in most cases, when features are fused from both narrations.

4.6 Detailed analysis

We proceed to analyze videos that were miss-classified by the best model. We find that few videos are misclassified, among them one of an apathetic individual and another one from a healthy individual. Both videos are of a short duration <20 sec, for which reason we assume that misclassification occurred. In other words, we believe that misclassification is due to the short duration of the videos.

The remaining analysis is conducted for the best feature combination, as reported in Table 4, namely Gaze+AU+Pose+VA+emotion. Given that we deal with an imbalanced-class classification problem, F1 is the representative performance metric to evaluate the accuracy of the different feature combinations. However, the best feature combination exhibited also superiority w.r.t. the other metrics too. The associated confusion matrix is depicted in Figure 7, further demonstrating the effectiveness of our proposed method. The obtained results are highly encouraging and have been received with excitement by involved clinicians.

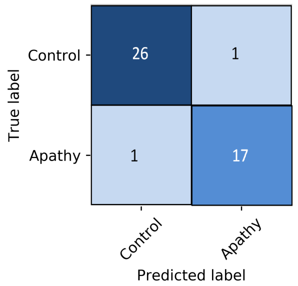


Fig. 7. The confusion matrix of the proposed method for best feature combination.

The loss curves of the best classification model employing the proposed GRU architecture are illustrated in Figure 8. It can been concluded from the curves that the proposed method converges by 30 epochs, hence it does not require large training duration, even when trained from scratch.

TABLE 4
Apathy classification accuracy of different feature fusion by proposed version of GRU architecture based classification with and without the fusion of features from positive and negative narration.

Features used	GRU type	Without fusion of positive and negative narration			After fusion of positive and negative narration		
		Accuracy	F1-score	AUC	Accuracy	F1-score	AUC
Global+Local Motion	Single layered	62.79	0.588	0.629	69.76	0.639	0.698
	Single-bi	65.11	0.599	0.658	72.09	0.649	0.716
	Deep layered	67.44	0.599	0.676	72.09	0.657	0.716
	Proposed	69.76	0.623	0.699	74.41	0.669	0.715
Global+Local Motion+VA	Single layered	79.06	0.789	0.793	81.39	0.817	0.818
	Single-bi	81.39	0.799	0.806	83.72	0.837	0.838
	Deep layered	83.72	0.827	0.828	86.04	0.861	0.867
	Proposed	86.04	0.857	0.859	88.37	0.870	0.871
Global+Local Motion+AU	Single layered	79.06	0.791	0.793	81.39	0.821	0.826
	Single-bi	81.39	0.789	0.811	83.72	0.845	0.845
	Deep layered	83.72	0.810	0.831	86.04	0.870	0.877
	Proposed	86.04	0.846	0.856	88.37	0.881	0.886
Global+Local Motion+Emotion	Single layered	81.39	0.801	0.801	83.72	0.839	0.821
	Single-bi	83.72	0.827	0.822	83.72	0.841	0.841
	Deep layered	86.04	0.841	0.841	86.04	0.861	0.867
	Proposed	88.37	0.861	0.869	88.37	0.889	0.887
Gaze+AU+Pose	Single layered	83.72	0.828	0.831	86.04	0.860	0.858
	Single-bi	86.04	0.851	0.851	88.37	0.882	0.870
	Deep layered	88.37	0.881	0.880	90.69	0.904	0.901
	Proposed	90.69	0.900	0.894	93.02	0.930	0.923
Gaze+AU+Pose+Emotion	Single layered	86.04	0.856	0.850	88.37	0.873	0.875
	Single-bi	88.37	0.870	0.878	90.69	0.901	0.901
	Deep layered	90.69	0.891	0.891	93.02	0.923	0.922
	Proposed	90.69	0.891	0.889	93.02	0.932	0.921
Gaze+AU+Pose+VA	Single layered	86.04	0.858	0.859	88.37	0.881	0.880
	Single-bi	88.37	0.873	0.881	90.69	0.903	0.904
	Deep layered	90.69	0.895	0.897	93.02	0.930	0.929
	Proposed	90.69	0.896	0.893	93.02	0.939	0.929
Gaze+AU+Pose+VA	Single layered	86.04	0.857	0.859	88.37	0.879.	0.882
+Emotion	Single-bi	88.37	0.879	0.881	90.69	0.903	0.901
	Deep layered	90.69	0.895	0.899	93.02	0.928	0.929
	Proposed	90.69	0.905	0.901	**95.34**	**0.945**	**0.949**

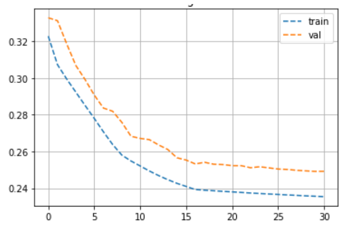


Fig. 8. Loss curve of the proposed method for best feature combination.

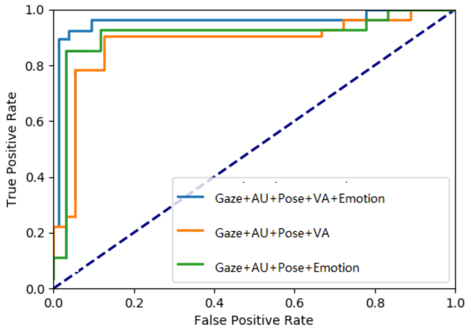
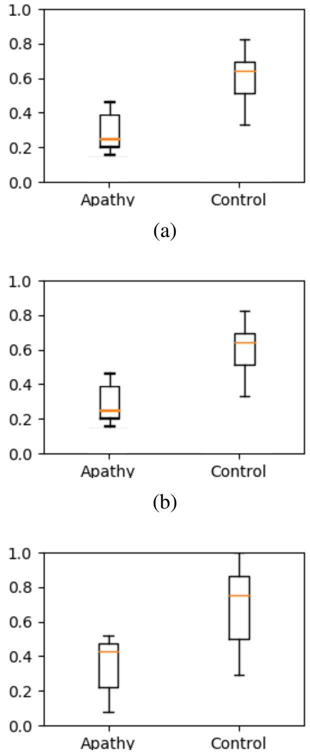


Fig. 9. ROC-curve for the proposed architecture for best feature combination.

Next, we analyze the receiver operating characteristic curve (ROC) of different proposed architectures, for the purpose of determining the best feature combination. Figure 9 provides the detailed ROC plots, indicating the effectiveness of our framework.

We here present a fine-grained analysis of the classification confidence as indicated by the box plots in Figure 10 and Figure 11. On the y-axis we have the prediction confidence of the best model and on the x-axis are the groups of healthy and apathetic individuals. The obtained confidence of output prediction is in the range of y[0 1] scale for each prediction (the original prediction from the binary cross entropy loss). A prediction closer to zero signifies stronger apathy, whereas closer to 1 signifies control or healthy individuals.

We observe that the median probabilities are gravitating towards better prediction (closer to 0 for apathy, and closer to 1 for control individuals) for most settings. This exhibits that our model entails a satisfactory discrimination in predictions. In addition, we do not find outliers in the plot. Further, we proceed to analyze the plot for each video category (positive and negative narration), as well as related fusion for male and then for female. The plots indicate that males exhibit more distinct features as opposed to females. However, for both, male and female, the fusion of positive and negative narration videos brings to the fore higher apathy classification accuracy. While negative videos are more informative than positive videos for female population whereas for male population it was the opposite.

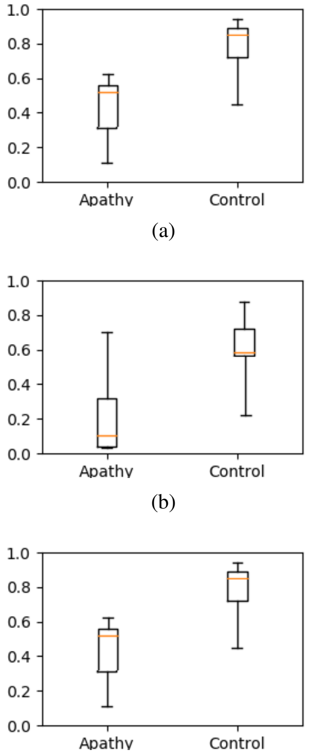


(a)

(b)

(c)

Fig. 10. Box plots of classification confidence for **females** pertained to (a) fused negative and positive narration video, (b) negative narration videos, and (c) positive narration videos.



(a)

(b)

(c)

Fig. 11. Box plots of classification confidence for **male** pertained to (a) fused negative and positive narration video, (b) negative narration videos, and (c) positive narration videos.

4.7 Comparison with the state-of-the-art

For comparison with state-of-the-art, we select the most related work to ours. All reported results are conducted with the same dataset, as well as the same evaluation strategy. We provide comparison results w.r.t. F1 and AUC, as summarised in Table 5. Our proposed method outperforms other models, such as the one presented by Chung et al. [9], as well as by Happy et al., [7], [8]. In particular, we note that the classification-accuracy of our method is 10-15% superior, 0.17-0.11 better with respect to F1 and 0.11-0.15 better in terms of AUC. Moreover, in our previous work we employed clinical scores along with mid-level features related to facial behavior based emotion and face dynamics. Hence, such a solution necessitates clinical information and the participation of practitioners to some extent. In contrast, we only use features extracted from face videos. Moreover, our proposed method showcases the lowest false positive and false negative rates. In addition, we compare our work with four face dynamic-based emotion recognition works [19]–[22], as well as an emotion based cognitive health analysis work [43], on the same dataset and using the same protocol. Our proposed method outperforms named works.

5 Conclusions

This work aims at advancing current apathy diagnosis methods, which require the patient's presence in a clinic and necessitate time-consuming clinical interviews, which are inherently costly and inconvenient for both, patients and clinical staff. We presented a novel automatic apathy detection method, employing statistical descriptors based on facial emotion and dynamic features, classified by a GRU. We validated our model on videos from healthy and apathetic individuals, who narrated negative and positive episodes in their lives. To characterize apathy based on such videos, we extracted a set of features including expression, eye gaze, pose and AUs. An apathy classification model was firstly trained without and then with fusing positive and negative videos. Best classification accuracy was yielded by fusing features pertained to emotion, VA, AU, gaze and pose and classifying them by a deep-bidirectional version of GRU. Our model significantly outperformed state-of-art frameworks.

Our work carries the premise for automatic and efficient apathy diagnostics. Future work involves the development of models related to few shot learning and weakly annotation refinement,

TABLE 5
Comparison with state-of-the-art methods w.r.t. recognition accuracy, F1 and AUC.

Method	Accuracy	F1	AUC
[7]	83.72	0.836	0.833
[9]	79.06	0.771	0.790
[8]	80.00	0.786	0.791
[19]	67.00	0.711	0.721
[22]	71.00	0.706	0.722
[20]	73.00	0.733	0.729
[21]	80.00	0.816	0.801
[43]	71.00	0.717	0.731
Proposed method	**95.34**	**0.945**	**0.949**

in order to mitigate dependence on annotation. Further, we will explore the correlation of features from different facial regions for apathy detection.

We note that the diagnosis and follow-up of affect and motivation disorders such as apathy is an important issue, especially when associated with cognitive impairment. Today, in addition to biological markers such as brain imaging, there are already many sensors of potential interest. The simultaneous assessment of speech and facial expression is particularly important. In this context the combination of video and audio sensors is a challenge that must be addressed in future.

ACKNOWLEDGEMENT

A. Das was supported by the Franco-Chinese research project FER4HM funded by Inria and CAS. SL Happy was supported by the French Government (National Research Agency, ANR) under grant agreement ANR-17-CE39-0002. H. Han was supported in part by the External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843, FER4HM).

REFERENCES

[1] P. Robert, K. Lanctôt, L. Agüera-Ortiz, P. Aalten, F. Bremond, M. Defrancesco, C. Hanon, R. David, B. Dubois, K. Dujardin et al., "Is it time to revise the diagnostic criteria for apathy in brain disorders? the 2018 international consensus group," European Psychiatry, vol. 54, pp. 71–76, 2018.

[2] L. Agüera-Ortiz, J. A. Hernandez-Tamames, P. Martinez-Martin, I. Cruz-Orduña, G. Pajares, J. López-Alvarez, R. S. Osorio, M. Sanz, and J. Olazarán, "Structural correlates of apathy in alzheimer's disease: a multimodal mri study," International journal of geriatric psychiatry, vol. 32, no. 8, pp. 922–930, 2017.

[3] J. Pagonabarraga, J. Kulisevsky, A. P. Strafella, and P. Krack, "Apathy in parkinson's disease: clinical features, neural substrates, diagnosis, and treatment," The Lancet Neurology, vol. 14, no. 5, pp. 518–531, 2015.

[4] G. Cipriani, C. Lucetti, S. Danti, and A. Nuti, "Apathy and dementia. nosology, assessment and management," The Journal of nervous and mental disease, vol. 202, no. 10, pp. 718–724, 2014.

[5] P. H. Robert, F. R. Verhey, E. J. Byrne, C. Hurt, P. P. De Deyn, F. Nobili, R. Riello, G. Rodriguez, G. B. Frisoni, M. Tsolaki et al., "Grouping for behavioral and psychological symptoms in dementia: clinical and biological aspects. consensus paper of the european alzheimer disease consortium," European Psychiatry, vol. 20, no. 7, pp. 490–496, 2005.

[6] H. Hampel, R. Frank, K. Broich, S. J. Teipel, R. G. Katz, J. Hardy, K. Herholz, A. L. Bokde, F. Jessen, Y. C. Hoessler et al., "Biomarkers for alzheimer's disease: academic, industry and regulatory perspectives," Nature reviews Drug discovery, vol. 9, no. 7, p. 560, 2010.

[7] S. L. Happy, A. Dantcheva, A. Das, R. Zeghari, P. Robert, and F. Bremond, "Characterizing the state of apathy with facial expression and motion analysis," in 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2019.

[8] S. L. Happy, A. Dantcheva, A. Das, F. Bremond, R. Zeghari, and P. Robert, "Apathy classification by exploiting task relatedness," in 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2020.

[9] J. Chung, S. A. Chau, N. Herrmann, K. L. Lanctôt, and M. Eizenman, "Detection of apathy in alzheimer patients by analysing visual scanning behaviour with rnns," in ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018, pp. 149–157.

[10] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[11] C. Theleritis, A. Politis, K. Siarkos, and C. G. Lyketsos, "A review of neuroimaging findings of apathy in alzheimer's disease," International psychogeriatrics, vol. 26, no. 2, pp. 195–207, 2014.

[12] E. Hill, P. Dumouchel, and C. Moehs, "An evidence-based toolset to capture, measure and assess emotional health." 2011.

[13] J. M. F. Montenegro, A. Gkelias, and V. Argyriou, "Emotion understanding using multimodal information based on autobiographical memories for alzheimer's patients," in Asian Conference on Computer Vision. Springer, 2016, pp. 252–268.

[14] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," IEEE Transactions on Multimedia, 2018.

[15] K. Anis, H. Zakia, D. Mohamed, and C. Jeffrey, "Detecting depression severity by interpretable representations of motion dynamics," in International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 739–745.

[16] Y. Wang, A. Dantcheva, J.-C. Broutart, P. Robert, F. Bremond, and P. Bilinski, "Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders," in Workshops of the European Conference on Computer Vision (ECCVW). Springer, 2018, pp. 144–157.

[17] A. Dantcheva, P. Bilinski, H. T. Nguyen, J.-C. Broutart, and F. Bremond, "Expression recognition for severely demented patients in music reminiscence-therapy," in 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017, pp. 783–787.

[18] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 8, pp. 1548–1568, 2016.

[19] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," Image and Vision Computing, vol. 31, no. 2, pp. 153–163, 2013.

[20] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "Lstm for dynamic emotion and group emotion recognition in the wild," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 451–457.

[21] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on eeg using lstm recurrent neural network," Emotion, vol. 8, no. 10, pp. 355–358, 2017.

[22] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia, "Facial emotion recognition using light field images with deep attention-based bidirectional lstm," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3367–3371.

[23] Q.-B. Hong, C.-H. Wu, M.-H. Su, and K.-Y. Huang, "Exploring macroscopic fluctuation of facial expression for mood disorder classification," in Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). IEEE, 2018, pp. 1–6.

[24] J. M. F. Montenegro, B. Villarini, A. Gkelias, and V. Argyriou, "Cognitive behaviour analysis based on facial information using depth sensors," in International Workshop on Understanding Human Activities through 3D Sensors. Springer, 2016, pp. 15–28.

[25] M. Del Coco, M. Leo, P. Carcagnì, P. Spagnolo, P. L. Mazzeo, M. Bernava, F. Marino, G. Pioggia, and C. Distante, "A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders," in IEEE International Conference on Computer Vision Workshops, ICCVW 2017, vol. 2018-January, 2018, pp. 1401–1407.

[26] M. D. Samad, N. Diawara, J. L. Bobzien, J. W. Harrington, M. A. Witherow, and K. M. Iftekharuddin, "A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 2, pp. 353–361, 2018.

[27] R. Prashanth and S. D. Roy, "Novel and improved stage estimation in parkinson's disease using clinical scales and machine learning," Neurocomputing, vol. 305, pp. 78–103, 2018.

[28] Z. Hammal and J. F. Cohn, "Intra-and interpersonal functions of head motion in emotion communication," in Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges. ACM, 2014, pp. 19–22.

[29] A. Adams and P. Robinson, "Automated recognition of complex categorical emotions from facial expressions and head motions," in International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2015, pp. 355–361.

[30] Z. Hammal, J. F. Cohn, C. Heike, and M. L. Speltz, "Automatic measurement of head and facial movement for analysis and detection of infants' positive and negative affect," Frontiers in ICT, vol. 2, p. 21, 2015.

[31] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis, "Stress and anxiety detec-

tion using facial cues from videos," Biomedical Signal Processing and Control, vol. 31, pp. 89–101, 2017.

[32] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," IEEE journal of biomedical and health informatics, vol. 22, no. 2, pp. 525–536, 2018.

[33] M. Ghayoumi and A. K. Bansal, "Unifying geometric features and facial action units for improved performance of facial expression analysis," arXiv preprint arXiv:1606.00822, 2016.

[34] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, 2017.

[35] O. M. Parkhi, A. Vedaldi, A. Zisserman et al., "Deep face recognition." in BMVC, vol. 1, no. 3, 2015, p. 6.

[36] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in Automatic Face & Gesture Recognition (FG 2017), 2017 8th IEEE International Conference on. IEEE, 2017, pp. 118–126.

[37] H. Han, S. Shan, X. Chen, and W. Gao, "Illumination transfer using homomorphic wavelet filtering and its application to light-insensitive face recognition," in IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2008, pp. 1–6.

[38] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[39] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," arXiv preprint arXiv:1702.01923, 2017.

[40] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[42] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.

[43] X. Niu, H. Han, J. Zeng, X. Sun, S. Shan, Y. Huang, S. Yang, and X. Chen, "Automatic engagement prediction with gap feature," in Proceedings of the 2018 on International Conference on Multimodal Interaction. ACM, 2018, pp. 599–603.

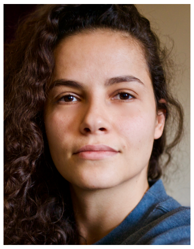
Abhijit Das is currently working as a visiting scientist at the Indian Statistical Institute, Kolkata. Previously, he has worked as a researcher at the University of Southern California, as a Post-Doc Researcher at Inria Sophia Antipolis– Méditerranée, France and as a Research Administrator at University Technology of Sydney, Australia. He has completed his PhD from the School of Information and Communication Technology, Griffith University, Australia. During his research career, he has published several scientific articles in conferences, journals and a book chapter and has also received several awards.

Xuesong Niu received the B.E. degree from Nankai University and is pursing the Ph.D. degree from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. His research interests include computer vision, machine learning and affective computing

Antitza Dantcheva is a Research Scientist (CRCN) with the STARS team of INRIA Sophia Antipolis, France. Previously, she was a Marie Curie fellow at Inria and a Postdoctoral Fellow at the Michigan State University and the West Virginia University, USA. She received her Ph.D. degree from Télécom ParisTech/Eurecom in image processing and biometrics in 2011. Her research is in computer vision and specifically in designing algorithms that seek to learn suitable representations of the human face in interpretation and generation. She is recipient among others of the Best Poster Award at IEEE FG 2019, winner of the Bias Estimation in Face Analytics (BEFA) Challenge at ECCV 2018 (in the team with Abhijit Das and Francois Bremond) and Best Paper Award (Runner up) at the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2017).

S L Happy has completed the joint MS – PhD degree from Indian Institute of Technology Kharagpur, India in 2018. Currently, he is working as a postdoctoral researcher at Inria Sophia Antipolis, France. His research interests include machine learning, computer vision, hyperspectral image classification, medical image analysis, and facial expression analysis.

Hu Han is an Associate Professor of the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). He received the B.S. degree from Shandong University, and the Ph.D. degree from ICT, CAS, in 2005 and 2011, respectively, both in computer science. Before joining the faculty at ICT, CAS in 2015, he has been a Research Associate at PRIP lab in the Department of Computer Science and Engineering at Michigan State University, and a Visiting Researcher at Google in Mountain View. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics and medical image analysis. He has authored or co-authored over 60 papers in refereed journals and conferences including IEEE TPAMI/TIP/TIFS/TMI/TBIOM, CVPR, ECCV, NeurIPS, and MICCAI. He was a recipient of the IEEE Signal Processing Society Best Paper Award (2020), IEEE FG 2019 Best Poster Presentation Award, and CCBR 2016/2018 Best Student/Poster Awards. He is a member of the IEEE.



Radia Zeghari a post doc researcher at CoBTeK. Her PhD was related to the use of new technologies to assess apathy in neurocognitive disorders. She is interested in new assessment methods of cognitive dysfunctions and neuropsychiatric disorders using digital biomarkers for instance. She is also currently working on projects involving remote cognitive and psychiatric assessment.

Philippe Robert is professor of Psychiatry at the Nice School of Medicine, Director of the Nice Memory Centre for Care and Research (CMRR), Director of the Cognition, Behaviour & Technology Unit (CoBTeK) at the Nice-Sophia Antipolis, and president of the Association IA (Innovation Alzheimer – affect – Autism). His domains of expertise include behavioural and psychological symptoms of dementia, apathy assessment and treatment, and the use of new technologies for diagnosis and stimulation of patients with neuropsychiatric diseases.

Shiguang Shan received Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He has been a full Professor of this institute since 2010 and now the deputy director of CAS Key Lab of Intelligent Information Processing. He is also a member of CAS Center for Excellence in Brain Science and Intelligence Technology. His research interests cover computer vision, pattern recognition, and machine learning. He has published more than 300 papers, with totally more than 15,000 Google scholar citations. He served as Area Chair for many international conferences including ICCV'11, ICASSP'14, ICPR'12/'14/'19, ACCV'12/'16/'18, FG'13/'18/'20, BTAS'18 and CVPR'19/'20. And he has been Associate Editor of several journals including IEEE T-IP, Neurocomputing, CVIU, and PRL. He was a recipient of the China's State Natural Science Award in 2015, and China's State S&T Progress Award in 2005 for his research work.

Francois Bremond received the PhD degree from INRIA in video understanding in 1997, and he pursued his research work as a post doctorate at the University of Southern California (USC) on the interpretation of videos taken from Unmanned Air-borne Vehicle (UAV). In 2007, he received the HDR degree (Habilitation a Diriger des Recherches) from Nice University on Scene Understanding. He created the STARS team on the 1st of January 2012. He is research director at INRIA Sophia Antipolis, France. He has conducted research work in video understanding since 1993 at Sophia-Antipolis. He is author or co-author of more than 140 scientific papers published in international journals and conferences in video understanding. He is a handling editor for MVA and a reviewer for several international journals (CVIU, IJPRAI, IJHCS, PAMI,AIJ, Eurasip, JASP) and conferences (CVPR, ICCV, AVSS, VS, ICVS). He has (co-)supervised 13 PhD theses. He is an EC INFSO and French ANR Expert for reviewing projects.

Xilin Chen is a professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 300papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is currently an associate editor of the IEEE Transactions on Multimedia, and a Senior Editor of the Journal of Visual Communication and Image Representation, a leading editor of the Journal of Computer Science and Technology, and an associate editor-in-chief of the Chinese Journal of Computers, and Chinese Journal of Pattern Recognition and Artificial Intelligence. He served as an Organizing Committee member for many conferences, including general co-chair of FG'13 / FG'18, program co-chair of ICMI 2010. He has been an area chair of CVPR 2017 / 2019 / 2020, and ICCV 2019. He is a fellow of the ACM, IEEE, IAPR, and CCF.