Characterizing the State of Apathy with Facial Expression and Motion Analysis

S L Happy, Antitza Dantcheva, Abhijit Das, Radia Zeghari, Philippe Robert, Francois F Bremond

Characterizing the State of Apathy with Facial Expression and Motion Analysis

S L Happy[1], Antitza Dantcheva[1], Abhijit Das[1], Radia Zeghari[2], Philippe Robert[2], and Francois Bremond[1]

[1] INRIA Sophia Antipolis, France

[2] CoBTeK, Memory center CHU, University Cote d'Azur, association IA, France

***Abstract*— Reduced emotional response, lack of motivation, and limited social interaction comprise the major symptoms of apathy. Current methods for *apathy diagnosis* require the patient's presence in a clinic, and time consuming clinical interviews and questionnaires involving medical personnel, which are costly and logistically inconvenient for patients and clinical staff, hindering among other large scale diagnostics. In this paper we introduce a novel machine learning framework to classify apathetic and non-apathetic patients based on analysis of facial dynamics, entailing both *emotion and facial movement*. Our approach caters to the challenging setting of current apathy assessment interviews, which include short video clips with wide face pose variations, very low-intensity expressions, and insignificant inter-class variations. We test our algorithm on a dataset consisting of 90 video sequences acquired from 45 subjects and obtained an accuracy of 84% in apathy classification. Based on extensive experiments, we show that the fusion of emotion and facial local motion produces the best feature set for apathy classification. In addition, we train regression models to predict the clinical scores related to the *mental state examination (MMSE)* and the *neuropsychiatric apathy inventory (NPI)* using the motion and emotion features. Our results suggest that the performance can be further improved by appending the predicted clinical scores to the video-based feature representation.**

I. Introduction

Apathy is defined as the quantitative reduction of goal-directed activity either in behavioral, cognitive, emotional or social dimensions in comparison to the patients previous level of functioning in these areas [1]. Apathy represents a pervasive neuropsychiatric symptom of the majority of neurocognitive, neurodegenerative, and psychiatric disorders including Alzheimer's disease (AD) [2], Parkinson's disease [3], and mild cognitive impairment [4]. It has been estimated that nearly 65% of dementia patients experience apathy [5]. While experts in the field suggest that the early indication of apathy could improve the intervention effects and decrease the global burden of the disease [6], apathy has been highly underdiagnosed. This is the case due to the only recently established criteria, as well as due to the high level of subjectivity in evaluation of human observers. Specifically, apathy diagnosis is based on interviews with patients and their caregivers, where the clinicians seek to assess among others the loss of, or diminished emotion in patients. Towards assisting such subjective assessment, an objective and automated analysis carries the promise to *enable early apathy diagnostics, leading to improved intervention effects, potentially increasing the performance of apathy detection in a non-invasive and efficient manner, relieving national health-care systems from excessive workload and allowing for large scale early and remote diagnostics*.

This analysis has been carried out in the context of the MNC3 program (Medecine Numrique Cerveau Cognition Comportement) from the University Cote d'Azur Idex.

Motivated by the above, in this work we introduce an automated system for classification of patients with and without apathy assisting clinicians in the diagnostics of apathy based on facial behavior analysis. Three dimensions of apathy were identified in a recent medical paper by Robert et al. [1]: (a) behaviour / cognition, (b) emotion, and (c) social interaction. We here aim at recognizing (b) the *emotional dimension of apathy*, characterized by exhibiting 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 gleaned from *facial movements*. Patients suffering from 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 were elicited by asking all subjects to briefly narrate past positive and negative experiences. The clinical diagnosis of the subjects was carried out by the psychiatrists along with the recording of facial videos during positive and negative narration. We explore the video data for apathy classification using the facial motion and emotion information.

In a nutshell, our proposed approach analyzes the facial expressions and face movement patterns in elderly subjects to infer the apathy state. The video-level representation of both motion and emotion features are extracted and utilized to estimate the apathy related clinical scores using regression models. Further, we test appending the prediction scores of clinical attributes (such as mini mental state examination (MMSE) [7] and neuropsychiatric apathy inventory (NPI-apathy) [8]) to the motion and emotion feature vectors to improve the class discrimination. With extensive experiments, we show that the fusion of the features extracted from two type of videos (video of positive and negative event narration by the subjects) obtains the best performance.

To the best of our knowledge, we are among the first to propose an approach for apathy detection based on facial behavioral analysis. Very recently Chung et al. [9] introduced

an approach for apathy diagnosis based on visual scanning behavior, i.e., the visual attention of patients given emotional and non-emotional visual stimuli. Deviating from their work, we here present a framework for apathy diagnosis based on facial motion and expression analysis from videos.

The contributions of the paper are listed below.

- This paper is the first to investigate facial behavior in an automated manner towards classifying apathetic and non-apathetic patients.
- Specifically, we investigate the variation of facial expressions and facial movements as cues for apathy detection. We report the performance of various feature combinations with extensive experiments.
- We further show that the feature representation can be improved by appending the estimated clinical scores to it. We use regression models to estimate the clinical scores (such as MMSE and NPI-apathy) from facial feature representation, and append the estimation scores to the feature vector to improve the classification performance.

This work is organized as follows: Section II revisits existing work on apathy diagnostics, as well as facial and head motion analysis. Section III describes our proposed method. We firstly introduce the dataset in Section III-A), where we elaborate on the video data, as well as the clinical scores, we utilize to train the machine learning models for apathy classification. Then, the motion and emotion feature extraction methods are discussed in Section III-B. The regression and classification models are explained next, followed by an overall representation of the proposed methodology. We present the experiments in Section IV and the related results in Section IV-C and finally conclude in Section V.

II. Related Work

Works on apathy diagnosis: Currently, clinical interviews or questionnaires are the only methods of a reliable apathy diagnosis [3], which are known to be time consuming and require the presence of patients at a hospital. Though several biomarkers for apathy are discussed in Hampel et al. [6], automated apathy diagnosis is a novel research area of high impact and hence interest. The computer vision based analysis of face and gesture has shown to provide abundant information about different neurodegenerative disorders [10], [11], [12], [13], which we here aim at exploiting for apathy diagnosis.

The only work on computer vision based apathy diagnosis is by Chung et al. [9] and is based on visual scanning behavior analysis. This paper analyzes the sequences of fixations and saccades within and between regions of interest on visual stimuli. In particular, recurrent neural networks (RNN) are employed to learn group difference and individual difference in visual scanning process towards the emotional and non-emotional stimuli. Some literature reports the use of the neuroimaging modalities for apathy diagnosis [2], [14]. The correlation of apathy to AD is studied in Aguera-Ortiz et al. [2] using magnetic resonance image analysis. Structural and functional alteration of frontal-subcortical networks is observed among AD patients suffering from apathy and has been analyzed based on single-photon emission computed tomography, positron emission tomography, and diffusion tensor imaging, which are reviewed and discussed in detail in [14].

Facial expression recognition and its applications in disease diagnosis: The emotional health, i.e. the ability to express emotions and identify others emotions, plays a major role in cognitive behavioral therapy [10], [15], [16]. Facial expression recognition, as an indicator of internal emotional state, has been widely explored in the last two decades [17], [18], [19]. Montenegro et al. [11] analyzed emotion recognition based on 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 from facial expression analysis [20]. Montenegro et al. [21] studied emotion recognition from facial depth image analysis to infer the cognitive and emotional behavior. A computational approach for diagnosis and assessment of autism spectrum disorders was proposed by Coco et al. [22] with facial analysis. Similarly, Samad et al. [23] investigated automatic detection of the autism spectrum disorder using spontaneous expression analysis and concluded that uncontrolled manifestation of smile without proper visual engagement could be an appropriate sign of impairment in social communication. Reduced facial expressions or hypomimia was found to be a major cue for estimating the stage severity of Parkinson's disease [24]. The facial expression features (facial appearance and dynamics) were used to estimate the clinical depression scores [12].

Head and face movements: According to Hammal and Cohn [25], head motion also plays an important role in emotion communication. Following this, both facial expression and head motion were studied by Adams and Robinson [26] for classifying complex categorical emotions. Hammal et al. [27] explored the dynamics of head and face movements for their connections with positive and negative affects. Facial movements from 49 facial landmarks and head movements in terms of pitch, toll, and yaw were analyzed and found to have strong correlation with the type of affect. Head movement, lip deformation, and eyebrow movements were categorized as major facial cues for anxiety detection in [28]. The results indicated that the use of specific head and face movements due to activities of eyes and mouth are discriminative indicators of anxiety and stress. The face and head movements along with the speech features were used by Dibeklioğlu et al. [29] to detect the severity of depression by encoding behavior patterns. Anis et al. [13] computed head movements by tracking facial landmarks. The histograms of velocity and acceleration intensities were used as features to estimate the three levels of chronic depression severity.

III. Proposed method

Deviating from the above we here propose a novel approach for apathy detection in two short video sequences per subject, comprising of subjects describing positive and negative episodes of their life. In what follows, we will

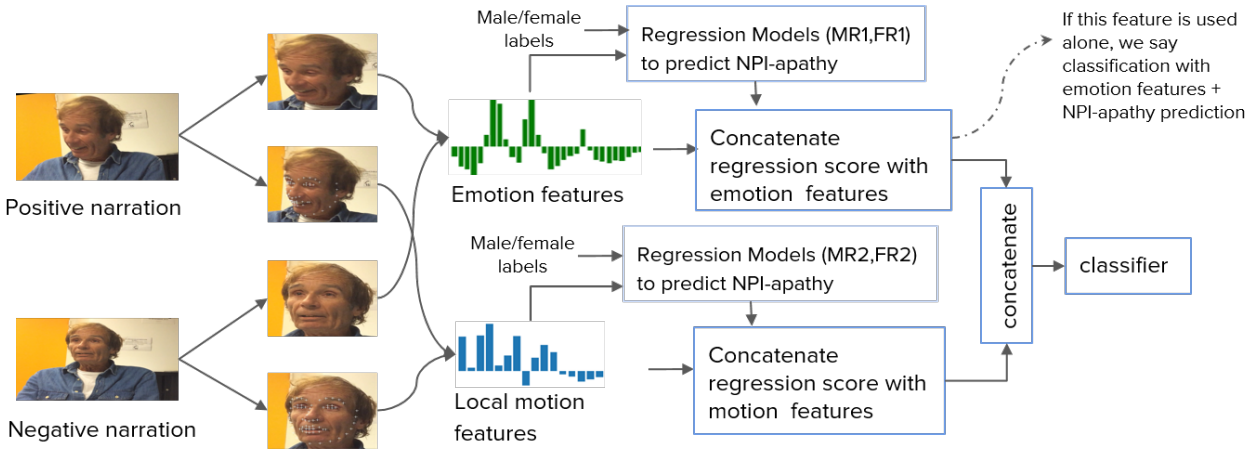


Fig. 1: Overall framework for apathy detection from facial videos.

describe the basic steps of our proposed method, illustrated in Figure 1.

A. 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 I. Among the 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 the mini mental state examination (MMSE) [7] and neuropsychiatric apathy inventory (NPI-apathy) [8]. MMSE has been used for the cognitive assessment and apathy quantitative assessment used the NPI-apathy scores. To elicit spontaneous facial expressions, in (iii) 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). The video data was recorded with a tablet controlled by the clinician. Though most of the videos have near-frontal face, there exists a lot of pose variation and facial occlusions in the dataset. Moreover, the expressions were very subtle and the average video length was around one minute.

TABLE I: 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)

B. Feature Extraction

While performing video level classification, researchers usually leverage the temporal dynamics [30] using long short-term memory (LSTM) or recurrent neural network (RNN). However, literature, which leverages the temporal patterns, predominantly focuses on categories like human activities (walking, bowling, typing, etc.) or visual content (parade, wedding dance, bird, birthday, etc.). These categories are quite distinguishable from human perspective. However, the problem at hand (apathy classification from facial videos) is more challenging, as opposed to the categories discussed above. We here note, that even psychology experts are challenged in predicting the apathy state, by analyzing merely the face of a subject. Since there is no particular temporal pattern associated with the collected video data, we approach the problem with a bag of visual words model [31], in which features extracted from each frame are further pooled for a codebook based representation of the whole video. Though this model lacks temporal relation, it is advantageous for us as the present data possess no particular temporal pattern.

1) Emotion Features: As facial expressions are related to internal emotions, facial expression recognition has been widely researched for emotion analysis [18]. While expression recognition is predominantly based on the six-expression model [18], as agreed with involved clinicians, we here use three categories of expressions, namely: *positive, negative, and neutral*. This choice stems from the highly limited expressions expressed by the participants in the relative short video sequences under clinical conditions. Thus, we trained a convolutional neural network (CNN) model for *expression classification* with these *three categories*.

The use of pretrained VGG-Face [32] is a prominent architecture choice among recent works on face analysis. Since VGG-Face is trained with 2.6 million faces, it has been reported as a robust facial feature extractor, achieving promising results in facial expression recognition [33],[34],[35],[36]. In such works, the last few layers of the VGG-Face are typically fine-tuned for respective applications. In our experiments we found that using a set of skipped

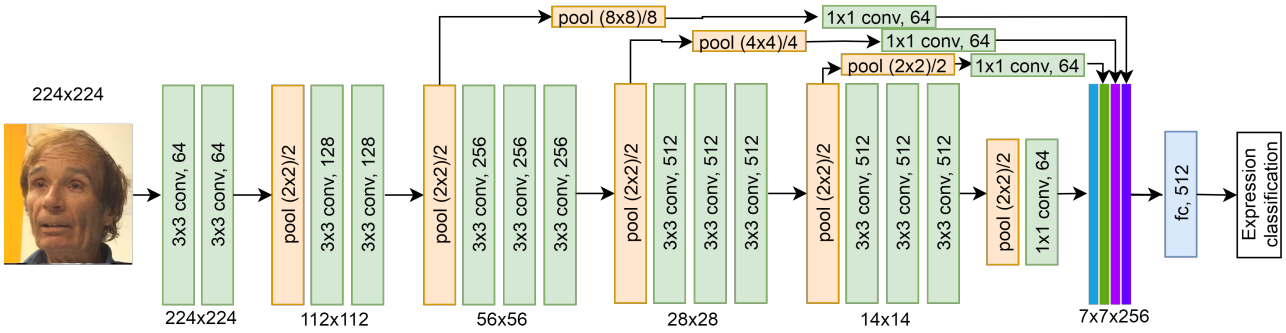


Fig. 2: Facial expression recognition framework. (conv: convolutional layer, fc: fully convolutional layer, pool: pooling layer)

connection in VGG-Face architecture further improves the expression classification accuracy. The experimental validation of the proposed emotion recognition model is beyond the scope of this paper. However, we here describe the details of the CNN architecture used in our experiments, illustrated in Figure 2. As can be seen, the skip connections are made from the pooling layers and further pooling is applied to reduce the spatial resolution to 7×7, after which they are concatenated channel-wise. The 1×1 convolutional layers are used to reduce the channel length, thus accumulating the necessary features across channels. In our experiments, the first five convolutional blocks use the pretrained weights of VGG-Face, while the skip connection layer weights and the last fully connected layers are trained with the publicly available datasets.

Publicly available expression datasets generally contain the universal classes ('anger', 'disgust', 'fear', 'happy', 'sad', and 'surprise') and we here directly use 'happy' and 'neutral' samples from the dataset during training, while grouping the 'sad', 'anger' and 'disgust' samples into the negative class.

Symptoms of apathy include reduced emotional responses. Hence, we hypothesize that the *expression intensity distribution* and the *duration of each expression* throughout the video can be utilized to infer the apathy state. Therefore, we concatenated these two types of features for both positive and negative narration videos, and call them "emotional features". We proceed to describe these two features, as well as the associated extraction methods.

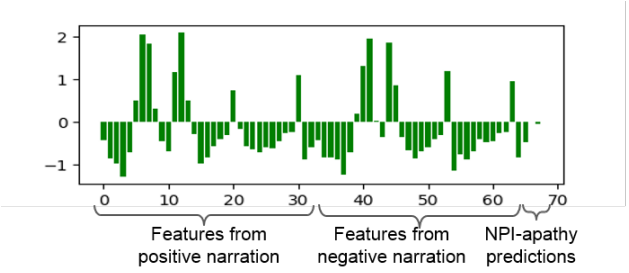
Expression intensity representation: We assume that the log probabilities of the softmax layer represent the emotion intensities corresponding to each category and pooled the frame-wise expression intensities into a histogram vector. Thus, we obtained a histogram vector (b bins in each histogram) for each expression, which are further combined together ($3 \times b$ dimensional feature vector for 3 classes) as a representation of expression intensities for the whole video. Similar to the bag of words analysis, here the histogram features represent the probable occurrence of an expression with certain intensity. As per our hypothesis, apathetic subjects will show less expressions with subtle intensities, thereby having higher bin counts in the first few bins. Therefore, the expression histograms of apathetic persons would hypothetically look like a right skewed distribution, whereas it would be left-skewed for non-apathetic persons.

Expression duration: The duration of dominant expressions in a video is an important cue for accessing the overall emotional display. If e-th expression is dominant for n_e number of frames out of total N number of video frames, then we used $t_e = \frac{n_e}{N}$ as the expression duration of e-th expression. The expression durations $(t_{pos}, t_{neg}, t_{neut})$ are appended to the expression representation, resulting in a $3 \times (b+1)$ dimensional feature vector. As shown in Figure 3a, the features of positive narrations include both expression intensity representation and the expression duration.

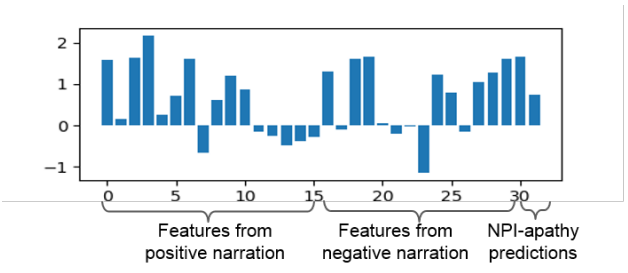
Combining features of positive and negative narration: We here seek to answer the question of whether positive or negative expressions are more likely to be expressed by apathy patients. From psychological point of view, apathetic persons are indifferent toward any emotions and hence expressions. However, the healthy subjects exhibit limited expressions in a clinic environment (such as ours) as well, which make the classification task difficult. Hence, the presence of both positive and negative narrations per subject is pertinent, and combining the features extracted from both videos-sequences provides a broader spectrum of facial expressions instrumental for apathy classification.

Training regression models: We also investigated the utilization of clinical scores (such as MMSE and NPI-apathy) in our experiments. These scores are estimated by the clinical experts through interviews and questionnaires. Instead of directly using the scores provided by the clinicians, we estimated these scores using the motion and emotion features. We use the support vector regression (SVR) with radial basis kernels [37] for estimating the clinical scores (MMSE an NPI-apathy). Although the estimation of these scores from emotion features is not very accurate, these predicted scores are appended to the feature vector extracted from videos to further improve the classification performance,

Motivated by depression analysis, where current literature [38] reported the difference between the acoustic features pertinent to depression in male and female speech and results suggested that audio-based depression recognition benefited from gender-dependent feature extraction, we here investigate apathy score prediction with two separate gender models. Considering NPI-apathy prediction from emotion features, we first train two gender dependent regression models. In order to make the system independent of gender, we use the NPI-apathy predictions of both the regression



(a) Normalized emotion features



(b) Normalized facial motion features

Fig. 3: Feature extraction: The (a) emotion features and (b) motion features are extracted for each subject by concatenating the features from videos of positive and negative narrations, followed by the concatenation of the predicted clinical scores. Note that the fusion of both emotion and motion features is utilized for final apathy classification.

models irrespective of the gender of the participant. In other words, the regression models MR1 and FR1 are trained with emotion features from male and female subjects separately; however, the prediction scores of both MR1 and FR1 for all participants are considered as features during classification. Thus, the gender information is not required during testing phase. The addition of NPI-apathy prediction from both the models to the emotion features is shown Figure 3a, thereby making a feature vector of $3 \times (b+1) + 2$ dimensions.

2) Motion Features: From the psychological point of view, facial motion can also be a prominent indicator of apathy. Since apathy is characterized by limited verbal or nonverbal interaction along with lack of interest in surrounding environment, we investigate head and facial movements for apathy detection. Inspired by Hammal et al. [27], the dynamics of head and facial landmarks were extracted. We estimate the rigid head movements by tracking the facial points, which do not change with non-rigid facial deformations. We use the nose and inner eye corners as the landmarks for computing rigid head movement, denoting it as global head motion. Similarly, the non-rigid facial landmark movements are associated with lips, eyes, eyebrows, and chin, while having a conversation or showing an expression. Specifically, the average movement of facial landmarks around these regions in successive frames are computed as the local motion feature. The facial landmarks used for computing local and global features are shown in Figure 4.

Video sequences in our dataset are not time-limited and hence entail different video length, we compute a video-based representation for the further analysis. We obtain the statistical features (such as, minimum, maximum, mean, median, standard deviation, skewness, and kurtosis) as the motion representation using the motion information (separately for global and local motion). In addition, we append the b-bin histograms of motion values to preserve the motion intensity distribution information in motion representation, thereby making a vector of $b + 7$ dimensions. Further, we concatenate the motion features of positive and negative narration.

Similar to emotion features (see Section III-B.1), we estimate the NPI-apathy and MMSE scores from motion features with two separate models for male (MR2) and female (FR2), which are appended to the motion representation for performance improvement.

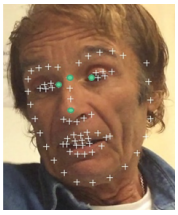


Fig. 4: Motion features obtained by averaging the movements in subsequent frames. The '+' and green 'o' landmarks are used for computation of local and global motion features respectively.

C. Overall Framework

Based on the above considerations, the block diagram of the proposed method is shown in Figure 5. A simplified description of the methodology is provided in this Section in terms of the operations in the training and test phase.

1) Training Phase: The proposed framework takes five inputs: (1) video of positive narration, (2) video of negative narration, (3) apathy / control label, (4) male / female label, (5) NPI-apathy score or MMSE. The emotion and motion features are extracted from both, (1) and (2). The clinical scores are estimated using regression models trained for male and female separately, which are further concatenated to individual features (as shown in Figure 3). After obtaining the emotion and motion representations, the representation level fusion [31] is carried out for further apathy / non-apathy classification.

2) Test Phase: During the test phase, the system relies only on the videos of positive and negative narration, (1) and (2). Using the emotion and motion features computed from both these videos, the framework is able to predict the apathy state of the patient.

IV. Experiments and Discussion

A. Preprocessing

Prior to the main framework, we detect faces from the dataset-videos using MTCNN [39], followed by face alignment by positioning both eyes at a fixed distance parallel to the horizontal axis. The aligned faces are re-sized to $224 \times$

TABLE II: Performance of different feature extraction techniques with and without the fusion of features from positive and negative narration. The fusion of features from both videos improves the performance.

Features used	Without fusion of positive and negative narration			After fusion of positive and negative narration		
	Accuracy	F1-score	AUC	Accuracy	F1-score	AUC
Local Motion Features	58.88	0.505	0.527	57.77	0.555	0.558
Global Motion Features	56.66	0.516	0.523	55.55	0.537	0.537
Local + Global Motion Features	51.11	0.458	0.467	62.22	0.602	0.601
Emotion features	52.68	0.532	0.518	64.44	0.622	0.62
Emotion features + Local Motion Features	53.86	0.526	0.552	77.77	0.757	0.75
Emotion features + Local + Global Motion Features	54.57	0.532	0.537	71.11	0.68	0.676

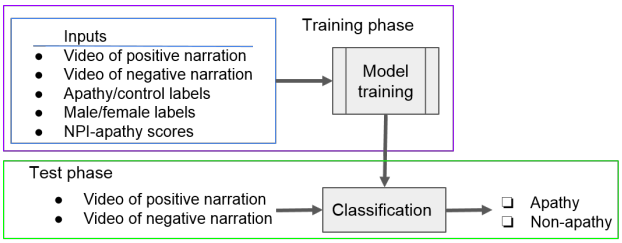


Fig. 5: The simplified block diagram of proposed method for apathy detection.

TABLE III: Performance of different feature extraction techniques when features from positive and negative narration are fused. The use of clinical score predictions as features improves classification accuracy.

Features used	Accuracy	F1-score	AUC
Local Motion Features + NPI-apathy prediction	73.33	0.722	0.722
Global Motion Features + NPI-apathy prediction	77.77	0.765	0.768
Local + Global Motion Features + NPI-apathy prediction	68.88	0.676	0.678
Emotion features + MMSE prediction	77.77	0.757	0.75
Emotion features + NPI-apathy prediction	66.66	0.649	0.648
Emotion features + Local Motion Features + MMSE prediction	68.88	0.675	0.675
Emotion features + Local Motion Features + NPI-apathy prediction	**84.44**	**0.836**	**0.833**
Emotion features + Local and Global Motion Features + MMSE prediction	68.88	0.66	0.657
Emotion features + Local and Global Motion Features + NPI-apathy prediction	77.77	0.757	0.75

224 resolution, constituting the input for the CNN model. The CNN model is trained to classify the face into three expression classes, namely: positive, negative and neutral. We use in-the-wild (AffectNet[40]) datasets 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 deep model. The facial landmarks are detected using DLIB [41] in order to compute the motion features. In all our experiments, we consider the histograms with 10 bins ($b = 10$) for both motion and emotion feature extraction. The extracted features are further normalized to zero mean and unit variance before feeding into the classifier.

B. Evaluation Strategy

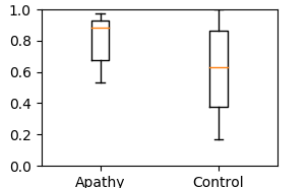
We evaluate the efficiency of the extracted features for distinguishing apathy patients from the control group. Support Vector Machines (SVM) with linear kernel are used as classifier in all experiments to demonstrate the efficiency of the proposed feature extraction method with a simple classifier. Since there are a limited number of samples in the dataset, we evaluate the performance using leave-one-subject-out (LOSO) strategy [29], in which samples from one patient constitute the testing set, while the remaining data is used to train the model. All the regression and classification models in our experiments are trained with LOSO. We report the macro average of accuracy, F1-score, and area under the curve (AUC) as the performance metrics.

C. Experimental Results

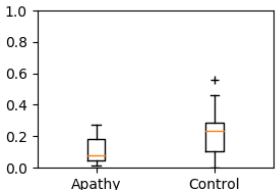
We here proceed to describe the extensive experimental study of various feature combinations and their related performances. Firstly, we present a performance comparison of various feature combinations with and without fusing the features of positive and negative narrations in Table II. Here the performance without fusion is obtained by using features from individual videos for classification. Thus, two videos per subject (positive and negative narration videos) are used for testing under LOSO cross-validation, while using the remaining set for training. Note that the clinical score predictions were not included in feature vectors for these experiments. As can be seen in Table II, all the performance metrics improved in most cases when the features are fused from both narrations. We also observe the improvement in F1-score and AUC even with motion features.

As per Table II, the framework achieves an accuracy of 64% when the emotion features are used alone, while achieving an accuracy of 62% when utilizing motion features alone. The best performance (accuracy = 77.77% and F1-score = 0.757) is achieved when emotion features are combined with local motion features. This proves the complementary information present in motion and emotion features. However, the performance is reduced when both global and local motions features are taken into account. This might be due to the redundant information that are encoded into local and global motion features. Another pertinent observation is that the use of emotion features alone or with combination of motion features always outperforms the motion features.

The performance improvement by including the clinical score predictions in feature vector is reported in Table III. Note that all results reported in Table III are obtained by combing the features from both positive and negative narration. We obtained the best performance (accuracy = 84.44% and F1-score = 0.83), when the emotion features are concatenated with local motion features and NPI-apathy prediction scores. However, we show that the performance

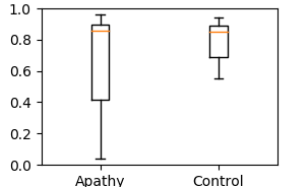


(a) Neutral or Positive emotion duration

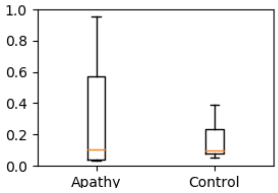


(b) Negative emotion duration

Fig. 6: The duration of positive/neutral and negative expressions during the negative narration of female subjects.



(a) Neutral or Positive emotion duration



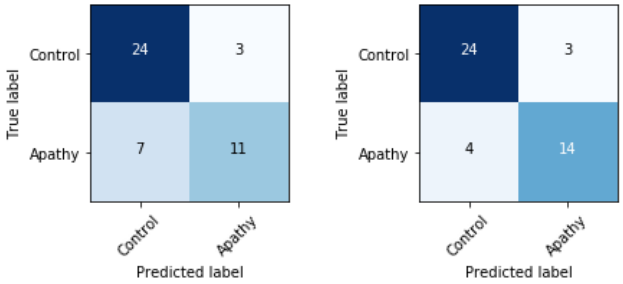
(b) Negative emotion duration

Fig. 7: The duration of positive/neutral and negative expressions during the positive narration of male subjects.

of global motions with NPI-apathy predictions and emotion features with MMSE predictions are comparable. As opposed as the results in Table II, we observe a huge performance gain, when the clinical score estimations are used as features. For instance, the accuracy of local features improved from 57.77% to 73.33%, when combined with NPI-apathy prediction. Similarly, the F1-score of emotion features improves from 0.602 to 0.75, when the MMSE predictions are utilized. This signifies that the prediction of clinical scores has an important role in performance improvement. Overall, the combination of features from positive and negative narrations and the use of clinical score estimations was found to improve the model performance significantly.

As discussed in Section III-B.1, we use emotion duration as one feature. Observations from the duration of different emotions of female and male subjects during negative and positive narrations are provided in Figure 6 and Figure 7 respectively. As can be observed from Figure 7, during positive narration, male subjects with apathy show both positive and negative emotions, whereas the control group showed more positive expressions compared to negative expressions. During negative narration (Figure 6), female subjects with apathy showed less negative emotion compared to the control group. However, it is clear that these observations alone are not enough for apathy state classification because of significant overlapping of features between control and apathy patients.

The confusion matrix of two feature extraction methods is reported in Figure 8. As can be observed, the use of NPI-apathy prediction scores improves the recall score of apathy detection from 0.61 to 0.77. The receiver operating characteristic curve (ROC) of different feature extraction methods is provided in Figure 9.



(a) (b)

Fig. 8: The confusion matrices for (a) classification with motion and emotion features, (b) classification with motion, emotion features, and NPI-apathy prediction scores.

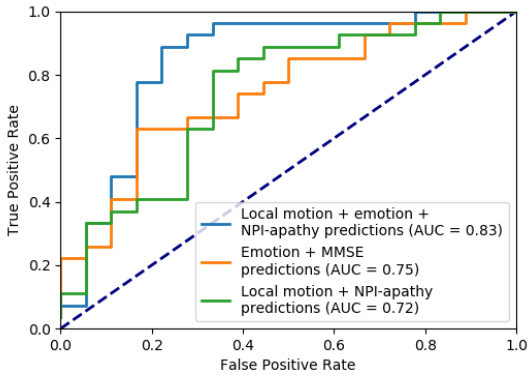


Fig. 9: ROC plot for selected feature extraction methods.

V. Conclusions

We presented a new automatic apathy detection method, which analyzes facial emotion and motion. Machine learning models are first trained with emotion and motion features along with clinical scores, as well as considering the subject's gender. During the testing phase, only the facial videos are fed to the trained model, in order to predict apathy. While *emotion* based apathy detection achieved an accuracy of 64%, jointly emotion and motion features boosted the performance to 77%. Our framework benefits from positive, as well as negative expressions. We obtained best results by fusing emotion and local motion features in combination with the NPI-apathy score estimations.

In future, we intend to explore the correlation of features from different facial regions for apathy detection. For example, the use of lips and eyebrow movement separately might improve the performance. The use of spatio-temporal features with deep models for apathy prediction is another unexplored field. Fusing information from multi-modal data including video, speech, thermal, galvanic skin response, as well as neuroimaging modalities could be an interesting future direction in this context.

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] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "mini-mental state: a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[8] J. L. Cummings, M. Mega, K. Gray, S. Rosenberg-Thompson, D. A. Carusi, and J. Gornbein, "The neuropsychiatric inventory: comprehensive assessment of psychopathology in dementia," *Neurology*, vol. 44, no. 12, pp. 2308–2308, 1994.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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 *European Conference on Computer Vision*. Springer, 2018, pp. 144–157.

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

[17] S. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Transactions on Affective Computing*, 2017.

[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] S. Happy and A. Routray, "Robust facial expression classification using shape and appearance features," in *International Conference on Advances in Pattern Recognition*. IEEE, 2015, pp. 1–5.

[20] 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.

[21] 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.

[22] M. Del Coco, M. Leo, P. Carcagn, 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis, "Stress and anxiety detection using facial cues from videos," *Biomedical Signal Processing and Control*, vol. 31, pp. 89–101, 2017.

[29] 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.

[30] Y.-G. Jiang, Z. Wu, J. Tang, Z. Li, X. Xue, and S.-F. Chang, "Modeling multimodal clues in a hybrid deep learning framework for video classification," *IEEE Transactions on Multimedia*, 2018.

[31] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

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

[33] 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 12th IEEE International Conference on*. IEEE, 2017, pp. 118–126.

[34] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 367–374.

[35] S. Albanie and A. Vedaldi, "Learning grimaces by watching TV," in *British Machine Vision Conference (BMVC)*, 2016.

[36] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, and Z. Luo, "Conditional convolution neural network enhanced random forest for facial expression recognition," *Pattern Recognition*, vol. 84, pp. 251–261, 2018.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant," in *Proc. of Conference on Artificial Intelligence in Medicine. Springer*, vol. 5, 2017.

[39] 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.

[40] 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.

[41] 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.