Audio-Visual Weakly Supervised Approach for Apathy Detection in the Elderly

Garima Sharma Jyoti Joshi
Human-Centered Artificial Intelligence Group
Monash University
{garima.sharma1, jyoti.joshi}@monash.edu

Radia Zeghari
CoBTeK lab
Université Côte d'Azur
radia.zeghari@gmail.com

Rachid Guerchouche
INRIA
rachid.guerchouche@inria.fr

***Abstract*—Apathy is manifested as lack of feelings or emotions in several neurological and psychological disorders. Hence, directly impairing the display of emotion through facial expressions and speech. Current practices of prediction of apathy heavily rely on clinical diagnosis, an expert interviewing a patient or reports from patients' family members. The dependence on an expert and the human bias in its examination results in under-diagnosis of condition. In this paper, a multimodal multi-instance learning based method is proposed for automatic apathy detection. There are several challenges present while automating the process. Some of which are - recognizing emotions in elderly people, correct identification of emotions in a conversation and identifying differences between emotional responses from apathetic and non apathetic cohorts. The proposed method uses the audio and visual information in a weakly supervised manner to learn the apathetic behaviour in order to address these challenges. Features from facial expressions, action units, facial landmarks and audio signals are extracted for training. The fusion of multiple modalities in a weakly supervised method achieves 75.71% accuracy for apathy detection in elderly people. The experiments show that multimodal fusion is able to leverage on the presence of complimentary information across different modalities.**

***Index Terms*—Apathy detection, Emotion recognition, Multiple instance learning, Digital health**

I. Introduction

Emotions are crucial in our everyday life. They are used not only to express our feelings but are also an indicator of our mental well-being. Marin [1] defined apathy as absence or lack of feeling, emotion, or concern. Robert et al. [2] revised the diagnostic criteria of apathy as significant reduction of goal directed activities in behavioural/ cognitive, emotions and social interactions. Apathy is associated with several psychological and affective mood disorders (such as Depression and Schizophrenia), neurological disorders (such as Dementia, Delirium, Amnestic disorders, Huntington's disease, Akinetic mutism), etc [1]. In a study by Simons et al. [3], it is found that people with Parkinson's disease show reduced spontaneous expressions towards an external stimuli as compared to healthy people. In an another study [4], it is found that there is a high chance of occurrence of apathy in various mental disorders such as 73% in Alzheimer's disease, 53% in depression, 32% in right hemispheric stroke, 22% in left hemispheric stroke, and 7% in normal population.

The diagnosis of apathy involves series of interviews, which are conducted by an expert to identify a change in a person's behaviour and loss of interest and activities. The family members or the carer of a patient are generally involved in this procedure. In many cases it is observed that a person is reluctant to visit an expert [5]. Studies also show that apathy is often misdiagnosed or confused with depression [6]. These factors show the relevance of studying apathy and identifying a suitable low cost method for its diagnosis.

Generally, the mental health disorders are recognized by pre-defined self-report scales or by an interview conducted by a clinician. The careful examination of a person's facial expressions, body gesture, gaze movements, etc. are done to identify the symptoms of any disorder. Computer vision based approaches have already achieved good performance in analyzing human's facial features and expressions. Several studies proposed an efficient vision based solutions targeting various aspects of different psychological disorders [7]–[10]. For the detection of Parkinson's disease, researchers are now using vision based methods for the pose detection of patients [11], [8]. The common characteristics observed in patients with dementia and depression such as prevalent low-valence emotion can be identified by analyzing a patient's facial expressions, speech and head movements [9] [12], [10].

According to literature in psychology, apathy can be identified by analyzing the emotions [13]. In this paper, an automatic multimodal system is proposed to detect apathy. The proposed method contains four parts to effectively analyze the emotion exploiting the facial expressions, facial action units, speech features and local facial motion information. The proposed method is validated on the videos of elderly people, narrating a positive and negative emotional event to a clinician.

There could be several applications of the proposed system. Besides in clinical settings, an automatic apathy detection system can be deployed in facilities such as aged-care, to analyse the behaviour of residents and provide an early level warning in case apathetic behaviour is predicted. Further, the result can be forwarded to an expert for the validation.

The main contributions of the paper are as follows:

- A non-invasive automatic audio-visual modality based model for apathy detection in elderly.
- A multiple instance learning based method for encoding distinct subtle variations in emotion.

The rest of the paper is organized as follows: Section II discusses the relevant studies in the direction of apathy

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

detection, emotion detection and multiple instance learning. Section IV presents the architecture of the proposed method. Section V explains the different experiments performed and the associated results. Conclusion and the future directions are discussed in Section VI.

II. Related Work

This section describes prior work in the area of apathy detection. Commonly used methods using appearance information to detect apathy are discussed. The concept of multiple instance learning is discussed along with the recent studies using it.

A. Apathy Detection

Automatic affective computing systems have been noted to be useful in detecting several mental health disorders [14]. Some systems use the appearance based information somewhat similar to a clinician observing the behaviour of a person in an interview visually [15]. Several studies exploit the appearance and audio information to recognize the indicators of disorders such as Dementia, Parkinson's disease etc. Parekh et al. [7] proposed an automatic visual system to measure engagement level in dementia. Their system analysis gaze, emotion and behaviour, while a person with dementia uses an application in a tablet.

López-de-Ipiña et al. [16], proposed an approach for early level diagnosis of Alzheimer's disease by using spontaneous speech. The study showed that use of such non invasive methods for early diagnosis of Alzheimer's disease can be deployed for a low cost method. Osborne-Crowley et al. [15] studied the facial expressions in apathetic patients at early stage of Huntington's disease. Their analysis showed that there is a little difference in the cognitive functioning of apathetic and non-apathetic people. However, people with apathy were found to be impaired in recognizing happy expression. Along with facial expressions, speech of a person can also be used to detect the state of apathy. Konig et al. [17] used prosodic, formant, source and temporal features from the speech to characterize and detect the apathy. The paper showed that speech is a reliable source to predict the apathetic state.

In a recent study, Happy et al. [18] identified the state of apathy by using facial expressions and the local facial motion on a similar data. Their method uses the positive and negative narration video as a set to extract emotion and local motion features. A different regressor is learned for each type of features which is concatenated for final classification. The authors also used the clinical scores such as Neuro Psychiatric apathy (NPI-Apathy) inventory [19] and Mini Metal State Examination (MMSE) [20] score as a feature to train the samples for apathy prediction. In present work, same data as of Happy et al. [18] is used although the number of samples is larger in our case. The proposed approach also differs from [18] as follows -

1) Positive and negative narration stimuli response videos are trained separately (without having any dependence) such that the prediction can be made in the absence of one.
2) The training is done without using any clinical scores.
3) The audio level information is fused for better representation of emotion.
4) The proposed model is independent of any extra meta-data information such as male/female labels, positive/negative narrated interviews labels. Only state of apathy label is used i.e. apathetic or non-apathetic.

The model is carefully designed by training only on appearance and audio level information. This is to ensure the scalable utilisation of the proposed model in the real-world settings to flag an early level warning for apathetic behaviour.

B. Emotion Detection

The emotional state of a person is widely used in various vision based cognitive applications [21]–[23]. A number of studies have already proposed different methods to identify emotion based on detected facial expressions from images [24], [25]. Barrett et al. [26] argued that facial expressions alone can't estimate the emotion of a person. Hence, the use of other modalities is essential. Audio and facial features are combined to identify different emotions in several studies [27]. However, these methods are found to have a bias on the training data. It has also been found that it is difficult to recognize facial expressions of an old person as compared to a young person due to the wrinkles and folds in the face [28]. Thus the proposed approach is motivated to investigate a bimodal approach to analyse emotions in elderly.

C. Multiple Instance Learning

Multiple Instance Learning (MIL) was introduced by Dietterich et al. [29] for drug activity prediction. In MIL, the training data is divided into multiple sets and one label is used for the complete set without having the labels for each data point [30]. The technique allows to learn on weakly labeled data. MIL has already achieved success in various applications. Andrews et al. [31] proposed two SVM based MIL methods for classification. The methods are named mi-SVM (for instance-level classification) and MI-SVM (bag-level classification). There are several studies [32], [33] which use neural networks to explore this problem. Most of the computer vision tasks such as face detection [34], segmentation [35], etc. can fit into multiple instance learning framework.

In a recent study, Xu et al. [36] proposed a weakly supervised deep learning based MIL method in medical image processing. Further, Zhu et al. [37] proposed a MIL based method with salient windows focused on unsupervised object detection. Wu et al. [38] used both CNN and DNN based MIL methods for image classification as well as image auto-annotation task. Zhu et al. [39] proposed a deep multi-instance framework (sparse label assignment) for the breast cancer classification task. In a recent study, Ilse et al. [40] proposed an attention based MIL framework for learning Bernoulli's distribution (at bag level).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

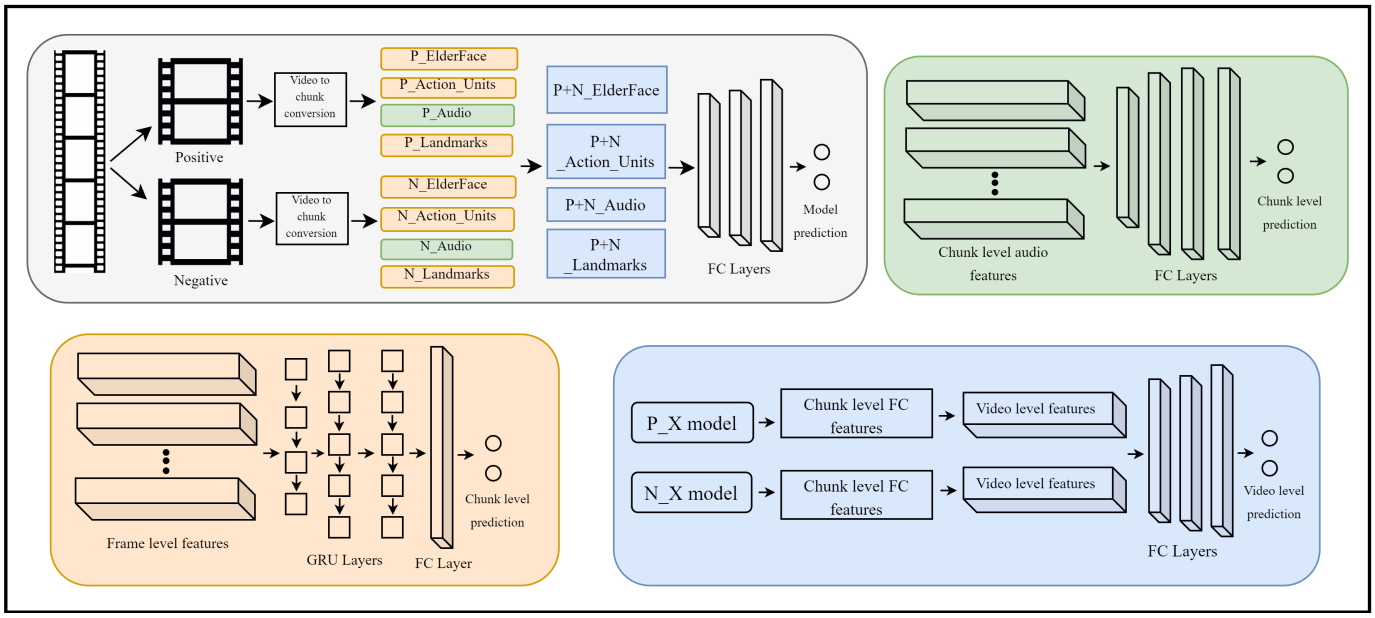


Fig. 1: The proposed network architecture for apathy prediction. The upper left network shows the overall network. The upper right network (in green) shows the training of audio features. Lower left network (in orange) is trained for ElderFace, AU and motion features. Finally lower right network (in blue) is used for the fusion of multiple features and to obtain the video level prediction. Here, P and N denotes positive and negative videos, respectively. (Best viewed in color)

III. Dataset

The data used in this study is collected in Nice Memory Research Center in Nice University Hospital. Videos are recorded in an interview where a person is narrating a positive and a negative experience to a clinician. People with age 65 or above participated in this experiment. The data is recorded without imposing any constraint to the participants. Thus, the videos recorded have a variation in head pose. The videos are then labeled to be apathetic or non-apathetic by an expert after observing the person. The dataset contains videos of 70 participants among which 28 are apathetic and 42 are non apathetic. For every participant two videos are recorded, one for positive and another for negative narration. Hence, total 140 videos are used in this study from 70 participants. This clinical dataset is referred as *apathy dataset* in rest of the paper.

Emotion recognition for elder people is a challenging task as compared to the younger ones. To make the trained model able to learn face level information of elder people, Faces dataset [41] is used in this study. The dataset contains the facial images of young, middle aged and older people with six basic expressions. The images available in this dataset are used first in the model architecture to pre-train and thus fine tune it on apathy data.

IV. Model Architecture

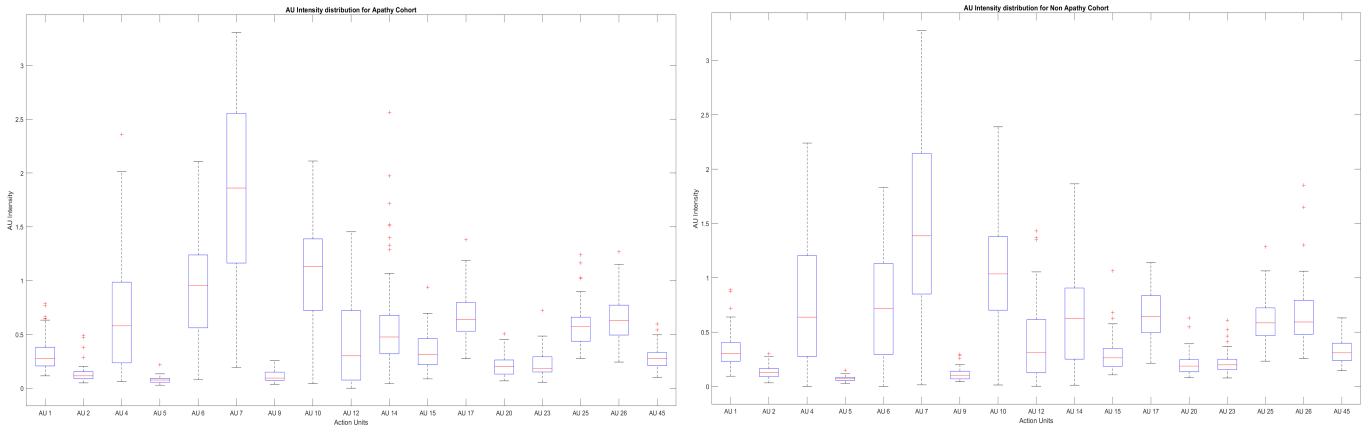
This section discusses the architecture of the proposed model. Consider V is a set of original videos, where $V = \{V_1, V_2, ...{}_n\}$. Corresponding to each video, there are two response videos for positive and negative narration. Consider the set of positive and negative videos as V_P and V_N, respectively. Each video in set $\{V_P, V_N\}$ is divided into equal sized chunks. The division of chunks from original videos has several observed benefits - (i) uniform sized clips are obtained from original videos of non-uniform duration; (ii) throughout the recording there can be a dominant and secondary emotions. Dividing in chunks may capture a wider gamut of emotions; (iii) it helps in generating large number of clips from a limited number of original videos. Consider C as a set of chunks, where $C = \{c_{P1}, c_{P2}, ...c_{Ps}, c_{N1}, c_{N2}, ...c_{Nt}\}$. Here, P and N denote a positive and a negative video, respectively and *s, t* denotes the total number of chunks corresponding to positive and negative videos, respectively. The video chunks, C, is used as an input for further training. After training a model from the features described below, chunk level prediction is converted into video level prediction using MIL. This process is further detailed in Section V.

Fig. 1 shows the proposed network architecture. The upper left network in Fig. 1 shows the overall network. The videos are divided into chunks (segments) to extract various features. Positive and negative videos are trained separately with a small network. This is to address and capture the subtle differences in expressions with varying intensity in response to positive and negative stimuli. It also enables the classifier to recognize apathetic behaviour for different emotion type.

A. Visual Features

Facial expression recognition is a significant step to identify the state of apathy in a person. Most of the state of the art models for facial expression detection are trained using data recorded from young people. Such models fail to recognize the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE



(a) Action units distribution for apathy cohort

(b) Action units distribution for non apathy cohort

Fig. 2: Distribution of action units and their intensities.

expressions of elder and aged people. The process of ageing leads to number of wrinkles and folds on a face which makes it difficult to recognize facial expressions [28]. Considering all these points, visual information is used in following two ways to obtain significantly discriminating features:

- To improve facial expression recognition in elderly, first the training is performed on a separate data of elderly people [41]. This training is performed by initializing the model with VGGFace [42] 6^{th} layer features. This pre-trained model on separate elderly data is then used to extract features for the apathy dataset. First each face is aligned and resized to 224×224 before extracting the features on apathy dataset. These features encode rich facial level information. For simplicity, this network is referred to as 'ElderFace' in the rest of the paper.
- Action Units (AU) encode the small muscle movements in a face, which are useful in recognizing the facial expression of a person. To leverage this, the intensity of AU is computed from OpenFace 2.0 toolkit [43] at frame level. OpenFace is an open source framework to extract face level information and is widely used in many studies [44], [45]. Frame wise AU intensities are extracted and used as a feature for training. The distribution of AU intensities is shown in Fig. 2. The x-axis shows 17 action units whose details are shown in Table I. The y-axis represent the AU intensities. These 17 AU's are commonly used for face level analysis [46]. The distribution shows high intensity for some AU's in non apathetic cohort. However, in some cases an apathetic person may express more, thus this subjectivity issue make it a challenging task.

Both of the above mentioned visual features are extracted and trained separately using a Gated Recurrent Unit (GRU) [47] network to learn the temporal changes. The GRU network has 5 layers with 128, 256, 512, 1024 and 2048 dimensions. The architecture of this training is shown in lower left part of Fig. 1. A Fully Connected (FC) layer with size 2048 is used after GRU layers and before using a final prediciton layer. This network is trained to learn two classes for the detection of apathetic and non apathetic behaviour in video chunks.

B. Audio Features

Analysis of audio information has been widely used for emotion recognition. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [48] defines a minimalistic feature set which is widely used to extract features from audio signals to recognize the emotion.

The GeMAPS feature set consist of 18 low-level descriptors based on frequency, energy and spectral parameters. The features included in this set are mentioned in Table II. These features are found to be beneficial to encode the emotion of a person [49]. The features are extracted for each video using OpenSMILE [50] toolkit. A small deep neural network is trained which has 5 FC layers having size 128, 256, 1024 and 2048 (upper right network in Fig. 1).

TABLE I: Description of each AU used in this study [51].

AU Number	AU Description
1	Inner brow raiser
2	Outer brow raiser
4	Brow lowerer
5	Upper lid raiser
6	Cheek raiser
7	Lid tightener
9	Nose wrinkler
10	Upper lip raiser
12	Lip corner puller
14	Dimpler
15	Lip corner depressor
17	Chin raiser
20	Lip stretcher
23	Lip tightener
25	Lips part
26	Jaw drop
45	Blink

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

C. Motion Features

Facial expressions can also be analysed by identifying the location of facial components such as nose, mouth, etc. and then focusing on specific locations. The use of geometric features for emotion recognition is already discussed in various studies [18], [52] and have shown promising outcomes. Motivated by this, 68 2D facial landmarks are extracted from OpenFace 2.0 [43]. The landmark points are converted to a 1D tensor for each frame. This represents and encodes local motion within the face. The training of these motion features is done similar to other visual features (lower left network in Fig. 1).

D. Fusion of Features

The features from the above mentioned models (ElderFace, AU, audio and motion) are combined together for the end-task of final video level apathy prediction. The fusion of features is depicted in lower right network in Fig. 1. In this figure, X (as in P_X) denotes any feature type such as P_Action_Units or P_Audio.

First, the model trained on positive and negative data are combined together (such as P_Audio and N_Audio are combined to form P+N_Audio). To accomplish this, chunk level features for P_X and N_X are computed from the pre-final layer of the trained model. These 2D chunk level features are converted to 3D video level features and then are concatenated together. Three FC layers are used then for training with kernel size 128, 512 and 1024 to encode this combination of models trained on positive and negative data.

After combining the positive and negative video level features (P+N_X), four components used in this study (Elderface, action units, audio and landmarks) are fused together. This fusion is also done after concatenating the pre-final layer features of the trained model and then training it for three FC layers (similar to the combination of P_X and N_X). The final model obtained by this is used to make apathetic and non-apathetic predictions.

V. Experiments and Results

A. Pre-processing

Each video is divided corresponding to positive and negative narration based on the given timestamps. FFmpeg library is used for video clipping and audio extraction. The size of each chunk is fixed to be 32 frames. The idea of selecting only 32 frames is to keep a trade off between performance and computational efficiency. Throughout the paper, term video is used to refer the complete given positive or negative video whereas, chunk is used to refer the subset of video having 32 frames. Each video has different duration, hence, multiple number of chunks are extracted from each video. Total 5906 such chunks are produced after splitting the videos. Further details of the data are given in Table III. To extract the facial level information, OpenFace 2.0 [43] is used to obtain frame wise aligned faces, active action units (AU) and 2D landmarks. Audio features are extracted by using OpenSMILE [50] toolkit.

TABLE II: Details of the audio features used in experiments [48].

Frequency	Energy	Spectral	Temporal features
Pitch	Shimmer	Alpha ratio	Rate of loudness peaks
Jitter	Loudness	Hammarberg Index	Mean length and standard deviation of voiced regions
Formant 1, 2, 3 frequency	Harmonic to noise ratio	Spectral Slope 0–500 Hz and 500–1500 Hz	Mean length and standard deviation of unvoiced regions
Formant 1		Formant 1, 2, and 3 relative energy	No. of continuous voiced regions per second
		Harmonic difference H1–H2 and H1–A3	

TABLE III: Details of the data used in experiments.

	No. of participants	No. of videos	No. of chunks
Apathetic	28	56	2231
Non apathetic	42	84	3675
Total	70	140	5906

B. Evaluation metric

Similar to [18], Leave One Subject Out (LOSO) cross-validation strategy is adopted while training. The accuracy and weighted F1-score is computed from the LOSO experiments. First, the network is trained by using 3 fold cross validation across each modality. After performing the hyper-parameter tuning, experiments are performed by LOSO cross validation. The results mentioned in this paper are obtained from LOSO cross validation.

C. Experimental Details

The experiments are performed by using binary cross entropy loss function and swish activation. Softmax activation is used in the prediction layer of the model. The SGD optimizer is used along with 0.01 learning rate while training the chunks. For MIL, Adam optimizer is used with learning rate 0.001. These parameters are selected by first training a network in 3 fold cross validation. Later, the final experiments are performed for fixed number of epochs. All the results mentioned in the paper are provided for videos rather than the chunks.

D. Results

Table IV shows the accuracy and the F1-score of positive (P), negative (N) and fusion of positive and negative (P+N) videos. The performance is mentioned before and after fusing each type of features i.e. ElderFace, action units, audio and landmarks. For action units and landmarks, features from negative videos are found to be more discriminative. In case of ElderFace and audio features, the difference between the accuracy of positive and negative videos is limited. The combination of positive and negative features for ElderFace and landmark features, respectively, achieves only 54.28% accuracy in both the cases. It is evident that these features significantly discriminate apathetic behaviour in positive or

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE IV: Performance of the proposed approach for the apathy prediction. Here, P and N represents positive and negative videos, respectively.

Features	ElderFace			Action Units			Audio			Landmarks			Combined
	P	N	P+N	P	N	P+N	P	N	P+N	P	N	P+N	
Accuracy (%)	60.00	60.00	54.28	51.42	57.14	60.00	54.28	52.85	61.42	55.71	57.14	54.28	**75.71**
F1-score	0.602	0.603	0.517	0.510	0.571	0.602	0.535	0.529	0.617	0.560	0.575	0.547	**0.736**

negative videos, separately yielding higher accuracy (Table IV). However, combining the positive and negative narration makes it difficult to differentiate solely relying on these visual features. This performance difference can be attributed to the fact that while narrating a positive and negative narration, there are differences in the exhibited expressions and the intensity of the expressions varies as well. It is highly probable that a person with apathy may not express high valence and positive emotion with the same intensity as low valence or negative emotions and vice versa.

The results also show that only expression level information is not sufficient given the complexity of the apathy detection task. It is clear from these results that it is a challenge to recognize the facial expressions of elderly people due to the wrinkles present in the face. There is an increase in the accuracy after combining positive and negative videos in case of action units and audio features respectively (Table IV).

The largest increase in the accuracy before and after combining positive and negative narrations, is observed in the case of the audio features. It shows the efficiency of audio features in emotion and apathy detection. The fusion of multiple modalities produced a hike in performance. The combined model gives 75.71% accuracy and 0.736 F1-score. The result shows that each feature learned by different modality is complementary to other and are equally contributing for the task of automatic apathy detection.

VI. Conclusion

An automatic audio-visual method for apathy detection is proposed in this paper. The proposed MIL based method exploits facial expressions, action units, facial landmarks and audio to detect apathetic and non-apathetic behaviour. The experiments are performed to provide an insight towards using vision based methods to accurately detect apathetic behaviour. Although, vision based facial expression recognition methods have achieved a very high accuracy, it is still difficult to use them to detect apathy in elder people. Analyzing emotions of elderly people is still a challenging task. To address this, pre-training is performed on a separate elderly faces dataset and thus fine tuned for better emotion recognition in elderly.

Another challenge is subjectivity issue. The intensity of expressions may vary for positive and negative videos. There is a possibility that apathetic people express differently for high and low valence emotions and with varying intensities. Its also plausible that an apathetic person could be more expressive in general as compared to some non-apathetic participants in the cohort. The problem is not only challenging due to its subjective dependency, however the varying range of emotion exhibition in different genders adds another complexity. These range of issues increase the intricacies of the automatic apathy detection task.

In future, the proposed network will be improved by considering the intensity of expressions which will help to improve the distinction of apathetic and non apathetic behaviour. Changes will be made in the network to utilize the gender information without using any extra labels for training. Further investigation in common and prevalent AUs and their intensities between apathetic and non-apathetic cohorts would be another interesting avenue to explore.

References

[1] R. S. Mann, "Differential diagnosis and classification of apathy," *Am J Psychiatry*, vol. 147, no. 1, pp. 22–30, 1990.

[2] 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.

[3] G. Simons, H. Ellgring, and M. Smith Pasqualini, "Disturbance of spontaneous and posed facial expressions in parkinson's disease," *Cognition & Emotion*, vol. 17, no. 5, pp. 759–778, 2003.

[4] P. H. Robert, E. Mulin, P. Malléa, and R. David, "Apathy diagnosis, assessment, and treatment in alzheimer's disease," *CNS neuroscience & therapeutics*, vol. 16, no. 5, pp. 263–271, 2010.

[5] W. J. Weiner, L. M. Shulman, and A. E. Lang, *Parkinson's disease: A complete guide for patients and families*. JHU Press, 2013.

[6] L. Nobis and M. Husain, "Apathy in alzheimer's disease," *Current opinion in behavioral sciences*, vol. 22, pp. 7–13, 2018.

[7] V. Parekh, P. S. Foong, S. Zhao, and R. Subramanian, "Aveid: Automatic video system for measuring engagement in dementia," in *23rd International Conference on Intelligent User Interfaces*, 2018, pp. 409–413.

[8] M. H. Li, T. A. Mestre, S. H. Fox, and B. Taati, "Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation," *Journal of neuroengineering and rehabilitation*, vol. 15, no. 1, p. 97, 2018.

[9] S. Ahmed, M. Qaosar, R. W. Sholikah, and Y. Morimoto, "Early dementia detection through conversations to virtual personal assistant," in *2018 AAAI Spring Symposium Series*, 2018.

[10] M. Rodrigues Makiuchi, T. Warnita, K. Uto, and K. Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[11] Y. Liu, J. Chen, C. Hu, Y. Ma, D. Ge, S. Miao, Y. Xue, and L. Li, "Vision-based method for automatic quantification of parkinsonian bradykinesia," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 10, pp. 1952–1961, 2019.

[12] 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, 2017.

[13] S. E. Starkstein, "Apathy and withdrawal," *International Psychogeriatrics*, vol. 12, no. S1, pp. 135–137, 2000.

[14] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

[15] K. Osborne-Crowley, S. C. Andrews, I. Labuschagne, A. Nair, R. Scahill, D. Craufurd, S. J. Tabrizi, J. C. Stout, T.-H. Investigators *et al.*, "Apathy associated with impaired recognition of happy facial expressions in huntington's disease," *Journal of the International Neuropsychological Society*, vol. 25, no. 5, pp. 453–461, 2019.

[16] K. López-de Ipiña, J. B. Alonso, N. Barroso, M. Faundez-Zanuy, M. Ecay, J. Solé-Casals, C. M. Travieso, A. Estanga, and A. Ezeiza, "New approaches for alzheimer's disease diagnosis based on automatic spontaneous speech analysis and emotional temperature," in *International Workshop on Ambient Assisted Living*. Springer, 2012, pp. 407–414.

[17] A. König, N. Linz, R. Zeghari, X. Klinge, J. Tröger, J. Alexandersson, and P. Robert, "Detecting apathy in older adults with cognitive disorders using automatic speech analysis," *Journal of Alzheimer's Disease*, vol. 69, no. 4, pp. 1183–1193, 2019.

[18] S. Happy, A. Dantcheva, A. Das, R. Zeghari, P. Robert, and F. Bremond, "Characterizing the state of apathy with facial expression and motion analysis," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[19] 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.

[20] M. F. Folstein, L. N. Robins, and J. E. Helzer, "The mini-mental state examination," *Archives of general psychiatry*, vol. 40, no. 7, pp. 812–812, 1983.

[21] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[22] 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.

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

[24] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983–2991.

[25] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," *arXiv preprint arXiv:1711.04598*, 2017.

[26] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: challenges to inferring emotion from human facial movements," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

[27] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.

[28] M. Fölster, U. Hess, and K. Werheid, "Facial age affects emotional expression decoding," *Frontiers in psychology*, vol. 5, p. 30, 2014.

[29] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[30] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[31] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2003, pp. 577–584.

[32] A. d. Garcez and G. Zaverucha, "Multi-instance learning using recurrent neural networks," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012, pp. 1–6.

[33] Z.-H. Zhou and M.-L. Zhang, "Neural networks for multi-instance learning," in *Proceedings of the International Conference on Intelligent Information Technology, Beijing, China*, 2002, pp. 455–459.

[34] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Advances in neural information processing systems*, 2006, pp. 1417–1424.

[35] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu, "Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 256–263.

[36] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1626–1630.

[37] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 4, pp. 862–875, 2015.

[38] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3460–3469.

[39] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 603–611.

[40] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[41] N. C. Ebner, M. Riediger, and U. Lindenberger, "Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behavior research methods*, vol. 42, no. 1, pp. 351–362, 2010.

[42] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[43] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[44] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38–45.

[45] 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 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 599–603.

[46] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–6.

[47] 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.

[48] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[49] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[50] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[51] P. Ekman and W. V. Friesen, *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[52] N. Alugupally, A. Samal, D. Marx, and S. Bhatia, "Analysis of landmarks in recognition of face expressions," *Pattern Recognition and Image Analysis*, vol. 21, no. 4, pp. 681–693, 2011.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE