Automatic Detection of Self-Adaptors for Psychological Distress

Weizhe Lin[1], Indigo Orton[2], Mingyu Liu[3] and Marwa Mahmoud[4]
[1] Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[2] [4] Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom
[3] Department of Physics, University of Oxford, Oxford, United Kingdom
{[1]wl356@, [2]indigo.orton@cl., [4]mmam3@}cam.ac.uk; [3]mingyu.liu@queens.ox.ac.uk

Abstract— **Psychological distress is a significant and growing issue in society. Automatic detection, assessment, and analysis of such distress is an active area of research. Compared to modalities such as face, head, and vocal, research investigating the use of the body modality for these tasks is relatively sparse. This is, in part, due to the lack of available datasets and difficulty in automatically extracting useful body features. Recent advances in pose estimation and deep learning have enabled new approaches to this modality and domain. We propose a novel method to automatically detect self-adaptors and fidgeting, a subset of self-adaptors that has been shown to be correlated with psychological distress. We also propose a multi-modal approach that combines different feature representations using Multi-modal Deep Denoising Auto-Encoders and Improved Fisher Vector encoding. We also demonstrate that our proposed model, combining audio-visual features with automatically detected fidgeting behavioral cues, can successfully predict distress levels in a dataset labeled with self-reported anxiety and depression levels. To enable this research we introduce a new dataset containing full body videos for short interviews and self-reported distress labels.**

I. Introduction

Psychological distress and mental disorders are significant threats to global health. According to the World Health Organization (WHO), an estimated 450 million people around the world are affected by different kinds of psychological distress and mental disorders. Despite existing strategies for the treatment of distress, such as depression, it is estimated that nearly two-thirds of people suffering distress have never received help from a health professional [1].

Early detection of distress is consistently noted as a key factor in treatment and positive outcomes. Early detection requires an ongoing assessment to identify distress when it begins. Self-evidently, ongoing assessment at scale is prohibitive when performed manually. As such, automatic detection of psychological distress, and specific mental disorders, is an active area of research.

Currently, the most effective automated distress detection approaches utilize multi-modal machine learning. These modalities include facial, head, eye, linguistic (textual), vocal, and body. A brief review is presented in Section II.

There are significant challenges to body modality research, particularly within automatic distress detection, including the lack of relevant data, the inability to share much of the data, and the difficulty in gathering such data. Specifically, the combination of full-body data (either sensor-based or video-based) with psychological distress labels is rare. Compounding this rarity is the private and sensitive nature of the data, which means such datasets are rarely shared.

In this paper, we are primarily concerned with self-adaptors within the body modality, as body expressions have been shown to be predictive in a number of affective computing tasks [2]. Self-adaptors are self-comforting gestures including any kind of touching on other parts of the body, either dynamically or statically [3], [4]. Fidgeting, a subset of self-adaptors, is the act of moving about restlessly, playing with one's fingers, hair, or personal objects in a way that is not peripheral or nonessential to ongoing tasks or events [5]. Patients with depression often engage in self-adaptors [6]. Fidgeting has been seen and reported in both anxiety and depression [4], individuals with autistic spectrum disorder also exhibit fidgeting behaviors. With manually annotated data, Scherer *et al.* [7] reported a longer average duration of self-adaptors as well as fidgeting for distressed participants. More recent advances in the state-of-the-art for pose estimation [8] enable accurate pose data on a broader set of datasets and thus new approaches to fidgeting detection and broader incorporation of fidgeting in multi-modal systems.

Therefore, in this paper, we propose to use a hierarchical model to automatically detect self-adaptors as well as fidgeting, which has been shown to be predictive of psychological distress. A Multi-modal Deep Denoising Auto-Encoder (multi-DDAE) is utilized to encode per-frame features. Improved Fisher Vector encoding [9] is then used to generate per-sample representation. Finally, we demonstrate these features are discriminative in psychological distress detection.

The contributions of this paper can be summarized as follows:

1) We introduce a new audio-visual dataset containing recordings of non-clinical interviews along with distress labels from established psychological evaluation questionnaires.

2) We introduce a hierarchical model for automatic detection of self-adaptors (including fidgeting) from visual data. We validated this detector with a publicly available fidgeting dataset with manual labels.

3) As a step of concept-proof, we presented a multi-modal feature fusion framework to perform distress classification and thus demonstrated the importance of self-adaptor, specifically fidgeting, features. We evaluate this classifier for depression and anxiety prediction.

The full framework is available at:
https://github.com/LinWeizheDragon/AutoFidgetDetection

II. Related Work

In this section, we focus on related work of psychological distress detection, including its practical modalities and multi-modal fusion frameworks.

A. Facial and head modality

Facial Action Coding System (FACS) [10] has long been used to taxonomize human facial movements by their appearance on the face, which yields the concept of Facial Action Units (AUs). For example, the Audio/Visual Emotion Challenge (AVEC) used AUs features as a basic descriptor for its psychological distress detection tasks.

Much work has been done using the facial and the head modalities. For example, Yang *et al.* [11] proposed "Histogram of Displacement Range (HDR)", which is a measurement of the amount of facial landmark movements. Joshi *et al.* [12] presented a categorization analysis framework which consists of "bag of facial dynamics" and "histogram of head movements". Dibeklioğlu *et al.* [13] [14] feature-engineered dynamic representation (e.g. velocity, acceleration, and standard deviation of motion) for facial landmark movement and head motion.

Psychomotor retardation refers to a slowing-down of thought and a reduction of physical movements in an individual. Sobin *et al.* [15] demonstrated the correlation between psychomotor retardation and depression. Syed *et al.* [16] handcrafted descriptors using craniofacial movements in order to capture the psychomotor retardation, and then made predictions of depression.

Some other features such as smiling (intensity and duration) [17], eye blink rate [18], eye lid movement [16], gaze activity [19] [20], and gaze orientation [17] are also shown to be predictive of depression.

B. Audio modality

Acoustic features of speech can be predictive of distress irrespective of the speech content [21]. For example, Ozdas *et al.* [21] assessed the risk of suicide by detecting the fluctuations in the fundamental frequency of people's speech. Dibeklioğlu *et al.* [13] explored the use of vocal prosody for depression detection. Similarly, Syed *et al.* [16] investigated the use of turbulence in speech patterns.

In addition, in AVEC challenges, low-level descriptors of voice signals, such as Mel-frequency Cepstral Coefficients (MFCCs), are provided, leading to many multi-modal methods incorporating these acoustic features for distress and illness detection [11], [22].

C. Body modality

There are two primary approaches for representing the body modality: a) traditional computer vision feature detection algorithms, and b) skeletal models.

The first approach does not target a specific part of the body but instead extracts generic feature points from the recording to represent the body and gestures. For example, Joshi *et al.* [12] computed Histogram of Gradients (HOGs) and Histogram of Optical Flow (HOFs) around the generic Space-Time Interest Points (STIPs) extracted from the videos, and then generated a "Bag of Body Dynamics" feature for further depression classification.

The second approach extracts body modality-specific interest points, the most famous one of which is the skeletal model. Such models have gained popularity in the past few years for action recognition tasks and could be used to generate more specific and concrete features by feature engineering [23].

In terms of fidgeting detection, Mahmoud *et al.* [3] developed a novel framework for generating automatic multimodal descriptors of rhythmic body movement, which features its ability to recognize rhythmic body motion and rhythmic fidgeting. They extracted Speeded-Up Robust Features (SURFs) interest points around Microsoft Kinect pose points and then detected rhythmic behaviors from the trajectories of interest points.

However, there are two limitations in their automated system when applied to distress detection: 1) Their dataset is based on actors, so the behavior is not natural. For example, in real interviews, participants don't always fidget with a rhythmic pattern. 2) The trajectory data is noisy and their method could not sufficiently handle the complexity. As such, they were only able to achieve 59% recognition.

D. Multi-modal Learning

Psychological distress is expressed through all modalities. Many frameworks were proposed in AVEC 2017 and 2018 challenges to automatically detect psychological distress using multi-modal approaches [11], [22], [24], [25], [26], [22]. However, due to the limited data available in the challenges, most frameworks utilize only low-level features (e.g. the latent activation of CNN layers, MFCCs), rendering the frameworks uninterpretable. As such the basis of their decisions cannot be supported by psychological literature.

III. Dataset

This dataset is designed to enable investigation of the body modality for use in automatic detection of distress. Details of methodology (e.g. type of questionnaires and detailed implementation) are described by Orton [23].

A. Overview and design

Inspired by the DAIC dataset collection method [27], a human interviewer asks a series of open-ended conversational questions such that the participant expresses naturalistic behavior. In order to keep behaviors naturalistic, participants were not aware of the main goal of the study, which is automatic analysis of behavioral cues. Instead, they were told that the experiment is for building models that can help in mental well-being. This ensured that their behavior would be as natural as possible.

The dataset is labeled with participant responses to self-evaluation questionnaires for assessing distress and personality traits, as well as demographic labels such as gender. The distress questionnaires are: the PHQ-8 [28], [29] for depression, GAD-7 [30] for anxiety, SSS-8 [31] for somatic

Label	Range	Mean	Covariance with Depression
Distress			
Depression	0–19	7.43	-
Anxiety	0–19	7.00	86.15%
Perceived stress	1–30	18.17	84.00%
Somatic symptoms	1–27	9.06	74.16%
Personality			
Extraversion	3–31	16.37	-30.49%
Agreeableness	12–34	25.67	-42.21%
Openness	7–39	27.29	4.29%
Neuroticism	1–31	16.86	80.00%
Conscientiousness	10–36	21.46	-46.41%
Demographic			
Gender	-	-	9.47%
Age	18–52	25.40	-11.09%

TABLE I
GENERAL STATISTICS REGARDING QUESTIONNAIRE AND DEMOGRAPHIC RESULTS WITHIN THE DATASET. THIS TABLE DEMONSTRATES THERE ARE NO CONFOUNDING CORRELATIONS WITH THE DEPRESSION LABEL.

symptoms, and the PSS [32] for perceived stress. Personality traits are measured using the Big Five Inventory [33].

As a result, the dataset includes fully natural non-acted expressions, including facial expressions, body motion, gestures, and speech.

B. Preliminary Analysis

The dataset contains 35 interviewed participants with a total video duration of 07:50:08.

Study participants consisted primarily of the students and staff from the University of Cambridge. Participants were selected to balance distress scores (PHQ-8 and GAD-7 scores) and gender.

General statistics regarding the questionnaire and demographic results within the dataset are provided in Table I. Covariance is presented as normalized covariance values, also known as the correlation coefficient.

We assess confounding correlations based on the depression label, as much of the related work focuses on depression. While the distress measures, anxiety, perceived stress, and somatic stress, are strongly correlated with depression, the personality measures have below 50% covariance with the exception of neuroticism which has an 80% covariance. The demographic measures, gender, and age are negligibly correlated, with 9.47% and -11.09% covariance, respectively. Finally, the interview duration is not correlated with any questionnaire result (less than 25% covariance with all labels). Thus, we can be confident that there are no confounding correlations with personality scores or demographics.

For a thorough analysis and validation of the dataset contents see Orton [23].

IV. METHOD

Our method consists of four primary steps: feature extraction, fidgeting detection, feature encoding, and distress prediction.

A. Feature extraction

1) Visual Features: For each video, we use state-of-the-art tools, OpenPose [8] and OpenFace 2.2 [34], to extract pose estimation, AUs, and gaze directions.

However, OpenPose and OpenFace do not take into account the consistency of the keypoints across time, causing the keypoints to fluctuate highly and introducing noise to the real motion. Besides, there are some frames where OpenPose or OpenFace fail to extract all pose points or gaze features, respectively. To overcome these problems, we infer the missing data via Cubic Spline Interpolation across the whole sequence. We then smooth the data using a Savitzky-Golay filter [35] (window length is 11 and the order of the polynomial is 3).

2) Audio Features: Speaker diarisation involves partitioning an audio stream into homogeneous segments according to the speaker's identity. In order to distinguish the speech of the interviewer and the participant, we use the open-source Speaker-Diarization project [36] which utilizes an Unbounded Interleaved-State Recurrent Neural Network (UIS-RNN) [37], to extract speaker identities with respect to the time axis. We then conduct manual check to assign correct diarization labels to the participant and the interviewer.

We also use pyAudioAnalysis [38] to extract MFCCs.

B. Automatic Fidget Detection

In this section, we present our fidgeting detection system in three subsections. We start by exploring the self-adaptors/fidgeting encoding and the overall hierarchical design. Then we show the methods of building the two essential detectors of our hierarchical model in the following two subsections. For each detector, we demonstrate the detector design, and then present the labeling strategy which provides reliable labels for training and evaluation. In order to validate the effectiveness of our automated fidget detector before moving onto distress classification, we perform detector evaluation in this section.

1) Overall Design and Encoding: Given fidgeting lacks a definition with a broad agreement so far, we define fidgeting based on a two-step hierarchical model. As shown in Table II, we first identify self-adaptors, which we define as low-level location events (e.g. H2H, H2F). Secondly, action events (i.e. DYNAMIC, STATIC) of hand/leg are classified by the *DYNAMIC/STATIC Classifier*. Fidgeting is then defined as a combination of low-level self-adaptors and action events. Specifically, we define three types of fidgeting: cross hand fidgeting, single hand fidgeting, and leg/feet fidgeting.

2) Self-adaptor Detector:

a) Design: Each body location is represented using a bounding box. Self-adaptors are defined as overlapping bounding boxes. We represent the hand and face using the smallest rectangular box bounding all corresponding hand

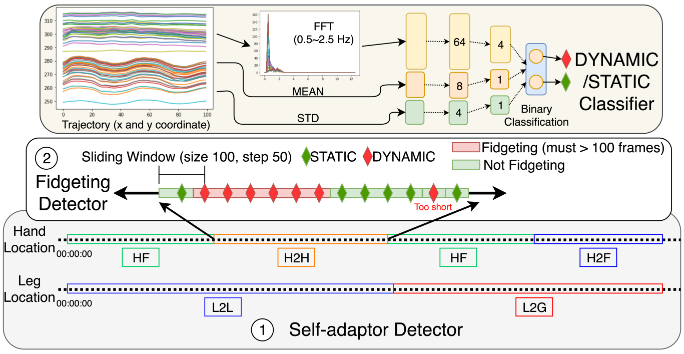

Fig. 1. Hierarchical self-adaptor detection workflow. (1) First detect hand/leg location (2) Classify motion using *DYNAMIC/STATIC Classifier* and then finally combine location and motion to give high-level fidgeting event. Figure shows the detection of H2H (Hand to hand) fidget. Same principle applies to other fidgets.

Self-adaptors	**Description**
H2H	Hand to hand
H2A	Hand to arm
H2L	Hand to leg
H2F	Hand to face
HF	Hand free (when not belong to any of above)
L2G	Both legs on ground
L2L	Leg on the other leg (crossed legs)
Action Events	**Description**
DYNAMIC	Moving obviously
STATIC	No obvious movement is observed

Fidgeting Type	**Combination**
CHF (Cross Hand Fidgeting)	H2H + DYNAMIC
SHF (Single Hand Fidgeting)	{H2A, H2L, H2F, HF} + DYNAMIC
SHF-L (to leg only)	H2L + DYNAMIC
SHF-F (to face only)	H2F + DYNAMIC
SHF-A (to arm only)	H2A + DYNAMIC
LFF (Leg/Feet Fidgeting)	{L2G, L2L} + DYNAMIC

TABLE II

SELF-ADAPTOR AND FIDGET ENCODING BOOK

or face keypoints. The forearms, upper arms, lower legs, and upper legs' bounding boxes' long sides are aligned with the connection between two joints from OpenPose, while the width is a free parameter tuned for the best automatic detection performance.

First, H2H self-adaptor events are detected (i.e. when the two hands' bounding boxes overlap). Then all other hand-based self-adaptor events are detected, for all segments of the video not containing H2H segments.

All self-adaptors, except for HF, must be longer than 100 frames (around 4 seconds with the frame rate of 26). This reduces noise from detected self-adaptor events.

b) Labeling and Evaluation: In order to validate our self-adaptor detector, we manually labeled 4 participants' videos, a total duration of 59 minutes. The inter-labeler agreement was checked using Krippendorff's alpha. Each frame was labeled with one of the self-adaptor codes from Table II. Within these videos, participants perform different self-adaptors and each event has a minimum total duration of 5 minutes, with the exception of H2F.

As shown in Table III, the alpha agreement for left-hand location is 0.823 for right-hand location is 0.888 and for leg location is 1.00. This suggests good agreement between the annotators and thus label reliability.

Hand Self-adaptors (left/right)

	Precision	Recall	F1 Score
H2H	1.00/1.00	0.99/0.99	1.00/1.00
H2A	1.00/NA	0.64/NA	0.79/NA
H2L	0.96/0.88	0.86/0.82	0.91/0.85
H2F	NA/1.00	NA/1.00	NA/1.00
HF	0.63/0.83	0.99/0.98	0.77/0.90
Alpha Score:	0.823/0.888		

Leg Location

	Precision	Recall	F1 Score
L2L	1.00	1.00	1.00
L2G	1.00	1.00	1.00
Alpha Score:	1.000		

TABLE III

SELF-ADAPTOR DETECTION EVALUATION

3) Fidgeting Detector:

a) Design: As shown in Fig. 1, the *DYNAMIC/STATIC Classifier* operates on **optical flow** from a sliding window across the video (size 100 frames, step 50 frames). To classify the action (DYNAMIC/STATIC), hand movements (especially fingers) and leg movements require optical flow to obtain smooth trajectories, given OpenPose estimations become unreliable when hands intersect or are occluded. We thus initialize the optical flow with the OpenPose estimations at the beginning of each slice.

We choose Fast Fourier Transform (FFT), standard deviation (STD), and mean values (MEAN) of point trajectories as our input features (in this case, number of trajectories is 2 × number of keypoints as we have 2-D data for each keypoint). For fidgeting, we are more interested in the cyclic motion with a frequency ranging from 0.5Hz to 2.5Hz [3]. Therefore, we extracted the spectrum data within the range $[0.5, 2.5]$ Hz. As we analyze on slices of length 100, the dimension of FFT spectrum data that is within $[0.5, 2.5]$ Hz is always fixed at $41\times$ number of trajectories. We averaged over the FFT values that have the same frequency to produce an FFT feature of length 41. As for the STD and MEAN features, we simply calculate along the time axis and give a vector with a length of the number of trajectories for each feature.

4) Labeling and Evaluation: To train and evaluate the *DYNAMIC/STATIC Classifiers*, accurate labeling is required. Three classifiers are required to cover the three categories of detected self-adaptors: {H2H}, {H2A, H2L, H2F, HF}, and {L2G, L2L}.

We labelled DYNAMIC/STATIC on each of the three categories. We randomly sampled and labeled approximately

30% of slices for each category in every video.

Two researchers labeled the data independently. As shown in Table IV, we first manually dropped the slices with a wrong category label (e.g. a slice is detected as H2H while it's in fact not). The number of slices that have a correct category label is shown as "Correct". Secondly, we labeled DYNAMIC/STATIC and dropped the slices that lack a consensus between two researchers. The number of slices with an agreement is shown as "Agreed". The high percentage of both "Correct" and "Agreed" suggests the good performance of our self-adaptor detection and also the high reliability of action labels.

Category	Total	Correct	Agreed
BOTH: H2H	3962	3922 (99%)	3793 (96%)
LEFT:{H2A, H2L, H2F, HF}	1614	1566 (97%)	1539 (96%)
RIGHT:{H2A, H2L, H2F, HF}	1620	1588 (98%)	1563 (96%)
{L2G, L2L}	6536	6536 (100%)	6196 (95%)

TABLE IV

HAND/LEG ACTION LABELLING OVERVIEW

Having reliable slice labels, we then partitioned participants into 5 folds and performed slice-level cross-validation. We report accuracy, F1 score, and their respective standard deviations.

Category	Acc.	Acc. Std.	F1	F1 Std.
BOTH: H2H	0.833	0.019	0.834	0.019
LEFT:{H2A, H2L, H2F, HF}	0.884	0.025	0.884	0.026
RIGHT:{H2A, H2L, H2F, HF}	0.895	0.026	0.894	0.026
{L2G, L2L}	0.875	0.022	0.871	0.021

TABLE V

DYNAMIC/STATIC CLASSIFIER EVALUATION (LEFT MEANS LEFT HAND, RIGHT MEANS RIGHT HAND, BOTH MEANS BOTH HANDS)

As shown in Table V, the detector achieved generally high accuracy and F1 score with low standard deviations. Though the hand actions are difficult even for researchers to label, the detector can successfully classify more than 80% of slices.

C. Feature encoding

1) Fidget feature processing: Having extracted low-level features from each frame we combine them to form high-level descriptors of fidgeting behavior (SHF, CHF, and LFF as shown in Table II). The Fidget_pure feature group is formed by {HCF, SHF-L(left hand), SHF-L(right hand), SHF-A(left hand), SHF-A(right hand), SHF-F(left hand), SHF-F(right hand), LFF}. The Fidget_pure group is combined with a participant speaking feature array to form the full fidget feature group, enabling us to investigate whether fidgeting and speaking co-occurrence is relevant. This participant speaking feature array indicates whether the participant is speaking during a frame. This is calculated using the previously described diarisation data.

After all the feature extraction, we have several feature groups shown in Table VI.

Feature Group	Dimension	Description
Fidget	9	fidget feature&speaking array
Fidget_pure	8	fidget feature only
Gaze	8	Gaze direction
AUs	35	Action Units
MFCCs	13	

TABLE VI

FEATURE GROUPS

2) Per-frame representation: In order to capture more useful feature representations and reduce the dimensionality, and inspired by our previous work [39], the modalities are combined using a Multi-modal Deep Denoising Auto-Encoder (multi-DDAE). As shown in Fig. 2, each modality is encoded through a dense layer and then all are concatenated to yield the last shared dense layer which provides the representation we use. The shared layer is then inversely decoded to generate each modality. We optimized the hyper-parameters of the auto-encoder via several experiments so that the dimensions of hidden layers are $\{0.5d, 0.25d, 0.5d\}$ where d represents the input dimension of each node, and the noise applied at the input is 0.1 Gaussian noise. The training optimization target is the joint Mean Square Error (MSE) of the MSEs of the feature group at each node (later we fixed the loss weights to be 0.35 for the fidget feature group while 0.1 for others, as we are more interested in fidgeting in our experiments).

3) Whole video representation: As the videos are of different lengths, it's necessary to unify the dimensionality of the per-video representation. Though Fisher Vector was originally proposed to aggregate visual features [9], it has become popular in social signal processing such as bipolar disorder [40] and depression recognition [41]. Inspired by these applications, we apply a Gaussian Mixture Model to cluster similar per-frame representations and then use an Improved Fisher Vector encoding to obtain a fixed-length representation. As a result, the feature is transformed from $n\_frames \times feature\_dim$ to $2 \times GMM\_Kernel\_num \times feature\_dim$.

D. Distress classification

We apply a Random Forest to select important features from the per-video representation. The selected features are used by the classifier. We experiment with two classifiers: 1) a logistic regression based classifier (LR) using a binary threshold of 0.5; and 2) a Multi-Layer Perception (MLP) with two softmax outputs for binary classification.

As the available samples are limited and the useful features vary across individual differences, label smoothing is applied to the MLP model in order to further boost the performance. More formally:

$$L\_new = L \times (1 - s) + \frac{s}{n} \tag{1}$$

where L is the one-hot label at softmax outputs, s is the smoothing parameter, and n is the number of classification

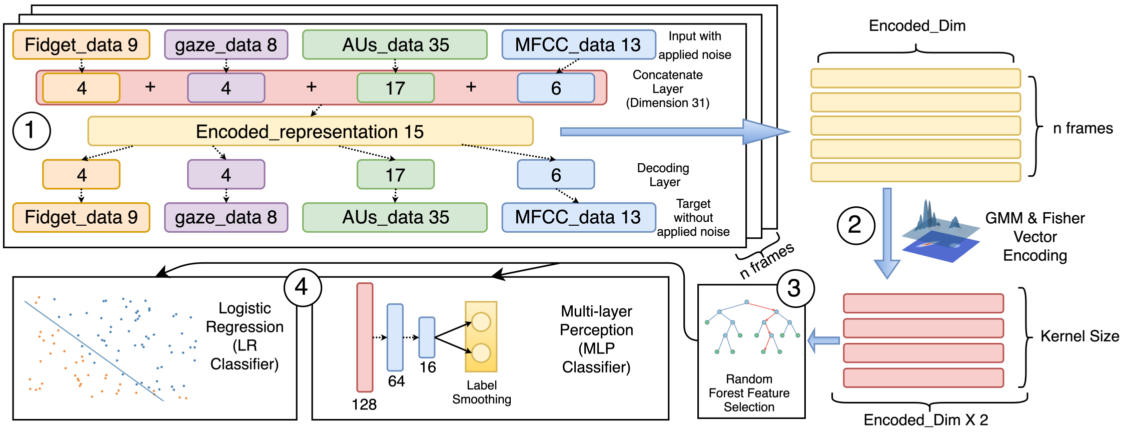

Fig. 2. Multi-modal fusion & classification pipeline. Dashed arrow represents a fully connected neural network between dense layers. Pose estimation, gaze, Action Units, and MFCC data are extracted from videos. Fidget features are computed using the method described in Section IV. (1) All features are fed into a Multi-modal Deep Denoising Auto-Encoder (multi-DDAE) to generate a compact per-frame encoded representation. (2) These per-frame features are then compressed into a whole video representation using a Gaussian Mixture Model (GMM) and Fisher Vector combination. (3) Random Forest feature selection is performed. (4) Finally, a classifier predicts a given label. We experiment with two classifiers, a logistic regression classifier and a Multi-layer Perception.

classes. For example, when smoothing is 0.2, the one-hot label $\{0, 1\}$ will become $\{0.1, 0.9\}$, which lowers the confidence of training samples but also reduces overfitting.

V. Evaluation and Results

We present our evaluations in three sections to demonstrate the validity and potential of fidget features. First, we present baseline distress classification results on our dataset. Next, we present results for our full multi-modal classifier pipeline, where we investigate the effects of hyper-parameters on the performance given a small dataset. Finally, we apply our fidget detector to a publicly available dataset [3] to demonstrate its accuracy and generalisability beyond our dataset.

All results are calculated as the mean of 3-fold cross-validation results. All experiments and cross validation are participant-independent.

A. Baselines

We present baseline models using Gaussian kernel Support Vector Machines (SVMs) on each individual feature group used in our multi-modal model (listed in Table VI). They are evaluated for a binary depression label and a binary anxiety label. These models provide a simple and common baseline for our dataset. For the baseline SVM, we use the mean value for each feature over the whole sample, thus providing a normalized representation with mean values of all the features. Results are presented in Fig. 4.

These baseline models demonstrate two points: first, classification within our dataset is complex; and second, our fidget features are not trivially predictive of distress, but rather require learnt representations.

B. Multi-modal distress classification

We present the best performance of different feature group combinations using our multi-modal fusion framework. As Random Forest takes in labels to find most discriminative features, this feature selection is only performed on the training set and selected features will be applied to the test set, which prevents label leaking.

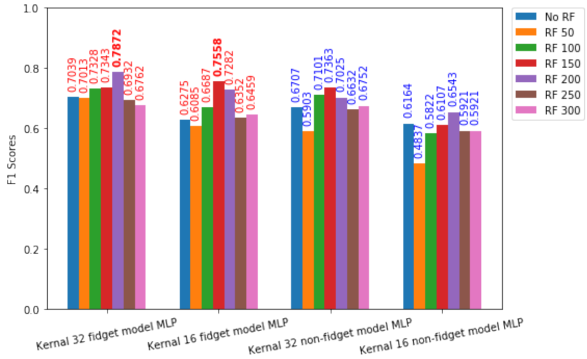

Fig. 3. Effects of hyper-parameters. Red denotes models incorporating fidget features and blue for non-fidget models. In general, models with fidget features perform better. (Error bars are not shown for better visualisation; best performance of each model is in **bold**).

1) Effects of some hyper-parameters: We fix the label smoothing parameter at 0.4. This value is obtained by conducting the classification step using smoothing parameters ranging from 0.0 to 0.6, and 0.4 is chosen as it generally improves performance more than others. We test different numbers of features selected by RF (RF_num), and different GMM kernel sizes.

As shown in Fig. 3, the performance is generally worse when RF_num is low (< 100) as it results in insufficient information. However, when RF_num is high (≥ 250), redundant features bias the classifier, decreasing performance.

Using 32 GMM kernels achieves better performance than 16 kernels. We believe this is due to the way GMM clusters similar per-frame features. More kernels mean more clusters and thus more predictive information. However, when kernel

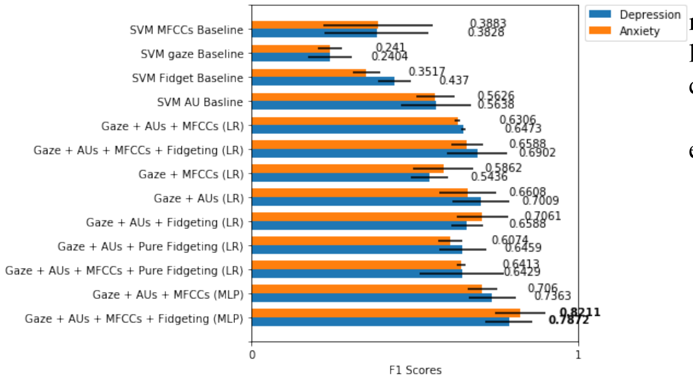

Fig. 4. Effects of feature groups and ablation analysis (error bars extend by the standard deviation in either side; best performance is in **bold**).

size is above 32, the fitting score is large (in GMM lower is better) and therefore increasing beyond 32 will not further improve performance.

2) Effects of feature groups: From Fig. 4, it is clear that fidget features improve most configurations' performance. It is also clear that performance decreases slightly without the participant speaking event. Leading us to conclude that the co-occurrence of speaking and fidgeting is relevant for distress detection.

3) Ablation Analysis: To better understand what is important for distress classification, we remove one or two feature groups from our framework and conduct the same experiments.

Without MFCCs features, the performance generally doesn't drop too much in depression and even increases in anxiety. This may suggest that MFCCs are not very important in depression and even distractive in anxiety detection.

AUs have long been proved to be predictive of distress and, as expected, we see a significant performance reduction when omitting it.

It is interesting to note that fidgeting, with the LR configuration, does not consistently improve performance but in anxiety it always boosts the classification results. Leading us to conclude that fidgeting is certainly important in anxiety, but is also predictive in depression when applying the suitable configuration.

C. Fidget detector cross-dataset validation

To further validate our fidget detector we apply to it a publicly available dataset from Mahmoud *et al.* [3] that has manual fidget labeling.

In this dataset, actors perform specific fidgets. While these fidgets are overemphasized compared to natural fidgets, their core movement is similar.

Segments of the video containing fidgeting are manually labeled in an action-exclusive manner. That is, the co-occurrence of fidgeting is not labeled. Given this, we measure the detector's accuracy in two phases: first we check that fidgeting, regardless of location, is detected during the periods of manually labeled fidgeting; and second, we calculate the recall for location-specific fidgeting. Precision would not make sense for location-specific fidgeting as the detected location may also be fidgeting, while the ground truth only considers another location.

Detected fidgeting segments shorter than 100 frames are excluded to reduce noise. As shown in Table VII, the recall of the non-fidget label is around 50%, but this due to the fact that the labels are generally assigned to a long continuous segment and do not accurately reflect the actions occurring per-frame. However, the recall of the fidget label is good, achieving 80%.

Step 1: Detect fidget only

fidget	precision	recall	f1-score	support
0	0.51	0.49	0.50	29440
1	0.79	0.80	0.80	69517

Step 2: Detect specific fidgeting (evaluated with recall)

Fidget type	Recall	Support
leg	0.784	32430
hand to face	0.865	10594
hand to arm	0.787	12794
hand cross	0.768	13699

TABLE VII
RESULTS OF FIDGET DETECTION ON MAHMOUD *et al.*'S DATASET [3].

We also improve upon Mahmoud *et al.*'s [3] recall for each fidget type, achieving above 75% for all types.

VI. Conclusion

We introduced a novel audio-visual distress dataset comprising recorded interviews and distress labels based on psychology questionnaires.

We then presented an automated fidgeting detection system to extract different fidgeting behaviors from real interview videos. We validated our automated system in a manually-labeled publicly-available fidgeting dataset.

We combined these features with three other modalities, AUs, gaze, and MFCCs, in a multi-modal distress classification pipeline. This pipeline utilized a Multi-modal Deep Denoising Auto-Encoder to compactly represent the modalities per-frame, a GMM to FV step to compactly represent the features across a whole video, and a random forest to select important features. Finally, we tested binary classification of distress labels in LR and MLP classifiers. This pipeline demonstrated the value of fidgeting behaviors in detecting psychological distress.

VII. Future work

As a concept-proof paper, we have demonstrated that fidgeting is useful and important. However, given the limitations of the small dataset, more work is required in utilizing fidget features. Though the recruitment of participants and interviewing is time-consuming and costly, more efforts will be contributed to enlarge the dataset. During the experiment, we treated all fidget features as a whole and validated its effectiveness, but the importance of each fidget behavior (e.g.

hand to arm fidget and hand to hand fidget) requires more exploration.

References

[1] https://www.mentalhealth.org.uk/, 2019, accessed 2019/06/04.

[2] B. De Gelder, "Why bodies? twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3475–3484, 2009.

[3] M. Mahmoud, L.-P. Morency, and P. Robinson, "Automatic multimodal descriptors of rhythmic body movement," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 429–436.

[4] L. A. Fairbanks, M. T. McGuire, and C. J. Harris, "Nonverbal interaction of patients and therapists during psychiatric interviews." *Journal of abnormal psychology*, vol. 91, no. 2, p. 109, 1982.

[5] A. Mehrabian and S. L. Friedman, "An analysis of fidgeting and associated individual differences," *Journal of Personality*, vol. 54, no. 2, pp. 406–429, 1986.

[6] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[7] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[8] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[9] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*. Springer, 2010, pp. 143–156.

[10] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, 1978.

[11] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 53–59.

[12] J. Joshi, R. Goecke, G. Parker, and M. Breakspear, "Can body expressions contribute to automatic depression analysis?" in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–7.

[13] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn, "Multimodal detection of depression in clinical interviews," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 307–310.

[14] 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.

[15] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *American Journal of Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[16] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression severity prediction based on biomarkers of psychomotor retardation," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 37–43.

[17] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency *et al.*, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[18] D. Ebert, R. Albert, G. Hammon, B. Strasser, A. May, and A. Merz, "Eye-blink rates and depression: Is the antidepressant effect of sleep deprivation mediated by the dopamine system?" *Neuropsychopharmacology*, vol. 15, no. 4, pp. 332–339, 1996.

[19] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[20] K. Anis, H. Zakia, D. Mohamed, and C. Jeffrey, "Detecting depression severity by interpretable representations of motion dynamics," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 739–745.

[21] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Analysis of fundamental frequency for near term suicidal risk assessment," in *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, vol. 3. IEEE, 2000, pp. 1853–1858.

[22] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan, "Multi-modality hierarchical recall based on gbdts for bipolar disorder classification," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 31–37.

[23] I. Orton, "Vision Based Body Gesture Meta Features for Affective Computing," Master's thesis, University of Cambridge, 2019.

[24] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 3–13.

[25] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[26] L. Yang, Y. Li, H. Chen, D. Jiang, M. C. Oveneke, and H. Sahli, "Bipolar disorder recognition with histogram features of arousal and body gestures," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 15–21.

[27] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. R. Traum, S. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews." *LREC*, 2014.

[28] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, Apr. 2009.

[29] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a Brief Depression Severity Measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[30] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A Brief Measure for Assessing Generalized Anxiety Disorder," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, May 2006.

[31] B. Gierk, S. Kohlmann, K. Kroenke, L. Spangenberg, M. Zenger, E. Brähler, and B. Löwe, "The Somatic Symptom Scale–8 (SSS-8)," *JAMA Internal Medicine*, vol. 174, no. 3, pp. 399–407, Mar. 2014.

[32] S. Cohen, T. Kamarck, and R. Mermelstein, "Perceived Stress Scale," 1983.

[33] O. P. John and S. Srivastava, "The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives," in *Handbook of personality Theory and research*. t.personality-project.org, 1999, pp. 102–138.

[34] 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.

[35] R. W. Schafer *et al.*, "What is a savitzky-golay filter," *IEEE Signal processing magazine*, vol. 28, no. 4, pp. 111–117, 2011.

[36] https://github.com/taylorlu/Speaker-Diarization, 2019, accessed 2019/09/20.

[37] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[38] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[39] Z. Zhang, W. Lin, M. S. Liu, and M. Mahmoud, "Multimodal deep learning framework for mental disorder recognition," in *2020 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020)*. IEEE, 2020.

[40] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated screening for bipolar disorder from audio/visual modalities," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 39–45.

[41] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 255–259.