Depression Recognition Using Remote Photoplethysmography From Facial Videos

Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López

Abstract—Depression is a mental illness that may be harmful to an individual’s health. The detection of mental health disorders in the early stages and a precise diagnosis are critical to avoid social, physiological, or psychological side effects. This work analyzes physiological signals to observe if different depressive states have a noticeable impact on the blood volume pulse (BVP) and the heart rate variability (HRV) response. Although typically, HRV features are calculated from biosignals obtained with contact-based sensors such as wearables, we propose instead a novel scheme that directly extracts them from facial videos, just based on visual information, removing the need for any contact-based device. Our solution is based on a pipeline that is able to extract complete remote photoplethysmography signals (rPPG) in a fully unsupervised manner. We use these rPPG signals to calculate over 60 statistical, geometrical, and physiological features that are further used to train several machine learning regressors to recognize different levels of depression. Experiments on two benchmark datasets indicate that this approach offers comparable results to other audiovisual modalities based on voice or facial expression, potentially complementing them. In addition, the results achieved for the proposed method show promising and solid performance that outperforms hand-engineered methods and is comparable to deep learning-based approaches.

Index Terms—Affective computing, depression detection, HRV features, image processing, machine learning, rPPG, remote photoplethysmography, signal processing.

I. INTRODUCTION

Major depressive disorder (MDD), also known as clinical depression, is one of the most common mental disorders with increasing prevalence that contributes significantly to the global healthcare burden [1]. Depression can lead to severe consequences for individuals both personally and socially [2], [3]. In addition, several studies suggest long-term and clinically significant depression as a trigger for other serious medical conditions and physiological changes such as cardiovascular disease, diabetes, osteoporosis, aging, pathological cognitive changes, including Alzheimer’s disease and other dementias, and even an increase in the risk of earlier mortality [4], [5].

Currently, depression screening is usually based on medical interviews described in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V), but depends on the subjectivity and experience of the psychiatrist and the subjective memory of the patient, a fact that can lead to misdiagnosis with its consequential social, physiological, or psychological side effects due to undertreatment or overtreatment of the illness.

In recent years, the assessment of depression from facial videos has aroused interest in the scientific community, since the clinical literature has documented particular visual cues and behaviors on faces and facial expressions triggered by major depressive disorder [6]. These facial signs go from reducing facial movements, eyebrow activity, eyes gaze, head pose, mood expressions occurrence, body gestures, or eyelid activity, among others. In addition, this discipline allows the development of a non-invasive and unobtrusive technology and modality that can support the medical diagnosis while the physician focuses exclusively on the patient. The literature studies based on facial visual information have concentrated mainly on three ideas: extracting features from textures and dynamic textures using handcrafted textural descriptors [7], extracting temporal features from the facial geometry and morphology to analyze facial expressions using Facial Action Coding System (FACS) and Action Units (AUs) [8], [9] or facial and head movement dynamics [10], and using deep learning approaches [11], [12], [13], which represent the state-of-the-art methods nowadays.

On the other hand, other objective biomarkers have been shown to be useful for physicians to evaluate and assess the level of depression of the patient in a more confident and precise manner. Recent studies have demonstrated the impact of depression on physiological biomarkers, such as heart rate variability (HRV) calculated from the electrocardiogram (ECG) [14], [15], HRV using photoplethysmography (PPG) signals [16], [17], electrodermal activity (EDA) [18] or acoustic physiological features from the speech [19].

Photoplethysmography (PPG) is a relatively simple and inexpensive optical technique that uses a light source and a photodetector to detect the blood volume changes at the skin surface. PPG is often used to monitor the heart rate (HR) and the blood oxygen saturation (SpO₂) but has been widely used in the scientific literature to estimate different physiological parameters such as Heart Rate Variability (HRV). Recent studies have utilized these PPG-derived parameters to detect...
affective states such as depression or pain [17], [20]. In particular, depression has been clinically found to correlate with parameters on both sympathetic and parasympathetic activity, including autonomic nerve transient responses [16], or the high frequency (HF) and low frequency (LF) components of the HRV [17]. PPG signals can be recorded using contact-based medical-graded devices (i.e., fingertip pulse oximeter) or wearable devices such as smartwatches, fitness trackers, or earphones [21]. The main advantage of this modality is that it is affordable, non-invasive, and portable. Additionally, it provides a more comfortable and less obtrusive user experience than ECG devices.

Remote PPG (rPPG) imaging is a contactless version of this technique that uses a video camera as sensor and ambient light sources [22]. Hence, rPPG can extract physiological signals remotely using only video streams. The technique consists in analyzing the subtle color variations or motion changes in skin regions [22]. Remote PPG has to deal with several challenges such as noise, illumination variations or the person’s movements, but allows for non-invasive, remote and unobtrusive evaluation and monitoring of the users. Hence, the technology offers significant advantages compared to contact-based devices [23], since it has shown comparable results to PPG methods using FDA-approved contact-based pulse oximeters [24]. A few studies have tried to use rPPG signals to assess different affective states such as pain [25] or stress [26] rPPG signals. However, they rely on reference signals for learning or evaluating the quality of the extracted rPPGs and features.

In this work, we aim to analyze the impact of different levels of depression on the physiological response of the blood volume pulse (BVP) signal. In particular, we aim to extract heart-related features from the BVP signal using remote photoplethysmography (rPPG) from facial videos in a fully unsupervised manner, using a non-learning based method that relies mostly on signal processing. Based on this, we propose, for the first time, a novel approach for automatic depression screening using these physiological signals extracted from facial videos and machine learning. Our main contribution can be summarized as follows:

- We assess depression scores by extracting remote photoplethysmographic signals (rPPG), and use them to compute a set of statistical and heart rate variability (HRV) features, including linear and non-linear geometrical parameters from the blood volume pulse (BVP), feeding them to machine learning regressors based on Random Forests and Multilayer Perceptrons.
- To demonstrate the validity of our approach, we evaluate our methods in two publicly available video-based datasets, typically used as a benchmark for depression assessment, AVEC2013 and AVEC2014. The results show that the new approach is feasible and shows more stable inter-video predictions than other modalities.
- To complement our study, we compare our approach with different audiovisual modalities. We prove that the combination of physiological signals with both texture-based and deep features is complementary and improves the results further.

II. PROPOSED METHODOLOGY

In this article, we propose a regression task to determine the level of depression of a person using remote photoplethysmography (rPPG). In this case, we use rPPG signals extracted from faces recorded with a user-graded RGB camera. The regression task comprises several steps: extracting the biosignals from the facial videos, pre-processing the extracted signals to convert them into physiological rPPG signals, extracting features from these rPPG signals, training the models using these features, and evaluating the performance of the models.

In the last decade, rPPG research has advanced significantly from simple signal processing of the raw RGB signals extracted from the video frames to sophisticated multi-step processing pipelines and end-to-end supervised learning methods with dedicated architectures. In general, we can divide the rPPG methods into two main categories: Unsupervised or non-learning-based methods and supervised or learning-based methods. The unsupervised rPPG methods focus on recovering the BVP signal by finding skin areas suitable to extract the raw RGB signals using face detection, tracking, and segmentation techniques. After that, these methods carefully process these raw RGB signals to separate the physiological signals contained in the subtle variations of the skin color from the rest of the information (motion, illumination changes, or facial expressions, among others) by applying filtering and different ways of combining the RGB signals into an rPPG signal. RGB to rPPG conversion methods are based on several ideas such as signal decomposition (PCA, OMIT [27]), chrominance information (Green, CHROM [28], POS [29]), or self similarity (LGI [30]).

Supervised rPPG methods are data-driven methods typically based on Deep Neural Networks (DNN). These methods are in general end-to-end solutions that focus on recovering the BVP signal from faces by learning to mimic the reference signals (BVP signals) captured with fingertip pulse oximeters during the training stage. Some of the well-known deep learning based rPPG methods are based on estimating the HR from sequences (HR-CNN [31]), attention mechanisms (DeepPhys [32]), video enhancement (rPPGNet [33]) transductive learning (Meta-rPPG [34]), and multitasking learning and autoencoders (MSTmaps [35]). In general, these methods represent the state-of-the-art in terms of performance, resulting in highly accurate models. However, there is a risk of overfitting to the training data [24].

A. Remote Photoplethysmographic Signal Extraction

To extract rPPG signals from facial videos, we utilize our unsupervised pipeline called Face2PPG [27]: This unsupervised (non-learning based) method for remote photoplethysmographic (rPPG) imaging is comprised of several steps: face detection and face alignment, skin segmentation, regions of interest (ROIs) selection, extraction of the raw signals from ROIs, filtering of the raw signals, RGB to PPG transformation and spectral analysis, and post-processing to compute different signal parameters such as the heart rate (HR), respiratory rate (RR), blood oxygen saturation (SpO2) or heart rate variability (HRV) [21]. The Face2PPG pipeline includes modules for movement and facial
expression stabilization based on geometric normalization using landmark points and dynamic selection of the facial ROIs that allows discarding those regions that present occlusion, low contrast or generally bad signals when compared with other regions, resulting in robust and accurate results in multiple datasets. An schematic of the pipeline can be seen in Fig. 1.

In particular, our configuration includes the following modules: First, it includes an accurate and robust deep learning-based face detection method based on a Single Shot Multibox Detection network (SSD) [36]. After that, the detected faces are aligned using a deep learning facial landmarks detector named Deep Alignment Network, [37] which gives exceptional performance in terms of accuracy even in challenging conditions [38]. Finally, these landmarks are used in a geometrical skin segmentation and normalization scheme that employs the 85 facial landmark points detected in the face by creating a fixed facial mesh composed of 131 triangles, fixing their coordinates in a normalized frontal pose. The results of the face normalization to extract the biosignals can be seen in Fig. 2. Images pixelated for privacy reasons.

Filtering of the raw signals on the frequency band of interest. The raw signals are then processed using an improved filtering module that includes detrending and bandpass filtering to remove artifacts and clean the raw signal to the frequency band of interest.

Finally, the framework incorporates a module to transform the RGB signals into rPPG signals. For the rPPG extraction we use an RGB to PPG conversion method based on chrominance (CHROM) [28]. This version of the Face2PPG framework has been evaluated extensively across several references databases. Table I shows the performance of the system, while complementary experiments can be seen in our previous work [27]. The evaluation shows that the expected HR error for rPPG signals when compared with reference PPG signals ranges from less than 1 beat per minute for simpler datasets with no movement (UBFC), to around 12 beats per minute for heavily compressed databases (MAHNOB). Although the lack of a reference signal in both depression datasets, makes a quantitative evaluation impossible, based on their video characteristics such as relatively free face movement and reasonable resolution and image quality, we could expect the error to be approximately in the middle of that range.

**B. Feature Extraction**

To train our regression models, we use rPPG signals extracted from visual information to compute 68 features along different windows of each 1-dimensional signal. We used windows of 6 seconds and a fixed sliding window of 0.33 seconds, which is equivalent to 10 video frames for a typical framerate of 30 fps. An example rPPG signal window, is shown in Table I. For each rPPG signal window, the extracted features include 9 statistical features for time-series, 49 heart-related features in time-domain, and landmarks. The right image shows the normalization of each detected face from the AVEC2014 database. The left image of each pair shows face detection and landmarks. The right image shows the normalization of each detected face in the videos. [27].

**Table I**

| Method          | LGI-PPG | COFACE | PURE | MAHNOB | UBFC-1 | UBFC-2 |
|-----------------|---------|--------|------|--------|--------|--------|
| Face2PPG        | 3.9     | 8.8    | 1.2  | 12.6   | 0.8    | 1.5    |

Fig. 1. Unsupervised methodology for remote photoplethysmographic (PPG) imaging using a RGB camera, comprising several steps: 1) Detection and alignment of the face at every frame. 2) Skin segmentation. 3) ROI selection. 4) Extraction of the raw signals from RGB channels at the regions of interest. 5) Filtering of the raw signals on the frequency band of interest. 6) Transformation of the filtered RGB signals to a pulse-type signal. 7) Computation of heart-related features using spectral analysis and post-processing.

Fig. 2. Normalization of the faces to fixed coordinates of two sample videos from the AVEC2014 database. The left image of each pair shows face detection and landmarks. The right image shows the normalization of each detected face in the videos. Images pixelated for privacy reasons.

Fig. 3. An example rPPG signal window extracted from a video included in the AVEC2014 dataset.

that do not meet quality extraction standards due to e.g., no face detected, excessive occlusion, or facial regions with poor SNR. The raw signals are then processed using an improved filtering module that includes detrending and bandpass filtering to remove artifacts and clean the raw signal to the frequency band of interest.
frequency-domain, and non-linear features, extending the 30 features used in our previous related work [39].

In particular, the statistical features, include the mean, min, max, std, dynamic range and four percentiles (10, 25, 75 and 90). The fractal analysis features include the Katz fractal dimension, Higuchi fractal dimension and detrended fluctuation analysis of the entire window, and the mean of the three fractal analysis features computed in sub-windows of 2 seconds of the whole window. The entropy analysis features include permutation entropy, spectral entropy, approximate entropy, sample entropy, Hjorth mobility and complexity and number of zero-crossings of the entire window. The heart and HRV related features include heart rate (HR), breathing rate (BR), interbeat interval (IBI), differences between R-R intervals (pNN20, pNN50), Poincare analysis, frequency domain components (VLF, LF, HF, LP/HF ratio), the standard deviation of NN intervals (SDNN), among others [40]. To compute them, we use the Numpy Python library to compute the statistical features, the Antropy Python package, a software tool for computing the complexity of time-series, to extract both fractal and entropy features [41] and two Python libraries, namely Neurokit2 [42] and HeartPy [43], to computer HRV related features.

For comparative purposes, we have also computed textural features from visual information. We have followed a similar approach to the AVEC2014 baseline [44], which employs the local dynamic appearance descriptor LGBP-TOP, employing fixed temporal windows of 10 consecutive frames. Following the baseline method, the extracted feature vector comprises features extracted only from the XY orthogonal plane. The computation of textural features employs a custom Python script based on the Bob signal processing and machine learning library [45].

C. Regressor Selection

For both physiological and textural features, we select regressors based on Random Forests and Multilayer Perceptrons, as included in Scikit-learn Python library. The Random Forest Regressor (RFR) uses $n_{\text{estimator}} = 550$, $\text{max}\_\text{depth} = 15$ and default values for the rest of the parameters of the model. The Multilayer Perceptron Regressor (MLPR) uses a topology that includes an input layer with the number of input features, three hidden layers, and an output layer with one neuron that corresponds to the regression value of the depression. The configuration used for the training includes: a "relu" (rectified linear unit function) for the activation function in the hidden layers, "Adam" solver for the weight optimization, a $\text{batch}\_\text{size} = 140$ with a learning rate "constant", an initial learning rate of 0.01 and default values for the rest of the parameters.

Again, for comparative purposes, we have implemented an end-to-end deep-learning regression model based on a ResNet-50 convolutional neural network [46], followed by a regression layer composed of two fully connected layers. Based on the literature [12], as input to the network, we have used all individual frames of each video by cropping the input frame to the facial rectangle.

We evaluate the performance of the regression models both individually and combined. First, we train individual models using extracted features from the rPPG physiological signals, and compare them with the performance of regressors based on textural features and the end-to-end regressor based on deep-learning. In addition, we combine these features and models in two different ways. First, using a feature-level fusion approach (pre-fusion) by creating a unique feature vector with features from both textural and physiological modalities, training a model with these feature vectors. Finally, we also use a score-level fusion approach (post-fusion) by combining the result of the inferences from the individual models using the average of the results.

III. EXPERIMENTAL ANALYSIS

A. Datasets and Protocol

To demonstrate the performance of the proposed method, we evaluate the trained models on two publicly available databases, namely the Audio/Visual Emotion Challenge (AVEC) 2013 [47] and 2014 [48]. The experiments were performed on the sets of the Depression Recognition Sub-Challenge (DSC) task, where the goal was to estimate the score of individuals on the Beck Depression Inventory (BDI-II). Both datasets are derived from a subset of the audiovisual depressive language corpus (AViD-Corpus) and they are divided in three partitions: training, development, and test set. Every video includes a label based on questionnaire answers following the Beck Depression Inventory-II (BDI-II) [6], resulting in a depression score of 0 to 63. According to the BDI-II score, the severity of depression can be classified into four levels: minimal (0-13), mild (14-19), moderate (20-28), and severe (29-63).

The AVEC2013 dataset contains 150 videos from 84 subjects, with 50 videos on each partition. However, in the AVEC2014 dataset, the individuals were recorded while performing two different tasks: Freeform and Northwind. The recordings are segmented into three parts in both tasks: training, development, and test set containing 50 videos in each partition for a total of 300 videos. The protocol for AVEC2014 evaluates the models using the two different tasks, both separately and jointly. For the separate task models, models are trained using the subsets of either the Northwind or Freeform tasks, while the joint models, simply combine the data from both tasks both in the training and testing phases.

B. Experimental Setup

We evaluated and analyzed the proposed methodologies to detect the level of depression using features extracted from remote photoplethysmography signals and visual features extracted from video frames from both benchmark data sets. We compare the results across different trained models using these features individually or in a fusion manner and compare them with state-of-the-art for both supervised and unsupervised methods. The experiments are performed using a computer that includes an AMD Ryzen(TM) 3700X 8-core processor at 3.6 GHz, with 64 Gigabytes of RAM, 4 TB SSD and two NVIDIA GeForce RTX(TM) 2080. We have also used the Puhti supercomputer at the IT Center for Science (CSC) in Finland to extract the visual
To further analyze the performance of the Freeform Northwind II [48], we use the following notation to refer to the machine learning algorithms: RFR for Random Forest Regressor and MLPR for a Multilayer Perceptron Regressor.

### Performance Metrics

#### C. Performance Metrics

To evaluate the performance of these models and make a fair comparison with the state-of-the-art methods, we provide the two most common metrics in the automatic depression assessment literature, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The overall predicted depression score for each input video is obtained by averaging the estimation scores for all its windows.

#### D. Experimental Results

In this section, we evaluate the performance and validity of the proposed modality and approach through a series of experiments in the benchmark databases. We compare them with other modalities and state-of-the-art approaches.

1) **Performance in AVEC2013 and AVEC2014:** In Tables II and III, we show the evaluation of the performance of the proposed approach using HRV and BVP features extracted from facial videos for both AVEC2013 and AVEC2014. We compare them with other unimodal methods based on appearance and texture. We observe that the results of the unimodal models corresponding to HRV features and textural features on the AVEC2013 and AVEC2014 test sets have similar performance, although textural features seem to provide slightly better information.

In addition, we also explore a multimodal fusion by combining the heart-related features with textural and deep features to complement the results.

The most remarkable output is that the combination of the features from both textural and physiological modalities, achieves the best results, supporting the hypothesis that both modalities are indeed complementary.

For AVEC2014, we can observe that for the Freeform task the regression models work slightly better than for the Northwind task, as expected according to the baseline results [48]. We can observe that the results of the individual models (using HRV features and textural features individually) when using the data joining both tasks are similar in both datasets.

We show results for individual modalities. We can observe that all modalities show similar results, while the deep learning-based approach (ResNet-50) has slightly better individual results than the models trained with handcrafted features extracted from either textural or rPPG features.

In addition, we show the fusion of HRV features with both textural and deep features. In AVEC14, score-level fusion also results in better performance than feature-level fusion although slightly worse than in AVEC2013. The combination of deep features and rPPG features at score-level shows a further improvement of the results. This proves that, in the same manner as textural and rPPG modalities, deep models provide for information that is also complementary to that extracted from physiological signals. In any case, the best results are obtained when fusing all three data modalities at score level.

2) **Error Analysis:** To further analyze the performance of the rPPG-based features, we display the error distribution in the AVEC2014 benchmark comparing them with the textural-based models, as shown in Fig. 4. The figure shows the mean absolute error for each of the 100 test videos sorted from the smallest to the largest.

In Fig. 4, it can be seen that for rPPG-based models (subfigure A), more than 60% of the videos show an error below a threshold of 15, results that will not result in heavy misclassification. Similar results can be seen for deep ResNet-50 models (subfigure B), while LGBP-TOP models shop up to 71% below the threshold but with a very uneven distribution of errors (subfigure C). The score-fusion model (subfigure D) shows improved results when compared with unimodal models, with 73% of the videos below the threshold, while also keeping a moderately uniform distribution of errors. The error distribution suggests the complementarity of the features and of texture, deep and rPPG based models.

3) **Qualitative Evaluation:** For a qualitative evaluation of the models, we show the different predictions per window for three different example videos, depicted in Fig. 5. We can observe that inference when using rPPG-based features to train the models is relatively stable and shows less variance for the different time windows that make up a single video. This is in contrast with the the inferences obtained from regressors trained with...
Fig. 4. Mean Absolute Error (MAE) distribution of the AVEC2014 Testing Video dataset (Northwind + Freeform). Error distribution of the depression ordered from smallest to largest error per video. From left to right, and top to bottom: A) error distribution when using: rPPG+HRV features + Random Forest regressor, B) ResNet-50 neural network, C) LGBP-TOP features + Multilayer Perceptron regressor and D) Score-fusion level of the models in A, B and C.

Fig. 5. Examples of the predicted depression level per window in two videos from the AVEC2014 test set. In the first row, estimation for the video 245_3 performing the Northwind task, and in the second row, the estimation for the video 319_2 performing the Freeform task. From left to right, estimations of: a) Random Forest regressor using rPPG features, b) ResNet-50 trained with input facial images, and c) Multilayer Perceptron regressor using the visual textural features.

visual textural features, that show high variability in the predictions, although a somehow accurate average. Models trained using deep learning, show a reasonable stability, but worse than HRV.

4) Computational Cost Analysis: We analyze the computational performance of the proposed method to detect depression using rPPG signals. Table IV shows the computational costs of each block that compose the method pipeline in terms of GFLOPs and time consumption per frame. We evaluate each block separately and compare the total cost of the proposed method with state-of-the-art end-to-end deep learning models to detect depression. In addition, we include the cost of the common frame preprocessing methods, namely face detection and alignment.
The measurement is performed using the desktop setup described in Subsection III-B. We used a floating point precision of 32 bits (FP32) and Python 3.8. To measure the computational costs, we used Perf, a profiler tool for Linux 2.6+ based systems that includes hardware level (CPU/PMU, Performance Monitoring Unit) features and software features (software counters, tracepoints).

The total computational cost of the proposed method is 0.091 GFLOPs for the part of pipeline including all processing modules, namely Face Normalization, raw RBG signal extraction and skin segmentation, RGB to BVP transformation, Feature Extraction and Model Inference. The Face Normalization module is the most time-consuming block, mostly due to intensive memory read and write operations.

The time consumed by pre-processing related blocks can vary depending on the face detection and alignment method. However, although they can account for most of the computational cost, they are also included in all end-to-end deep learning-based models. A direct comparison of our method, including rPPG extraction, feature computation and model inference is from 45 to 134 times more efficient when compared with the inference of other end-to-end deep learning models. These results are to be expected since our method focuses on the analysis of one-dimensional signals.

5) Impact of the Window Size: We compare the results of the proposed method using different window sizes to extract HRV features from the rPPG signals and a fixed sliding window of 0.33 seconds (10 video frames). We have carried out this experiment in AVEC2014 using the same Random Forest regressor as in Table III and the data from both tasks included in AVEC2014 (Freeform and Northwind data). We have tested on typical values five different window lengths: 5, 6, 8, 10, and 15 seconds. The summary of the results can be seen in Table V.

The results show that shorter windows that capture short term temporal changes shows a better performance than longer ones, while windows below 6 seconds, start showing problems worse performance due to the lack of sufficient pulse peaks to compute reliable statistics, especially when the subjects have a low heart rate.

6) Cross-Database Analysis: To observe the how rPPG-based models generalize when exposed to additional unseen data, we perform a cross-database analysis using the AVEC2013 and AVEC2014 databases. Although both signals are recorded using a similar setup, the test subset shows different videos. Table VI shows the results of the cross-database experiments for features obtained from visual information. We trained both Random Forest regressor (RFR) and Multilayer Perceptron (MLPR) regressor from rPPG features, using the training protocol suggested in the source dataset, testing the resulting models on the Test subset of the target database.

We can observe that results using the Random Forest regressor (RFR) and the Multilayer Perceptron regressor (MLPR) regressor from rPPG features, using the training protocol suggested in the source dataset, testing the resulting models on the Test subset of the target database.

We can observe that results using the Random Forest regressor (RFR) and the Multilayer Perceptron regressor (MLPR) models with rPPG features show similar behavior, with similar performance as when used in the source datasets (see Tables II and III). We compared them with models trained with LGBP-TOP features from textural information, which show to generalize worse to unseen data, especially when comparing the RMSE error. On the other hand, similar cross-database analysis using a deep features from a 3D-ResNet type architecture [49], have
shown to maintain a similar level of performance. These comparative experiments suggest that the rPPG-based models learn HRV features that are useful when used in other related, but different unseen data.

7) Performance Across Different Machine Learning Regression Models: We explore the performance across different regression models and summarize the results in Table VII. We have trained a set of Machine Learning regressors selected using an exploratory strategy that tried up to 15 different regressors, which we narrowed down to 6 based on their type and preliminary performance. We selected Random Forest regression (RFR) and Extremely Randomized Trees regression (ExTR) from ensemble learning methods, Logistic regression (LogR) and Support Vector Machine regression (SVR) as linear regressors, Stochastic Gradient Descent regression (SGDR) as iterative method and Multilayer Perceptron regression (MLPR) as neural network method. For each model, we have used the default parameters of the machine learning algorithms set by the Scikit-learn Python library, with the exception of an increased number of estimators and maximum depth for the models based on trees.

Similarly to the experiments shown in Tables II and III, we explore the results when training the different models with visual and rPPG features individually, and using two multimodal fusion approaches.

We can observe that in general the Random Forest regressor and the Multilayer Perceptron regressor obtain the best results.

The RFR works especially well when using the features extracted from the rPPG signals. The MLPR works especially well when using the visual features. We hypothesize that in the case of the HRV features, the RFR is able to find nonlinear relationships between the dependent and independent variables whereas the MLPR works better with linear relationships, assuming that the features extracted from dynamic textures of a face have a strong linear dependency. The logistic regressor works well when using the LGBP-TOP features but achieves poor performance when using the HRV features. As expected, extra-trees ensemble regressor has similar performance than the Random Forest, but slightly worst when using rPPG features and slightly better with the LGBP-TOP features, especially for the RMSE metric.

E. Comparison of Features and Sensor Modalities

We have compiled a series of previous works for each modality from baseline to state-of-the-art methods. The primary sensor modalities are based on the typically available sensor modalities such as audio and RGB video, as for AVEC2013 and AVEC2014 database benchmarks. However, the main differences are related to the type of information of interest and the way of computing features from it. Since we introduced a data and feature modality extracted from a remote facial video to regress the level of depression, namely remote physiological features from visual information, we focus on these comparisons. Table VIII shows a comparison of different approaches, sensors, and data modalities to infer depression levels in an unobtrusive manner automatically from audiovisual material. We have identified five types of features extracted from both audio and video sensors.

From the audio sensor, previous works have employed features extracted from:

- Speech signals as an audio time series. We have identified features such as handcrafted speech features (LLDs, MFCCs, statistical features, spectral features, etc.), deep learning features, or the conversion to spectral images to extract deep learning visual features.
- Speech as semantic information. Features such as linguistic and para-linguistic features or emotion recognition features.

From the RGB videos, we have identified in the literature four different data (feature) modalities:

- Geometrical features, mostly associated with motion and morphology of both the image and the facial landmarks. The approaches and methods that use these features focus primarily on translating the temporal information of the landmarks or head pose to images such as spectral heat maps, motion history images or motion maps. But other approaches use temporal and morphological information and facial landmark features, gaze, or Action Units (AU) to regress the level of depression.
- Texture features, mostly associated with the static visual features of only one frame. The approaches and methods use handcrafted visual descriptors such as LPQ or LBP features or deep learning features based on the facial appearance of one frame to infer an instantaneous level of depression from the appearance.

We have explored the performance across different regression models and summarize the results in Table VI. We tested in AVEC2014 using rPPG features and Two Different Regression Models in Cross-Dataset Setting. "TR13→TST14" means that the models are trained in AVEC2013 and tested in AVEC2014.

| Method        | Modality | TR13→TST14 | TR14→TST13 |
|---------------|----------|------------|------------|
| Ours (RFR)    | rPPG     | 7.52       | 9.48       | 7.45       | 9.64       |
| Ours (MLPR)   | rPPG     | 7.07       | 9.94       | 7.90       | 9.98       |
| LGBP-TOP      | Texture  | 9.01       | 12.97      | 8.33       | 10.81      |
| MDN-152 [49]  |          | 6.40       | 8.04       | 6.19       | 7.90       |

"TR14→TST13" means that the models are trained in AVEC2014 and tested in AVEC2013.

E. Comparison of Features and Sensor Modalities

We have compiled a series of previous works for each modality from baseline to state-of-the-art methods. The primary sensor modalities are based on the typically available sensor modalities such as audio and RGB video, as for AVEC2013 and AVEC2014 database benchmarks. However, the main differences are related to the type of information of interest and the way of computing features from it. Since we introduced a data and feature modality extracted from a remote facial video to regress the level of depression, namely remote physiological features from visual information, we focus on these comparisons. Table VIII shows a comparison of different approaches, sensors, and data modalities to infer depression levels in an unobtrusive manner automatically from audiovisual material. We have identified five types of features extracted from both audio and video sensors.

From the audio sensor, previous works have employed features extracted from:

- Speech signals as an audio time series. We have identified features such as handcrafted speech features (LLDs, MFCCs, statistical features, spectral features, etc.), deep learning features, or the conversion to spectral images to extract deep learning visual features.
- Speech as semantic information. Features such as linguistic and para-linguistic features or emotion recognition features.

From the RGB videos, we have identified in the literature four different data (feature) modalities:

- Geometrical features, mostly associated with motion and morphology of both the image and the facial landmarks. The approaches and methods that use these features focus primarily on translating the temporal information of the landmarks or head pose to images such as spectral heat maps, motion history images or motion maps. But other approaches use temporal and morphological information and facial landmark features, gaze, or Action Units (AU) to regress the level of depression.
- Texture features, mostly associated with the static visual features of only one frame. The approaches and methods use handcrafted visual descriptors such as LPQ or LBP features or deep learning features based on the facial appearance of one frame to infer an instantaneous level of depression from the appearance.

We have compiled a series of previous works for each modality from baseline to state-of-the-art methods. The primary sensor modalities are based on the typically available sensor modalities such as audio and RGB video, as for AVEC2013 and AVEC2014 database benchmarks. However, the main differences are related to the type of information of interest and the way of computing features from it. Since we introduced a data and feature modality extracted from a remote facial video to regress the level of depression, namely remote physiological features from visual information, we focus on these comparisons. Table VIII shows a comparison of different approaches, sensors, and data modalities to infer depression levels in an unobtrusive manner automatically from audiovisual material. We have identified five types of features extracted from both audio and video sensors.

From the audio sensor, previous works have employed features extracted from:

- Speech signals as an audio time series. We have identified features such as handcrafted speech features (LLDs, MFCCs, statistical features, spectral features, etc.), deep learning features, or the conversion to spectral images to extract deep learning visual features.
- Speech as semantic information. Features such as linguistic and para-linguistic features or emotion recognition features.

From the RGB videos, we have identified in the literature four different data (feature) modalities:

- Geometrical features, mostly associated with motion and morphology of both the image and the facial landmarks. The approaches and methods that use these features focus primarily on translating the temporal information of the landmarks or head pose to images such as spectral heat maps, motion history images or motion maps. But other approaches use temporal and morphological information and facial landmark features, gaze, or Action Units (AU) to regress the level of depression.
- Texture features, mostly associated with the static visual features of only one frame. The approaches and methods use handcrafted visual descriptors such as LPQ or LBP features or deep learning features based on the facial appearance of one frame to infer an instantaneous level of depression from the appearance.

We have compiled a series of previous works for each modality from baseline to state-of-the-art methods. The primary sensor modalities are based on the typically available sensor modalities such as audio and RGB video, as for AVEC2013 and AVEC2014 database benchmarks. However, the main differences are related to the type of information of interest and the way of computing features from it. Since we introduced a data and feature modality extracted from a remote facial video to regress the level of depression, namely remote physiological features from visual information, we focus on these comparisons. Table VIII shows a comparison of different approaches, sensors, and data modalities to infer depression levels in an unobtrusive manner automatically from audiovisual material. We have identified five types of features extracted from both audio and video sensors.

From the audio sensor, previous works have employed features extracted from:

- Speech signals as an audio time series. We have identified features such as handcrafted speech features (LLDs, MFCCs, statistical features, spectral features, etc.), deep learning features, or the conversion to spectral images to extract deep learning visual features.
- Speech as semantic information. Features such as linguistic and para-linguistic features or emotion recognition features.

From the RGB videos, we have identified in the literature four different data (feature) modalities:

- Geometrical features, mostly associated with motion and morphology of both the image and the facial landmarks. The approaches and methods that use these features focus primarily on translating the temporal information of the landmarks or head pose to images such as spectral heat maps, motion history images or motion maps. But other approaches use temporal and morphological information and facial landmark features, gaze, or Action Units (AU) to regress the level of depression.
- Texture features, mostly associated with the static visual features of only one frame. The approaches and methods use handcrafted visual descriptors such as LPQ or LBP features or deep learning features based on the facial appearance of one frame to infer an instantaneous level of depression from the appearance.
Dynamic texture features include the temporal information based on visual features from a sequence of frames. This is the most explored feature modality since it is known that temporal facial reactions or expressions throw more information about a person’s emotional state. The approaches focused on this modality have explored different features such as handcrafted spatio-temporal visual descriptors (LGBP-TOP, LBQ-TOP), different deep learning architectures that encode temporal information, or low-level deep learning features extracted from sequences of images.

And finally, to the best of our knowledge, we have introduced a new data (feature) modality that can be used on RGB videos. It consists on the extraction of physiological signals (BVP) from faces using the temporal RGB information. We use remote photoplethysmographic waveforms to extract features related to the pulse signal, such as heart rate variability and fractal analysis, which have been shown to have a significant impact on the monitoring and diagnosis of mental health disorders such as depression, stress, or anxiety.

From the comparative results, it can be seen that visual information seems to offer better cues for the assessment of depression than audio information. In particular, deep features that combine both spatial and temporal information offer the best overall performance, while other modalities such as geometrical features, behavioural signals and remote physiological

| Sensor modality | Feature type | Feature Extraction | Year | Method approach | Method | MAE | RMSE | Test dataset |
|-----------------|--------------|--------------------|------|-----------------|--------|-----|------|-------------|
| Audio           | Speech       | Handcrafted        | 2013 | Speech features (Baseline) | Violar et al. [47] | 10.35 | 14.12 | AVEC2013    |
| Audio           | Speech       | Handcrafted        | 2014 | Speech features (Baseline) | Violar et al. [48] | 10.04 | 12.37 | AVEC2014    |
| Audio           | Speech       | Handcrafted        | 2014 | MFFC + LR       | Jan et al. [51]  | 6.07 | 12.28 | AVEC2014    |
| Audio           | Speech       | Deep Learning      | 2020 | Spectrum images + STA network | Niu et al. [52]  | 7.14 | 9.50 | AVEC2013    |
| Audio           | Speech       | Deep Learning      | 2020 | Spectrum images + STA network | Niu et al. [52]  | 7.65 | 9.13 | AVEC2014    |
| Audio           | Speech       | Deep Learning      | 2021 | Speech signal + Spectrum images + ResNet | Dong et al. [53]  | 7.32 | 8.73 | AVEC2013    |
| Audio           | Speech       | Deep Learning      | 2021 | Speech signal + Spectrum images + ResNet | Dong et al. [53]  | 6.80 | 8.82 | AVEC2014    |
| Audio           | Speech       | Deep Learning      | 2021 | Attention: TCN-based (TDCA-Net) | Cai et al. [54]  | 6.90 | 9.22 | AVEC2013    |
| Audio           | Speech       | Deep Learning      | 2021 | Attention: TCN-based (TDCA-Net) | Cai et al. [54]  | 7.08 | 8.90 | AVEC2014    |
| RGB Video       | Geometrical  | Deep Learning      | 2018 | Motion + AlexNet (Landmark Motion History, Motion History Image, Curv Motion History) | Sattan et al. [55] | n/a  | n/a  | AVEC2014    |
| RGB Video       | Geometrical  | Deep Learning      | 2020 | Spectral heatmaps and vectors + CNN + ANN | Zhu et al. [54]  | 6.16 | 8.10 | AVEC2013    |
| RGB Video       | Geometrical  | Deep Learning      | 2020 | Spectral heatmaps and vectors + CNN + ANN | Zhu et al. [54]  | 5.95 | 7.15 | AVEC2014    |
| RGB Video       | Geometrical  | Handcrafted        | 2022 | (Landmark Motion Magnitude, Gaze, Action Units) | Rathi et al. [57]  | n/a  | n/a  | DAI-WOZ     |
| RGB Video       | Texture      | Handcrafted        | 2013 | LFPQ-TOP + c-SVR (Baseline) | Violar et al. [47]  | 10.88 | 13.41 | AVEC2013    |
| RGB Video       | Dynamic Texture | Handcrafted    | 2014 | LGBP-TOP + SVR (Baseline) | Violar et al. [48]  | 6.86 | 10.86 | AVEC2014    |
| RGB Video       | Dynamic Texture | Handcrafted    | 2015 | Facial LBQ-TOP + SVR | Men et al. [56]  | 6.22 | 10.27 | AVEC2014    |
| RGB Video       | Textures     | Deep Learning      | 2017 | Facial Appearance + DCNN | Zhu et al. [54]  | 7.18 | 9.49 | AVEC2013    |
| RGB Video       | Textures     | Deep Learning      | 2017 | Facial Appearance + DCNN | Zhu et al. [54]  | 7.18 | 9.19 | AVEC2013    |
| RGB Video       | Textures     | Deep Learning      | 2019 | Facial + ResNet-50 | Melo et al. [60]  | 6.30 | 8.25 | AVEC2013    |
| RGB Video       | Textures     | Deep Learning      | 2019 | Facial + ResNet-50 | Melo et al. [60]  | 6.15 | 8.23 | AVEC2013    |
| RGB Video       | Dynamic textures | Deep Learning | 2020 | Facial + Two-stream 2DCNN | Melo et al. [62]  | 6.06 | 7.97 | AVEC2013    |
| RGB Video       | Dynamic textures | Deep Learning | 2020 | Facial + Two-stream 2DCNN | Melo et al. [62]  | 6.20 | 7.94 | AVEC2014    |
| RGB Video       | Dynamic textures | Deep Learning | 2021 | Facial 2DCNN features + SVR | Niu et al. [53]  | 6.59 | 8.02 | AVEC2013    |
| RGB Video       | Dynamic textures | Deep Learning | 2021 | Facial 2DCNN features + SVR | Niu et al. [53]  | 6.14 | 7.98 | AVEC2014    |
| RGB Video       | Dynamic textures | Deep Learning | 2022 | Facial = DMSN | Melo et al. [50]  | 6.14 | 7.66 | AVEC2013    |
| RGB Video       | Dynamic textures | Deep Learning | 2022 | Facial = DMSN | Melo et al. [50]  | 5.69 | 7.50 | AVEC2013    |
| RGB Video       | Dynamic textures | Deep Learning | 2021 | Upper body images + CNN AlexNet | Ahmad et al. [11]  | 5.64 | 7.28 | AVEC2013    |
| RGB Video       | Physiological | Handcrafted        | 2022 | rTPG and HRV features + RF | Ours  | 7.54 | 9.75 | AVEC2013    |
| RGB Video       | Physiological | Handcrafted        | 2022 | rTPG and HRV features + RF | Ours  | 7.44 | 9.55 | AVEC2013    |
| Multimodal      | Speech + Textures | Handcrafted       | 2013 | Speech features = LPB + PLS | Ming et al. [61]  | 9.14 | 11.19 | AVEC2013    |
| Multimodal      | Speech + Dynamic textures | Handcrafted | 2014 | Speech features = LGBP-TOP + SVR | Violar et al. [48]  | 7.89 | 9.89 | AVEC2014    |
| Multimodal      | Speech + Dynamic textures | Geometrical + Textures | Handcrafted | Geometrical features + LPQ + k-NN | Kaya et al. [62]  | 7.86 | 9.72 | AVEC2013    |
| Multimodal      | Speech + Dynamic textures | Deep Learning | 2018 | MFFC + VGG-Face features + PLS | Jan et al. [51]  | 6.14 | 7.43 | AVEC2014    |
| Multimodal      | Speech + Dynamic textures | Deep Learning | 2020 | Speech spectrum images + Facial + STA network + EEP | Niu et al. [52]  | 6.14 | 8.16 | AVEC2013    |
| Multimodal      | Speech + Dynamic textures | Physiological | 2020 | Speech spectrum images + Facial + STA network + EEP | Niu et al. [52]  | 5.21 | 7.03 | AVEC2014    |
| Multimodal      | Speech + Dynamic textures | Physiological | 2022 | rTPG features (RF) + LGBP-TOP (MLFR) | Ours  | 6.43 | 8.01 | AVEC2013    |
| Multimodal      | Speech + Dynamic textures | Physiological | 2022 | rTPG features (RF) + LGBP-TOP (MLFR) | Ours  | 6.81 | 8.63 | AVEC2014    |
| Multimodal      | Speech + Dynamic textures | Physiological | 2022 | rTPG features (RF) + LGBP-TOP (MLFR) + ResNet-50 | Ours  | 6.57 | 8.49 | AVEC2014    |

Notation: TCN: Temporal Convolutional Network, SVM: Super Vector Machine, SVR: Super-Vector Regressor, MLP: Multilayer Perceptron, DCNN: Deep Convolutional Neural Network 2DCNN: 2-Dimensional Convolutional Neural Network, STA: Spatio-Temporal Attention, EEP: Eigen Evolution Pooling, LR: Linear Regression, PLS: Partial Least Square Regression, DMSN: Decomposed Multiscale Spatiotemporal Network.
signals (HRV) could offer complementary information, further improving the performance. For audio, deep models also outperform those created using handcrafted features. Overall, the multimodal combination of both audio and video shows the best individual performance.

**F. Comparison With Previous Work**

For modalities based only on visual information, we compare the results of our proposed method against state-of-the-art methods on AVEC2013 and AVEC2014 datasets and show them in Tables IX and X. We can observe that we can divide the previous works into two big groups, those based on hand-engineered representations and deep learning methods. In general, deep learning methods outperform methods that use handcrafted features. However, their black-box nature could result in decreased interpretability, missing cues that show where and when manifestations of depression are seen, something that could make them more useful as tools for medical practitioners.

Tables IX and X show, respectively, the performance of several of these methods on AVEC2013 and AVEC2014, both for (data) monomodal and multimodal approaches. The results of these methods seem to improve when using a multimodal approach with different feature modalities [62] where geometric and texture features are combined. Our proposed method builds on similar ideas, but combines novel physiological features with typical dynamic texture features to exploit mostly the complementary visual and physiological temporal information provided by each subject. The learning based methods mostly rely on exploiting also the temporal information using different different deep learning architectures that search for temporal cues in the stream of frames, potentially exploiting spatio-temporal relationships in the videos that could be indicative of depression.

For AVEC2013, the proposed modality in this study outperforms the hand-engineering "traditional" methods, even as a (data) monomodal approach, resulting on a 7.54 MAE. In addition, it has similar performance than one of the first learning-based method proposed to compute the depression level based in two DCNNs [65]. To show that our proposed modality and method extracts complementary information with other approaches based on visual information, we combined our results with other types of features. When our modality is fused with other textural or deep modalities, our results show results comparable (e.g.) to the state-of-the-art methods evaluated in AVEC2013, demonstrating the complementary of the information of both modalities.

For AVEC2014, our method, using exclusively the HRV features as the data modality, also outperforms traditional methods using handcrafted features from the RGB videos, and is very close to some deep learning-based methods such as Zhu et al. [65]. When we combine the features derived from the rPPG signal with deep or visual texture-based features, we achieve results comparable to the state-of-the-art methods in the detection of depression. The improvement of modality fusion at the score level is worse than when testing in AVEC2013, probably due to a smaller amount of data.

**IV. CONCLUSION**

This paper introduced the extraction of remote biosignals from RGB videos to be used in automatic screening of depression levels from facial videos, a novel visual data modality explored here for the first time. In this context, we have proposed a novel scheme that directly extracts physiological signals in an unsupervised manner, just based on visual information, removing the need for any contact-based device or reference signal. We have directly used these signals to compute physiological features such as blood volume pulse features or heart rate variability parameters, training different machine learning regression models. We evaluated our approach using the AVEC2013 and 2014 benchmark databases. Our results show that our method provides information that can help in the assessment of depression, proving that it can be combined with other visual data modalities to improve the performance further. In our analysis, we have shown

| Methods          | MAE  | RMSE |
|------------------|------|------|
| AVEC2013 Video Baseline [47] | 10.88 | 13.61 |
| MHH + LBP (Meng et al. [61]) | 9.14 | 11.19 |
| LPQ + SVR (Kächele et al. [63]) | 8.97 | 10.82 |
| LPQ-TOP + MFA (Wen et al. [64]) | 8.22 | 10.27 |
| LPQ + Geo (Kaya et al. [62]) | 7.86 | 9.72  |
| Two DCNN (Zhu et al. [65]) | 7.58 | 9.82  |
| C3D (Jazayeri et al. [66]) | 7.37 | 9.28  |
| ResNet-50 (Melo et al. [60]) | 6.30 | 8.25  |
| Four DCNN (Zhou et al. [67]) | 6.20 | 8.28  |
| 3DCNN + SVR (Niu et al. [13]) | 6.19 | 8.02  |
| Two-stream 2DCNN (Melo et al. [12]) | 5.96 | 7.97  |
| Ours (HRV) | 7.54 | 9.75  |
| Ours (HRV + LGBP-TOP) | 6.43 | 8.01  |

| Methods          | MAE  | RMSE |
|------------------|------|------|
| AVEC 2014 Video Baseline [48] | 8.86 | 10.86 |
| MHH + PLS (Jen et al. [68]) | 8.44 | 10.50 |
| LGBP-TOP + LPQ (Kaya et al. [69]) | 8.20 | 10.27 |
| Two DCNN (Zhu et al. [65]) | 7.47 | 9.55  |
| C3D (Jazayeri et al. [66]) | 7.22 | 9.20  |
| VGG + FWHI (Jen et al. [70]) | 6.68 | 8.04  |
| Four DCNN (Zhou et al. [67]) | 6.21 | 8.39  |
| ResNet-50 (Melo et al. [60]) | 6.15 | 8.23  |
| 3DCNN + SVR (Niu et al. [13]) | 6.14 | 7.98  |
| Two-stream 2DCNN (Melo et al. [12]) | 6.20 | 7.94  |
| Ours (HRV) | 7.44 | 9.55  |
| Ours (HRV + LGBP-TOP) | 6.81 | 8.63  |
| Ours (HRV + LGBP-TOP + Deep) | 6.57 | 8.49  |
The authors wish to acknowledge CSC, IT Center for Scientific, Finland, for computational resources.

REFERENCES

[1] A. Ferrari et al., “Global, regional, and national burden of 12 mental disorders in 1990-2019: A systematic analysis for the global burden of disease study 2019,” Lancet Psychiatry, vol. 9, no. 2, pp. 137–150, 2022.

[2] M. Trivedi, “The link between depression and physical symptoms,” Primary Care Companion J. Clin. Psychiatry, vol. 6, pp. 12–16, 2004.

[3] T. Elmer and C. Stadtfeld, “Depressive symptoms are associated with social isolation in face-to-face interaction networks,” Sci. Rep., vol. 10, no. 1, 2020, Art. no. 1444.

[4] B. Penninx, “Depression and cardiovascular disease: Epidemiological evidence on their linking mechanisms,” Neuron., vol. 74, pp. 277–286, 2016.

[5] J. Verhoeven, D. Révéz, J. Lin, O. Wolkwitz, and B. Penninx, “Major depression levels are linked to accelerometry and falls: Results from a large psychiatric cohort study.” Mol. Psychiatry, vol. 19, pp. 895–901, 2013.

[6] A. Pampouchidou et al., “Automatic assessment of depression based on visual cues: A systematic review,” IEEE Trans. Affect. Comput., vol. 10, no. 4, pp. 445–470, Fourth Quarter 2017.

[7] L. He, D. Jiang, and H. Sahli, “Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process filter encoding,” IEEE Trans. Multimedia, vol. 21, no. 6, pp. 1476–1486, Jun. 2019.

[8] M. Gavrilescu and N. Vizireanu, “Predicting depression, anxiety, and stress levels from videos using the facial action coding system,” Sensors, vol. 19, no. 17, 2019, Art. no. 3693. [Online]. Available: https://www.mdpi.com/1424-8220/19/17/3693

[9] A. Darzi et al., “Facial action units and head dynamics in longitudinal interviews reveal OCD and depression severity and disengagement: Results from a large psychiatric cohort study,” Front. Psychiatry, vol. 9, 2018, Art. no. 1444.

[10] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. Cohn, “Multimodal detection of depression in clinical interviews,” in Proc. 2015 ACM Int. Conf. Multimodal Interaction (ICMI’15). Assoc. Comput. Machinery, New York, NY, USA, pp. 307–310, 2015. [Online]. Available: https://doi.org/10.1145/2818346.2817076

[11] D. Ahmad, R. Goecke, and J. Ireland, “CNN depression severity level estimation from upper body vs. face-only images,” in Proc. Int. Conf. Pattern Recognit., 2021, pp. 744–758.

[12] W. Carneiro de Melo, E. Granger, and M. Bordallo Lopez, “Encoding temporal information for automatic depression recognition from facial analysis in the blinking modality,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 1080–1084.

[13] M. Niu, J. Tao, and B. Liu, “Multi-scale and multi-region facial discriminative representation for automatic depression level prediction,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2021, pp. 1325–1329.

[14] R. Hartmann, F. M. Schmidt, C. Sander, and U. Hegerl, “Heart rate variability as indicator of clinical state in depression,” Front. Psychiatry, vol. 9, 2019, Art. no. 735. [Online]. Available: https://www.frontiersin.org/article/10.3389/fspst.2018.00735

[15] S. Byun et al., “Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol,” Comput. Biol. Med., vol. 112, 2019, Art. no. 103381. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519302586

[16] S. Daganpuren et al., “Development and clinical application of a novel autonomic transient response-based screening system for major depressive disorder using a fingertip photoplethysmographic sensor,” Front. Biosci. Biotechnol., vol. 6, 2018, Art. no. 64.

[17] M. Kobayashi, G. Sun, T. Shinba, T. Matsui, and T. Kirimoto, “Development of a mental disorder screening system using support vector machine for classification of heart rate variability measured from single-lead electrocardiography,” in Proc. IEEE Sensors Appl. Symp., 2019, pp. 1–6.

[18] M. Sarchipone et al., “The association between electrodermal activity (eda), depression and suicidal behaviour: A systematic review and narrative synthesis,” BMC Psychiatry, vol. 18, 2018, Art. no. 22.

[19] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. S. Couto, and C. Oliveira, “Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan,” PLoS One, vol. 16, no. 4, pp. 1–20, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0248842

[20] D. Jiang, Y. Chu, J. Cai, Y. Tai, and C. Chuang, “Pain monitoring using heart rate variability and photoplethysmograph-derived parameters by binary logistic regression,” J. Med. Biol. Eng., vol. 41, pp. 669–677, 2021.

[21] T. Tamura, “Current progress of photoplethysmography and SPO2 for health monitoring,” Biomed. Eng., vol. 9, pp. 21–36, 2021.

[22] Y. Sun and N. Thakor, “Photoplethysmography revisited: From contact to noncontact, from point to imaging,” IEEE Trans. Bio-Med. Eng., vol. 63, no. 3, pp. 463–477, Mar. 2016.

[23] F.-T.-Z. Khanam, A. Al-Naji, and J. Chahf, “Remote monitoring of vital signs in diverse non-clinical and clinical scenarios using computer vision systems: A review,” Appl. Sci., vol. 9, no. 20, 2019, Art. no. 4474. [Online]. Available: https://www.mdpi.com/2076-3417/9/20/4474

[24] A. Dasari, S. K. Arul Prakash, L. Jeni, and C. Tucker, “Evaluation of bias in remote photoplethysmography methods,” NPJ Digit. Med., vol. 4, 2021, Art. no. 91.

[25] R. Yang, Z. Guan, Z. Yu, X. Feng, J. Peng, and G. Zhao, “Non-contact pain recognition from video sequences with remote physiological measurements prediction,” 2021, arXiv:2105.08822.

[26] R. M. Sabour, Y. Benezech, P. De Oliveira, J. Chappe, and F. Yang, “UBFC-Phys: A multimodal database for psychophysiological studies of social stress,” IEEE Trans. Affect. Comput., to be published, doi: 10.1109/TAFCC.2021.3056960.

[27] C. Alvarez Casado and M. Bordallo Lopez, “Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces,” 2022, arXiv:2202.04101.

[28] G. de Haan and V. Jeanne, “Robust pulse rate from chrominance-based rPPG,” IEEE Trans. Biomed. Eng., vol. 60, no. 10, pp. 2878–2883, Oct. 2013.

[29] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, “Algorithmic principles of remote ppg,” IEEE Trans. Biomed. Eng., vol. 64, no. 7, pp. 1479–1491, Jul. 2017.

[30] C. S. Pilz, S. Zannaseder, K. Krajewski, and V. Blazek, “Local group variance for heart rate estimation from face videos in the wild,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2018, pp. 1335–13358.

[31] R. Speltik, V. Franc, and J. Matas, “Visual heart rate estimation with convolutional neural network,” in Proc. Brit. Mach. Vis. Conf., 2018, pp. 3–6.

[32] W. V. Chen and D. J. McDuff, “DeepPhys: Video-based physiological measurement using convolutional attention networks,” 2018, arXiv: 1805.07888.

[33] Z. Wu, Y. Peng, X. Li, X. Hong, and G. Zhao, “Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 151–160.

[34] E. Lee, E. Chen, and C.-Y. Lee, “Meta-rPPG: Remote heart rate estimation using a transductive meta-learner,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 387–409.

[35] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, “Video-based remote physiological measurement via cross-verified feature disentangling,” 2020, arXiv:2007.08213.

[36] B. Liu et al., “SSD: Single shot multibox detector,” 2015, arXiv:1512.02325.

[37] M. Kowalski, J. Naujokaitis, and T. Tzimiropoulos, “Deep alignment network: A convolutional neural network for robust face alignment,” 2017, arXiv:1706.01789.

[38] C. Alvarez Casado and M. Bordallo Lopez, “Real-time face alignment: Evaluation methods, training strategies and implementation optimization,” J. Real-Time Image Process., vol. 18, pp. 2259–2267, 2021.

[39] C. Alvarez Casado, P. Paananen, P. Sirttiola, S. Pirttikangas, and M. Bordallo Lopez, “Meditation detection using sensors from wearable devices,” New York, NY, USA: Association for Computing Machinery, 2021, pp. 112–116.
