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, Remote Photoplethysmography, rPPG, Signal Processing.

1 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 (SpO_2) 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 pa-

Submitted on: Reviewed on: Published on:
This research has been supported by the Academy of Finland 6G Flagship program under Grant 346208 and PROFI5 HiDyn programunder Grant 326291
All authors are with the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, 90570 Oulu, Finland. Miguel Bordallo López is also with VTT Technical Research Centre of Finland Ltd, 90571 Oulu, Finland.(e-mail: constantino.alvarezcasado@oulu.fi; manuel.lage@oulu.fi; miguel.bordallo@oulu.fi)
Manuscript received April 19, 2005; revised August 26, 2015.

rameters to detect affective states such as depression or pain [20] [17]. 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 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.

2 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 multitask 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].

2.1 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 such as 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 Figure 1.

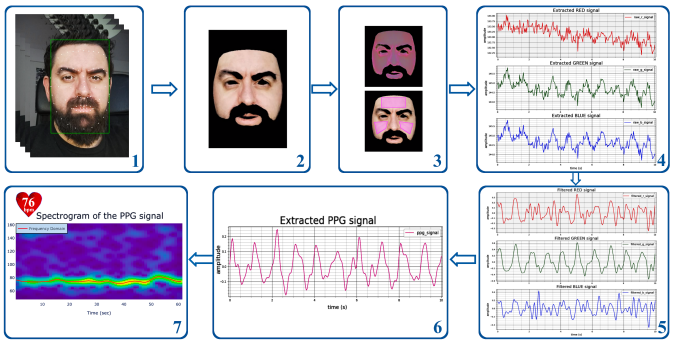

Fig. 1. Unsupervised methodology for remote photoplethysmographic (PPG) imaging using a RGB camera, comprising several steps: 1) Detection and alignment of the face at every frame. 2) Skin segmentation. 3) ROI selection. 4) Extraction of the raw signals from RGB channels at the regions of interest. 5) 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.

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 Figure 2. The normalized face is processed further using a dynamic multi-region selection scheme that extracts raw RGB signals from the best facial areas.

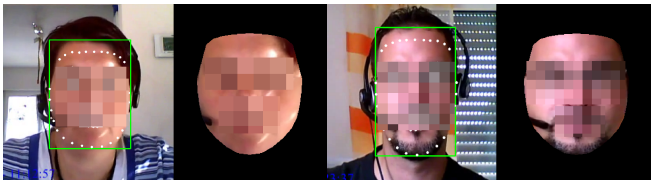

Fig. 2. Normalization of the faces to fixed coordinates of two sample videos from the AVEC2014 database. The left image of each pair shows face detection and landmarks. The right image shows the normalization of each detected face in the videos. [27]. Images pixelated for privacy reasons.

This selection is based on several signal statistics such the mean, standard deviation, variance, signal-to-noise ratio (SNR), Katz Fractal dimension (KFD), number of zero-crossings (Zc), sample entropy, detrended fluctuation analysis (DFA) and the energy in terms of local power spectral density (PSD). In addition, this step allows to discard those parts of the signals 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.

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 1 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.

TABLE 1
Performance of the selected rPPG extraction method, evaluated using the mean average error (MAE) of the heart rate, in beats per minute.

Method	Databases					
	LGI-PPGI	COHFACE	PURE	MAHNOB	UBFC-1	UBFC-2
Face2PPG	3.9	8.8	1.2	12.6	0.8	1.5

2.2 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 Figure 3. For each rPPG signal window, the extracted features include 9 statistical features for time-series, 6 fractal analysis features, 6 entropy analysis features, and 49 heart-related features in time-domain, 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

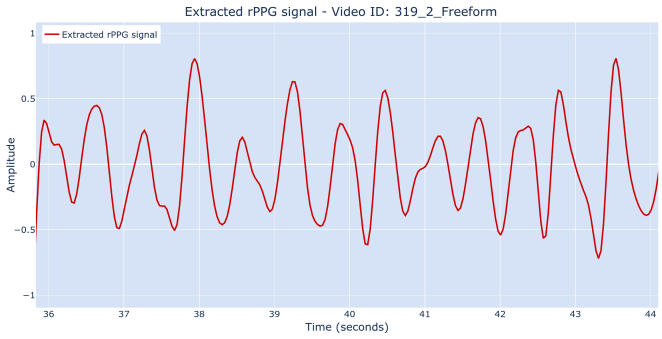

Fig. 3. An example rPPG signal window extracted from a video included in the AVEC2014 dataset.

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, LF/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].

2.3 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 $n\_estimator = 550$, $max\_depth = 15$ 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 $batch\_size = 140$ 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.

3 EXPERIMENTAL ANALYSIS

3.1 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 audio-visual 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.

3.2 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 3.6GHz, with 64 Gigabytes of RAM, 4 terabyte 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 texture-based features. We used Python 3.8 as the programming language.

3.3 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.

3.4 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.

3.4.1 Performance in AVEC2013 and AVEC2014

In Table 2 and Table 3, 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.

TABLE 2
Performance of the proposed method and models for depression recognition on AVEC2013, measured in mean absolute error (MAE) and root mean square error (RMSE). We use the following notation to refer to the machine learning algorithms: RFR for Random Forest Regressor and MLPR for a Multilayer Perceptron Regressor.

Modality	Features	Fusion	Model	MAE	RMSE
Unimodal	rPPG	-	RFR	7.97	9.98
	rPPG	-	MLPR	7.54	9.75
	Textural	-	MLPR	7.26	8.99
Multimodal	rPPG + Textural	Pre	MLPR	6.98	9.02
	rPPG + Textural	Post	RFR + MLPR	7.03	8.97
	rPPG + Textural	Post	MLPR + MLPR	**6.43**	**8.01**

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.

TABLE 3
Performances of the proposed methods for depression recognition considering single task and fusion of tasks on AVEC2014. Performance is measured in mean absolute error (MAE) and root mean square error (RMSE). Notation: RFR: Random Forest Regressor, MLPR: MultiLayer Perceptron Regressor, CNN: Convolutional Neural Network.

Modality	Features	Fusion	Model	Task	MAE	RMSE
Unimodal	rPPG	-	RFR	Freeform	7.74	9.68
	Textural	-	MLPR	Freeform	7.43	9.33
Multimodal	rPPG + Textural	Pre	RFR	Freeform	8.03	9.84
	rPPG + Textural	Post	RFR + MLPR	Freeform	7.37	8.72
Unimodal	rPPG	-	RFR	Northwind	8.28	10.76
	Textural	-	MLPR	Northwind	8.17	10.40
Multimodal	rPPG + Textural	Pre	RFR	Northwind	7.21	8.99
	rPPG + Textural	Post	RFR + MLPR	Northwind	7.62	9.64
Unimodal	rPPG	-	RFR	Joint-tasks	7.44	9.55
	Textural	-	MLPR	Joint-tasks	7.02	9.08
	Deep	-	CNN	Joint-tasks	6.83	9.06
Multimodal	rPPG + Textural	Pre	RFR	Joint-tasks	7.20	9.03
	rPPG + Textural	Post	RFR + MLPR	Joint-tasks	6.81	8.63
	rPPG + Deep	Post	RFR + CNN	Joint-tasks	6.90	8.88
	All	Post	All	Joint-tasks	**6.57**	**8.49**

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.

3.4.2 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 texture-based models, as shown in Figure 4. The figure shows the mean absolute error for each of the 100 test videos sorted from the smallest to the largest.

In Figure 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 missclassification. 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.

3.4.3 Qualitative evaluation

For a qualitative evaluation of the models, we show the different predictions per window for three different example videos, depicted in Figure 5. We can observe that inference

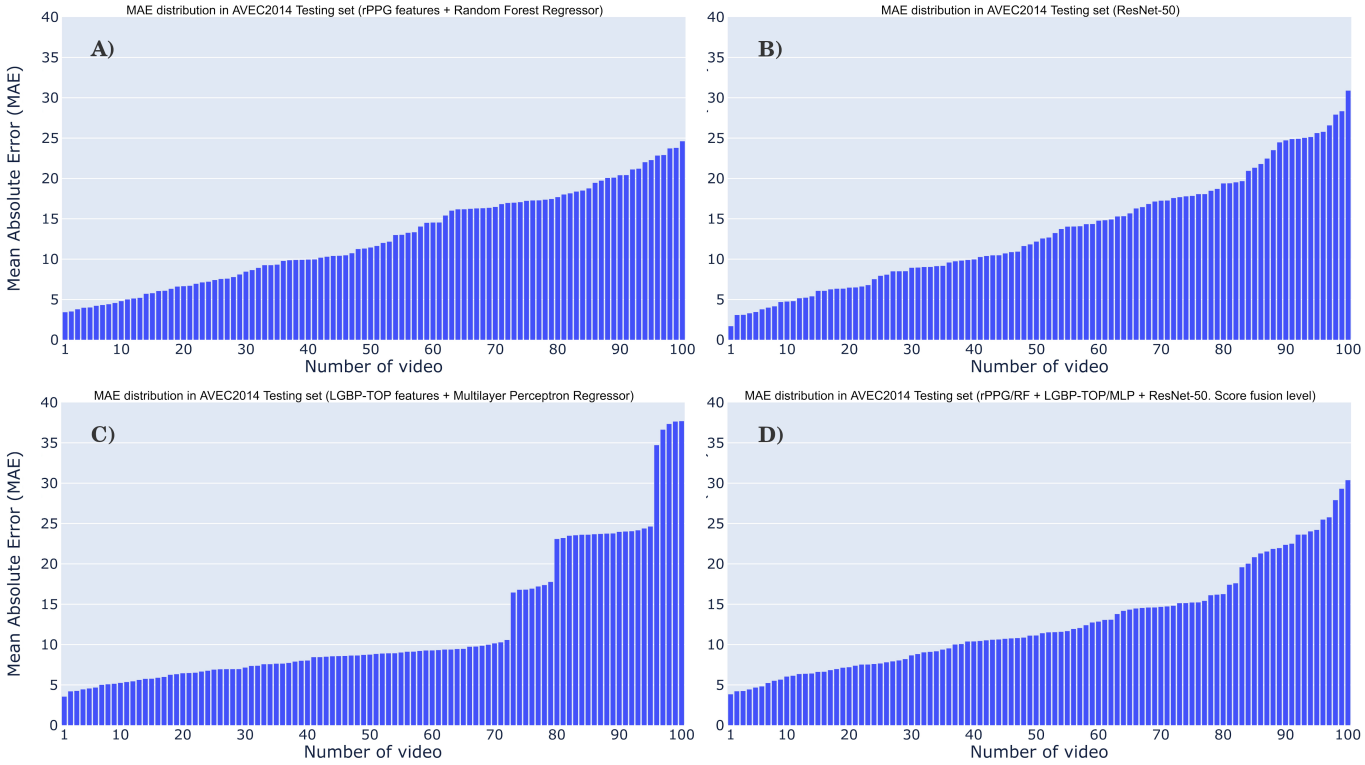

Fig. 4. Mean Absolute Error (MAE) distribution of the AVEC2014 Testing Video dataset (Northwind + Freeform). Error distribution of the depression ordered from smallest to largest error per video. From left to right, and top to bottom: A) error distribution when using: rPPG+HRV features + Random Forest regressor, B) ResNet-50 neural network, C) LGBP-TOP features + Multilayer Percepton regressor and D) Score-fusion level of the models in A, B and C.

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 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.

3.4.4 Computational Cost Analysis

We analyze the computational performance of the proposed method to detect depression using rPPG signals. Table 4 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 3.2. 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.

3.4.5 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 3 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 5. 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

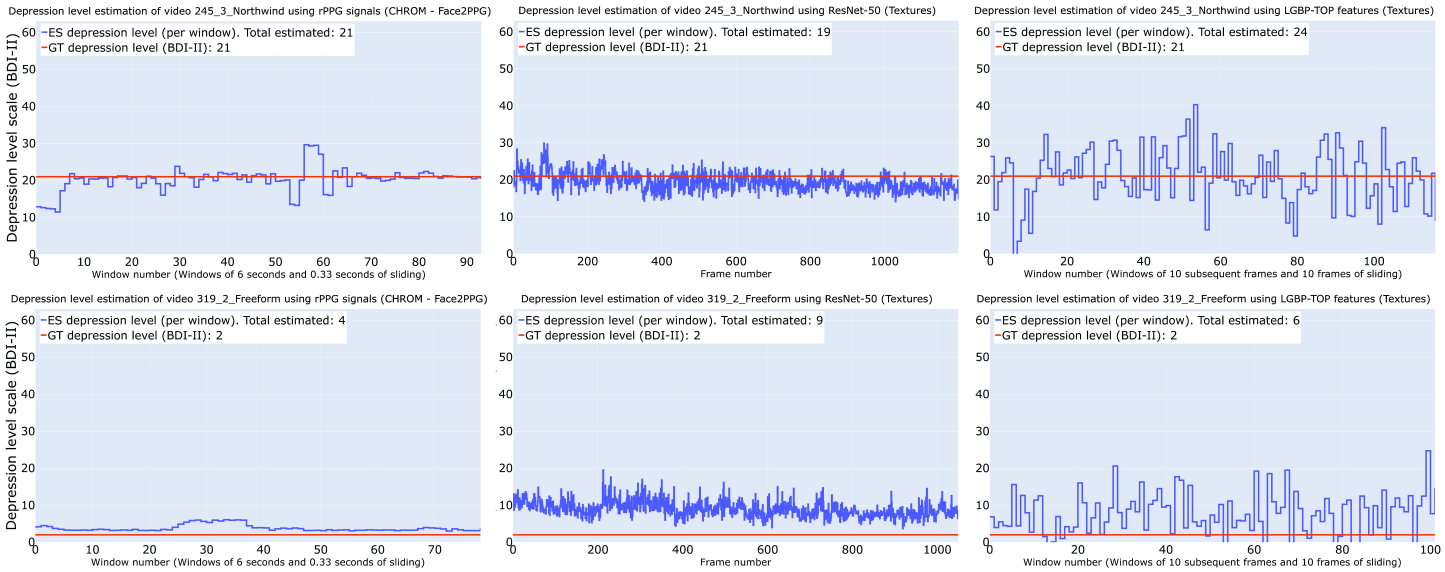

Fig. 5. Examples of the predicted depression level per window in two videos from the AVEC2014 test set. In the first row, estimation for the video *245_3* performing the *Northwind* task,and in the second row, the estimation for the video *319_2* performing the *Freeform* task. From left to right, estimations 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.

TABLE 4
Computational cost analysis comparison. Pre-processing blocks common to all methods (orange), state-of-the-art CNN-based methods (black) and main blocks of our proposed method (blue)

Block name	Block type	Method	GFLOPs	Time (ms.)	Processor	Resolution
Preproc1	Face Detection	OpenCV DNN	0.3173	12.18	CPU + GPU	300x300
Preproc2	Face Alignment	DAN	3.3581	121.77	CPU	112x112
CNN1	Inference	2D-ResNet50	4.13	28.37	GPU	224x224
CNN2	Inference	3D-ResNet50 [49]	12.22	91.45	GPU	224x224
CNN3	Inference	MDN [49]	7.40	55.53	GPU	224x224
CNN4	Inference	DMSN [50]	11.29	84.62	GPU	112x112
B1	Face Normalization	Face2PPG	0.0584	15.746	CPU	640x480
B2	RGB signal extraction	Face2PPG	0.0003	1.413	CPU	640x480
B3	RGB to BVP transform	CHROM	0.0023	0.001	CPU	640x480
B4	Feature Extraction	HRV features	0.0052	0.125	CPU	640x480
B5	Model Inference	Random Forest	0.0246	0.299	CPU	640x480
B1+B2+B3+B4+B5	rPPG-blocks	ours	**0.091**	**17.58**	CPU	640x480

pulse peaks to compute reliable statistics, especially when the subjects have a low heart rate.

TABLE 5
Performance of the regression model in AVEC2014 trained with HRV features extracted from the rPPG signal using different window sizes (in seconds) and a fixed sliding window of 0.33 seconds (10 video frames).

	Window Size				
Metrics	$w = 5$	$w = 6$	$w = 8$	$w = 10$	$w = 15$
RMSE	10.54	9.55	10.24	9.94	10.29
MAE	8.36	7.44	8.13	7.92	8.47

3.4.6 *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 6 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) models with rPPG features show similar behav-

TABLE 6
Performance of different methods, including the proposed method 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. "TR14→TST13" means that the models are trained in AVEC2014 and tested in AVEC2013

Method	Modality	TR13→TST14		TR14→TST13	
		MAE	RMSE	MAE	RMSE
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]	Deep	6.40	8.04	6.19	7.90

TABLE 7
Performance of different regression models in AVEC2014 using rPPG features, visual features (LGBP-TOP) features and the fusion of both at feature-level (pre-fusion) or score-level(post-fusion).

Metric	Features	regression model					
		RF	ExTR	LogR	SVR	SDGR	MLPR
MAE	rPPG	**7.44**	8.18	10.69	8.91	9.15	7.94
	Visual	8.15	7.99	7.85	7.92	8.21	**7.02**
	rPPG + Visual (Post)	7.66	7.89	8.42	8.17	8.41	**7.09**
	rPPG + Visual (Pre)	**7.20**	7.52	8.54	7.98	8.56	7.44
RMSE	rPPG	**9.55**	9.97	14.71	11.12	11.24	10.36
	Visual	9.96	9.61	10.71	10.08	10.37	**9.08**
	rPPG + Visual (Post)	9.57	9.55	11.45	10.26	10.27	**8.83**
	rPPG + Visual (Pre)	**9.02**	9.27	10.71	9.89	10.61	9.11

ior, with similar performance as when used in the source datasets (see Tables 2 and 3). 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.

3.4.7 Performance across different machine learning regression models

We explore the performance across different regression models and summarize the results in Table 7. 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 (SDGR) 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 Table 2 and Table 3, 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.

3.5 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 8 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

TABLE 8
Comparison of different sensor and data (feature) modalities for depresion estimation from audiovisual data. 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.

Sensor modality	Feature type	Feature Extraction	Year	Method approach	Method	MAE	RMSE	Test dataset
Audio	Speech	Handcrafted	2013	Speech features (Baseline)	Valstar et al. [47]	10.35	14.12	AVEC2013
Audio	Speech	Handcrafted	2014	Speech features (Baseline)	Valstar et al. [48]	10.04	12.57	AVEC2014
Audio	Speech	Handcrafted	2018	MFCC + LR	Jan *et al.* [51]	8.07	10.28	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 (Landmarks Motion History, Motion History Image, Gabor Motion History)	S'adan et al. [55]	n/a	n/a	AVEC2014
RGB Video	Geometrical	Deep Learning	2020	Spectral heatmaps and vectors + CNN + ANN	Zhu et al. [56]	6.16	8.10	AVEC2013
RGB Video	Geometrical	Deep Learning	2020	Spectral heatmaps and vectors + CNN + ANN	Zhu et al. [56]	5.95	7.15	AVEC2014
RGB Video	Geometrical	Handcrafted	2022	Facial landmarks motion + SVM (Landmarks Motion Magniture, Gaze, Action Units)	Rathi et al. [57]	n/a	n/a	DAIC-WOZ
RGB Video	Texture	Handcrafted	2013	LPQ-TOP + ϵ-SVR (Baseline)	Valstar et al. [47]	10.88	13.61	AVEC2013
RGB Video	Dynamic texture	Handcrafted	2014	LGBP-TOP + SVR (Baseline)	Valstar et al. [48]	8.86	10.86	AVEC2014
RGB Video	Dynamic texture	Handcrafted	2015	Facial LBQ-TOP + SVR	Wen et al. [58]	8.22	10.27	AVEC2013
RGB Video	Textures	Deep Learning	2017	Facial Apparence + DCNN	Zhu et al. [59]	7.88	10.19	AVEC2013
RGB Video	Textures	Deep Learning	2017	Facial Apparence + DCNN	Zhu et al. [59]	7.82	10.36	AVEC2014
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	AVEC2014
RGB Video	Dynamic Textures	Deep Learning	2020	Facial + Two-stream 2DCNN	Melo et al. [12]	5.96	7.97	AVEC2013
RGB Video	Dynamic Textures	Deep Learning	2020	Facial + Two-stream 2DCNN	Melo et al. [12]	6.20	7.94	AVEC2014
RGB Video	Dynamic texture	Deep Learning	2021	Facial 3DCNN features + SVR	Niu et al. [13]	6.19	8.02	AVEC2013
RGB Video	Dynamic texture	Deep Learning	2021	Facial 3DCNN features + SVR	Niu et al. [13]	6.14	7.98	AVEC2014
RGB Video	Dynamic texture	Deep Learning	2022	Facial + DMSN	Melo et al. [50]	6.14	7.66	AVEC2013
RGB Video	Dynamic texture	Deep Learning	2022	Facial + DMSN	Melo et al. [50]	5.69	7.50	AVEC2014
RGB Video	Dynamic texture	Deep Learning	2021	Upper body images + CNN AlexNet	Ahmad et al. [11]	5.64	7.28	AVEC2013
RGB Video	Physiological	Handcrafted	2022	rPPG and HRV features + RF	Ours	7.54	9.75	AVEC2013
RGB Video	Physiological	Handcrafted	2022	rPPG and HRV features + RF	Ours	7.44	9.55	AVEC2014
Multimodal	Speech + Textures	Handcrafted	2013	Speech features + LBP + PLS	Meng et al. [61]	9.14	11.19	AVEC2013
Multimodal	Speech + Dynamic textures	Handcrafted	2014	Speech features + LGBP-TOP + SVR	Valstar et al. [48]	7.89	9.89	AVEC2014
Multimodal	Geometrical + Textures	Handcrafted	2014	Geometrical features + LPQ + k-NN	Kaya et al. [62]	7.86	9.72	AVEC2013
Multimodal	Speech + Dynamic textures	Deep Learning + Handcrafted	2018	MFCC + 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	Deep Learning	2020	Speech spectrum images + Facial + STA network + EEP	Niu et al. [52]	5.21	7.03	AVEC2014
Multimodal	Physiological + Dynamic textures	Handcrafted	2022	rPPG features (RFR) + LGBP-TOP (MLPR)	Ours	6.43	8.01	AVEC2013
Multimodal	Physiological + Dynamic textures	Handcrafted	2022	rPPG features (RFR) + LGBP-TOP (MLPR)	Ours	6.81	8.63	AVEC2014
Multimodal	Physiological + Dynamic textures	Deep Learning + Handcrafted	2022	rPPG features (RFR) + LGBP-TOP(MLPR) + ResNet-50	Ours	6.57	8.49	AVEC2014

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 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.

3.6 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 Table 9 and Table 10. 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 9 and 10 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.

TABLE 9
Comparison of methods for predicting the level of depression on the AVEC2013 dataset.

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 (Jazaery *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

TABLE 10
Comparison of methods for predicting the level of depression on the AVEC2014 dataset.

Methods	MAE	RMSE
AVEC 2014 Video Baseline [48]	8.86	10.86
MHH + PLS (Jan *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 (Jazaery *et al.* [66])	7.22	9.20
VGG + FDHH (Jan *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

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.

4 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 graphical examples that suggest that the inference of the models trained with this type of feature modality is slightly more stable than those of other models, such as those that exploit textural or deep features. Extensive experiments indicated the usefulness of such modality, when compared to different methods present in the literature. Future work should explore the extraction of all kinds of visual information including textural, spatiotemporal, and rPPGs in a unified framework.

Acknowledgments

This research has been supported by the Academy of Finland 6G Flagship program under Grant 346208 and PROFI5 HiDyn under Grant 326291. The authors wish to acknowledge CSC, IT Center for Scientific, Finland, for computational resources.

References

[1] A. Ferrari, D. Santomauro, A. Herrera, J. Shadid, C. Ashbaugh, H. Erskine, F. Charlson, L. Degenhardt, J. Scott, J. McGrath, P. Allebeck, C. Benjet, N. Breitborde, T. Brugha, X. Dai, L. Dandona, R. Dandona, F. Fischer, J. Haagsma, and H. Whiteford, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet Psychiatry*, 01 2022.

[2] M. Trivedi, "The link between depression and physical symptoms," *Primary care companion to the Journal of clinical psychiatry*, vol. 6, pp. 12–6, 02 2004.

[3] T. Elmer and C. Stadtfeld, "Depressive symptoms are associated with social isolation in face-to-face interaction networks," *Scientific Reports*, vol. 10, 01 2020.

[4] B. Penninx, "Depression and cardiovascular disease: Epidemiological evidence on their linking mechanisms," *Neuroscience and Biobehavioral Reviews*, vol. 74, 07 2016.

[5] J. Verhoeven, D. Révész, J. Lin, O. Wolkowitz, and B. Penninx, "Major depressive disorder and accelerated cellular aging: Results from a large psychiatric cohort study," *Molecular psychiatry*, vol. 19, 11 2013.

[6] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis, "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Trans. on Affective Computing*, pp. 1–27, 2017.

[7] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1476–1486, June 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. [Online]. Available: https://www.mdpi.com/1424-8220/19/17/3693

[9] A. Darzi, N. R. Provenza, L. A. Jeni, D. A. Borton, S. A. Sheth, W. K. Goodman, and J. F. Cohn, "Facial action units and head dynamics in longitudinal interviews reveal ocd and depression severity and dbs energy," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, Dec 2021, pp. 1–6.

[10] H. Dibeklioglu, Z. Hammal, Y. Yang, and J. Cohn, "Multimodal detection of depression in clinical interviews," 11 2015.

[11] D. Ahmad, R. Goecke, and J. Ireland, *CNN Depression Severity Level Estimation from Upper Body vs. Face-Only Images*, 02 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 *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 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," *Frontiers in Psychiatry*, vol. 9, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyt.2018.00735

[15] S. Byun, A. Y. Kim, E. H. Jang, S. Kim, K. W. Choi, H. Y. Yu, and H. J. Jeon, "Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol," *Computers in Biology and Medicine*, vol. 112, p. 103381, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482519302586

[16] S. Dagdanpurev, G. Sun, T. Shinba, M. Kobayashi, N. Kariya, L. Choimaa, S. Batsuuri, S. Kim, S. Suzuki, and T. Matsui, "Development and clinical application of a novel autonomic transient response-based screening system for major depressive disorder using a fingertip photoplethysmographic sensor," *Frontiers in Bioengineering and Biotechnology*, vol. 6, p. 64, 05 2018.

[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 *2019 IEEE Sensors Applications Symposium (SAS)*, March 2019, pp. 1–6.

[18] M. Sarchiapone, c. m. Gramaglia, M. Iosue, V. Carli, L. Mandelli, A. Serretti, D. Marangon, and P. Zeppegno, "The association between electrodermal activity (eda), depression and suicidal behaviour: A systematic review and narrative synthesis," *BMC Psychiatry*, vol. 18, 01 2018.

[19] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-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, 04 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0248842

[20] D. Jhang, Y. Chu, J. Cai, Y. Tai, and C. Chuang, "Pain monitoring using heart rate variability and photoplethysmograph-derived parameters by binary logistic regression," *Journal of Medical and Biological Engineering*, 09 2021.

[21] T. Tamura, "Current progress of photoplethysmography and spo2

for health monitoring," *Biomedical Engineering Letters*, vol. 9, 02 2019.

[22] Y. Sun and N. Thakor, "Photoplethysmography revisited: From contact to noncontact, from point to imaging," *IEEE transactions on bio-medical engineering*, vol. 63, 09 2015.

[23] F.-T.-Z. Khanam, A. Al-Naji, and J. Chahl, "Remote monitoring of vital signs in diverse non-clinical and clinical scenarios using computer vision systems: A review," *Applied Sciences*, vol. 9, no. 20, 2019. [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 biases in remote photoplethysmography methods," *npj Digital Medicine*, vol. 4, 12 2021.

[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," *arXiv preprint arXiv:2105.08822*, 2021.

[26] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, 2021.

[27] C. Álvarez Casado and M. Bordallo López, "Face2ppg: An unsupervised pipeline for blood volume pulse extraction from faces," 2022. [Online]. Available: https://arxiv.org/abs/2202.04101

[28] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, Oct 2013.

[29] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, July 2017.

[30] C. S. Pilz, S. Zaunseder, J. Krajewski, and V. Blazek, "Local group invariance for heart rate estimation from face videos in the wild," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2018, pp. 1335–13 358.

[31] R. Špetlík, V. Franc, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proceedings of the british machine vision conference, Newcastle, UK*, 2018, pp. 3–6.

[32] W. V. Chen and D. J. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," *ArXiv*, vol. abs/1805.07888, 2018.

[33] Z. Yu, W. 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 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 *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 392–409.

[35] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling," *CoRR*, vol. abs/2007.08213, 2020. [Online]. Available: https://arxiv.org/abs/2007.08213

[36] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[37] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," *CoRR*, vol. abs/1706.01789, 2017. [Online]. Available: http://arxiv.org/abs/1706.01789

[38] C. Álvarez Casado and M. Bordallo López, "Real-time face alignment: evaluation methods, training strategies and implementation optimization," *Journal of Real-Time Image Processing*, pp. 1–29, 2021.

[39] C. Álvarez Casado, P. Paananen, P. Siirtola, S. Pirttikangas, and M. Bordallo López, *Meditation Detection Using Sensors from Wearable Devices*. New York, NY, USA: Association for Computing Machinery, 2021, p. 112–116. [Online]. Available: https://doi.org/10.1145/3460418.3479318

[40] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpubh.2017.00258

[41] R. Vallat and M. Walker, "An open-source, high-performance tool for automated sleep staging," *eLife*, vol. 10, 10 2021.

[42] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021. [Online]. Available: https://doi.org/10.3758%2Fs13428-020-01516-y

[43] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "Heartpy: A novel heart rate algorithm for the analysis of noisy signals," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 66, pp. 368–378, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369847818306740

[44] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014 - 3d dimensional affect and depression recognition challenge," *AVEC 2014 - Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Workshop of MM 2014*, pp. 3–10, 11 2014.

[45] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel, "Continuously reproducing toolchains in pattern recognition and machine learning experiments," in *International Conference on Machine Learning (ICML)*, Aug. 2017. [Online]. Available: http://publications.idiap.ch/downloads/papers/2017/Anjos_ICML2017-2_2017.pdf

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR 2016*, 2016, pp. 770–778.

[47] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013 - the continuous audio/visual emotion and depression recognition challenge," 10 2013, pp. 3–10.

[48] M. V. *et al.*, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *AVEC 2014*, 2014, pp. 3–10.

[49] W. Carneiro de Melo, E. Granger, and M. Bordallo Lopez, "Mdn: A deep maximization-differentiation network for spatio-temporal depression detection," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[50] W. C. de Melo, E. Granger, and M. B. Lopez, "Facial expression analysis using decomposed multiscale spatiotemporal networks," 2022. [Online]. Available: https://arxiv.org/abs/2203.11111

[51] A. Jan, H. Meng, Y. F. Abdul Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. PP, pp. 1–1, 07 2017.

[52] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 10 2020.

[53] Y. Dong and X. Yang, "A hierarchical depression detection model based on vocal and emotional cues," *Neurocomputing*, vol. 441, pp. 279–290, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221002654

[54] C. Cai, M. Niu, B. Liu, J. Tao, and X. Liu, "Tdca-net: Time-domain channel attention network for depression detection," 08 2021, pp. 2511–2515.

[55] M. S'adan, A. Pampouchidou, and F. Meriaudeau, "Deep learning techniques for depression assessment," 08 2018.

[56] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[57] S. Rathi, B. Kaur, and R. Agrawal, "Selection of relevant visual feature sets for enhanced depression detection using incremental linear discriminant analysis," *Multimedia Tools and Applications*, 03 2022.

[58] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1432–1441, July 2015.

[59] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 01 2017.

[60] W. C. de Melo, E. Granger, and A. Hadid, "Depression detection based on deep distribution learning," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 4544–4548.

[61] H. M. *et al.*, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *AVEC 2013*, 2013, pp. 21–30.

[62] H. Kaya and A. A. Salah, "Eyes whisper depression: A cca based multimodal approach," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 961–964. [Online]. Available: https://doi.org/10.1145/2647868.2654978

[63] M. K. *et al.*, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," in *ICPRAM 2014*, 2014, pp. 671–678.
[64] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. on Information Forensics and Security*, vol. 10, pp. 1432–1441, 2015.
[65] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Trans. on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2018.
[66] M. Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Trans. on Affective Computing*, pp. 1–8, 2018.
[67] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Trans. on Affective Computing*, pp. 1–12, 2018.
[68] A. Jan, H. Meng, Y. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *AVEC 2014*, 2014, pp. 73–80.
[69] H. Kaya, F. Çilli, and A. Salah, "Ensemble cca for continuous emotion prediction," in *AVEC 2014*, 2014, pp. 19–26.
[70] A. Jan, H. Meng, Y. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. on Cognitive and Developmental Systems*, vol. 10, pp. 668–680, 2018.

Constantino Álvarez Casado is a doctoral researcher at the Center for Machine Vision and Signal Analysis (CMVS) at the University of Oulu. His doctoral research is focused in embedded Artificial Intelligence and Machine Vision, especially in the implementation of real time algorithms and models that are energy efficient and suitable to be integrated in small devices. He received his M.Sc. degree in computer science from University of Oulu (Finland). He has several years of industrial experience in the development of real-time embedded computer vision algorithms, especially for face analysis, both in Spain and Finland. Contact him at constantino.alvarezcasado@oulu.fi

Manuel Lage Cañellas is a PhD candidate at Center for Machine Vision and Signal Analysis (CMVS) at the University of Oulu. His doctoral research is focused in multimodal representation learning. He obtained is Computer Science Engineering degree in Universidad Autónoma de Madrid (Spain) in 2010 and started his career as a software engineer in the Air Traffic Management industry. He worked as project engineer and team leader for the automotive and aerospacial industry. Contact him at manuel.lage@oulu.fi

Miguel Bordallo Lopez obtained his doctoral degree from the University of Oulu in 2014. During 15+ years he has worked at the Center for Machine Vision and Signal Analysis. He is currently Senior Scientist at VTT Technical Research Centre of Finland and Associate Professor at the University of Oulu. His research interests include face analysis, embedded AI, image-based real-time sensing and energy-efficient embedded computer vision. Contact him at miguel.bordallo@oulu.fi