MDN: A Deep Maximization-Differentiation Network for Spatio-Temporal Depression Detection

Wheidima Carneiro de Melo, *Student Member, IEEE*,
Eric Granger, *Member, IEEE*, and Miguel Bordallo López

Abstract—Deep learning (DL) models have been successfully applied in video-based affective computing, allowing, for instance, to recognize emotions and mood, or to estimate the intensity of pain or stress of individuals based on their facial expressions. Despite the recent advances with state-of-the-art DL models for spatio-temporal recognition of facial expressions associated with depressive behaviour, some key challenges remain in the cost-effective application of 3D-CNNs: (1) 3D convolutions usually employ structures with fixed temporal depth that decreases the potential to extract discriminative representations due to the usually small difference of spatio-temporal variations along different depression levels; and (2) the computational complexity of these models with consequent susceptibility to overfitting. To address these challenges, we propose a novel DL architecture called the Maximization and Differentiation Network (MDN) in order to effectively represent facial expression variations that are relevant for depression assessment. The MDN, operating without 3D convolutions, explores multiscale temporal information using a maximization block that captures smooth facial variations and a difference block that encodes sudden facial variations. Extensive experiments using our proposed MDN with models with 100 and 152 layers result in improved performance while reducing the number of parameters by more than $3\times$ when compared with 3D ResNet models. Our model also outperforms other 3D models and achieves state-of-the-art results for depression detection. Code available at: https://github.com/wheidima/MDN.

Index Terms—Affective computing, deep learning, convolutional neural networks, face analysis, depression detection

1 Introduction

Health care has attracted an increasing amount of interest from the computer vision and machine learning communities due to its large number of applications. It is anticipated that automatic diagnosis systems may provide effective decision support for clinicians in an explainable, unobtrusive, and objective manner regardless of the identity, gender, age, and ethnicity of the subject. Recently, much progress has been made towards this goal, specially for systems based on facial analysis [1]. Such systems leverage the fact that the facial modality acts as a mirror of the health condition which may expose symptomatic signs of particular diseases, including mental health conditions. For instance, Giannakakis *et al.* [2] explored facial cues obtained from eye activity, mouth activity and head movements for the recognition and analysis of stress and anxiety states.

An emerging field for automatic health care diagnosis methods is depression detection. Major Depressive Disorder, also known as depression, is a common mental disorder with an immense economic burden. Such mental disorder is associated with a negative state of mind which persists for a long time. It may cause alterations in appetite [3], sleep disturbances [4], limited ability to concentrate [3], headache [5], backache [5], stomach ache [6], anxiety [7], loss of pleasure and/or interest in persons or things [3]. In severe cases, depression leads to suicidal behavior and substance abuse [8]. Furthermore, depression may amplify the chances of developing and sometimes contribute to the progress of serious clinical states, such as diabetes, cardiovascular disease, and cancer [9].

Despite the gravity of depression, there are effective treatments for this disorder. Typical treatments include antidepressants, mood stabilizers, and psychotherapeutic approaches. Consequently, an accurate diagnosis of depression and its severity is crucial for immediate treatment and reduction of negative consequences. Normally, clinical practice is based on Diagnostic and Statistical Manual of Mental Disorders (DSM-5) specifications [10] that are analyzed under structured interviews. The severity of depression is determined by employing self-report inventories, e.g., Beck Depression Inventory (BDI), or an inventory such as Hamilton Depression Rating (HAM-D), usually managed by a clinician with experience in treating psychiatric patients. However, some studies have shown that clinicians have difficulties to recognize depression [11], [12].

- *Wheidima Carneiro de Melo is with the Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland. E-mail: wheidima.melo@oulu.fi.*
- *Eric Granger is with the Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), Department of Systems Engineering, École de technologie supérieure, Montreal, Quebec H3C 1K3, Canada. E-mail: eric.granger@etsmtl.ca.*
- *Miguel Bordallo López is with the VTT Technical Research Centre of Finland, 02044 Espoo, Finland. E-mail: miguel.bordallo@vtt.fi.*

Manuscript received 16 Sept. 2020; revised 25 Feb. 2021; accepted 3 Apr. 2021. Date of publication 12 Apr. 2021; date of current version 28 Feb. 2023.
(Corresponding author: Wheidima CarneirodeMelo.)
Recommended for acceptance by A. Dhall.
Digital Object Identifier no. 10.1109/TAFFC.2021.3072579

Inefficient depression detection has resulted in an alarming number of false-positives that posed serious consequences including the death of patients [12]. Moreover, the clinical interventions are normally labor-intensive, expensive and require considerable expertise in managing depressive states.

From this perspective, decision support diagnosis systems based on machine learning can provide an accurate and objective prediction of depression levels, contributing to the evaluation and monitoring of patients. Such methods may focus on visual-based nonverbal cues for depression detection. Indeed, studies have shown a set of visual cues which are correlated with depression [3], [13], [14]. These cues basically comprise specific facial expressions and dynamics, and limited levels of positive social behaviors (e.g., absence of smiles [3]). In this context, a commonly accepted approach is to automatically predict depression levels by exploring the facial information of subjects in videos.

Although several deep learning (DL) models have been proposed for depression detection from video-based facial analysis [15], [16], [17], [18], [19], [20], [21], these architectures may not consistently achieve a high level of performance. There are at least two reasons for this issue. The first one is that DL models are failing to encode the spatio-temporal information from facial expressions along depression levels. Two video sequences with distinct labels can exhibit small differences in the variations of facial expression. In this case, the use of models that explore fixed-range temporal information decreases the ability to produce discriminative representations. The second reason is the limited amount of annotated training data that is available to design the predictive architectures. When applied to depression detection, effective DL models for video representation based on 3D Convolutional Neural Networks (3D-CNNs) require optimizing a large number of parameters. Therefore, the risk of overfitting is high because of the relatively small size of training datasets.

In other applications domains, e.g., action recognition [22], [23], diverse 3D architectures have been proposed to capture spatio-temporal features [23], [24], [25], [26]. However, their high level of performance is typically achieved at the expense of high computational complexity. Hence, training these architectures requires large amounts of training data and computational resources. Some authors have proposed to reduce the computational cost of 3D models by using different forms of spatio-temporal convolutions [22], [27], [28], [29], but such approaches explore fixed-temporal information. We argue that this decreases the potential for generating discriminative feature representations for depression detection.

In this paper, an effective architecture, named Maximization and Differentiation Network (MDN), is proposed to explore facial expression variations at different temporal scales. This DL model is composed of a maximization block and a difference block. Given an input, the maximization block is employed to capture smooth transitions of facial structures, while the difference block encodes sudden spatio-temporal variations. These blocks do not rely on 3D filters, and their generated features are combined in a way that leads to a robust feature representation for depression detection. We design our MDN module by using residual-like structure since such skip connections have shown their effectiveness for training CNNs. For experimental validation, our MDN module is integrated into a 3D ResNet-like architecture, although it can be incorporated into other CNN architectures.

The main contributions of this paper are:

- The definition of maximization and difference blocks that encode the complementary smooth and sudden facial expression variations, respectively.
- The combination of the maximization and difference blocks into an efficient MDN module such that a wide range of spatio-temporal facial variations can be explored without employing complex 3D filters.
- An extensive experimental study indicating that our proposed MDN provides a cost-effective solution, outperforming different 3D-CNNs and state-of-the-art models on two publicly available benchmarking datasets: AVEC2013 and AVEC2014. Experiments also show that, for deeper networks, our MDN reduces the number of parameters by around $3.3\times$, and improves the performance over 3D ResNet.

The rest of this paper is organized as follows. Section 2 provides some background on models for depression detection, as well as for spatio-temporal recognition. Our proposed MDN is presented in Section 3. Finally, Sections 4 and 5 describe the experimental methodology (datasets, protocols and performance metrics), and results for validation while Section 6 draws the conclusions of the present work.

2 Related Work

2.1 Automatic Depression Estimation

People affected by depression have been demonstrated to exhibit higher chances of facial expression disturbances due to mood variations [3]. For example, the authors in [13] report the restriction of facial expressiveness (or emotional variability) associated with depressive states. In this context, sad facial expressions are shown to be more predominant [30], while depressed patients show to have limited eye contact [3] and less intense smiles [31]. The number of head movements and their intensity is also statistically lower when compared with healthy subjects [14]. All these visual cues have the potential to be automatically explored to support in the detection, diagnosis, and assessment of depression by using videos containing human faces and machine learning approaches. In the literature, conventional machine learning approaches have been proposed, often by applying hand-engineered feature representations, e.g., Local Binary Patterns (LBP), followed by regression analysis, such as Support Vector Regression (SVR). In contrast, DL approaches perform end-to-end learning, typically using a 2D-CNN followed by a recurrent network, or using a 3D-CNN, where a regression layer generates the output.

2.2 Hand-Engineered Methods

Recently, events like the Audio-Visual Emotion Challenge and Workshop, AVEC 2013 [32] and 2014 [33], have increased the interest and number of contributions in automated depression analysis. The baseline facial descriptor in AVEC2013 [32] was the Local Phase Quantization (LPQ), and SVR was employed for prediction of depression levels. The following researches on AVEC2013 dataset are relied on LPQ [34], LBP [35], Pyramid of Histogram of Gradients (PHOG) [36], Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) [37], and Canonical Correlation Analysis (CCA) [38]. The baseline facial descriptor in AVEC 2014 [33] was the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP). Following work from Jan *et al.* [39] extract three different texture representations employing LBP, LPQ, and Edge Orientation Histograms (EOH), while temporally mapping their variations using Motion History Histogram (MHH). Kaya *et al.* [40] compute LGBP-TOP and LPQ features, and further analyze them by using CCA whereas Dhall *et al.* [41] and Jain *et al.* [42] employ Fisher Vectors (FV) to derive the depression levels.

2.3 Deep Learning Methods

More recently, some deep neural networks have been proposed to address depression detection, and other applications in affective computing. In particular, Zhu *et al.* [15] proposed a DL architecture which is comprised of two streams using facial images and optical flow as inputs. In [43], the authors presented a Deep Transformation Learning (DTL) scheme to project facial features into a new feature subspace with the purpose to capture the non-linearity of the data. Jazaery *et al.* [18] used two Convolutional 3D (C3D) networks [25] to capture spatio-temporal features at two different scales. Extending on this idea, Melo *et al.* [19] employed two C3D to extract spatial and temporal features from two different facial areas. Jan *et al.* [21] employed a 2D-CNN to explore appearance information, while the variations of the features are encoded using Feature Dynamic History Histograms (FDHH).

Following recent trends, Residual Networks [44] have also been explored in depression detection. For example, the depression level was predicted by using a 50-layer residual network (ResNet-50), and deep distribution learning [20], whereas a ResNet-50 was used in [17] with an attention mechanism to combine facial features. In [16], four ResNet-50 were employed to estimate depression levels while providing the facial regions that provide most information about depression. Finally, Song *et al.* [45] presented an approach to explore behavior primitives (facial action units, head pose, and gaze directions) by transforming one-dimensional signals into their spectral representations. These representations are then fed to a DL network that performs the final regression of the depression levels. The majority of these methods exploit spatial and temporal information separately by using 2D-CNNs and some approach to explore the facial features. However, such an approach may deteriorate the intrinsic spatio-temporal relationships.

2.4 Modelling Spatio-Temporal Information

To directly encode the facial appearance and dynamics for depression detection in videos, it is essential to produce efficient representations. Several DL models have been proposed to model spatio-temporal information. Tran *et al.* [25] proposed the architecture called C3D which was one of the first methods to capture spatial and temporal information using 3D-CNN. Carreira *et al.* [24] proposed Inflated 3D-ConvNet (I3D) which is a transformation of 2D Inception model into 3D-CNN by inflating all the filters and pooling kernels. In [26], authors explored the effectiveness of diverse 3D-CNN architectures based on residual networks (3D ResNet). Feichtenhofer *et al.* [23] presented SlowFast network that is composed of a slow path to explore spatial semantics and a fast path to explore motion at fine temporal resolution.

In general, all these architectures have structures with fixed temporal depth. In this case, it is difficult to generate effective features representations for depression detection since the difference of spatio-temporal variations between the depression levels is often small. Moreover, the number of model parameters to optimize is typically very large, which increases the chances of overfitting due to the limited amount of annotated training data that is available for depression detection. Some authors have proposed different techniques of spatio-temporal convolutions. In particular, Tran *et al.* [27] factorized 3D convolution into two cascaded operations, a 2D convolution (spatial) and a 1D convolution (temporal). Xie *et al.* [28] investigated various forms of 3D-CNNs where Top-Heavy I3D, which employs 2D structures in the lower layers, and 3D structures in the upper layers, presents better performance. In [29], the authors proposed a Pseudo-3D Residual Network (P3D ResNet) by using a spatio-temporal decomposition on a residual learning module. Finally, Jiang *et al.* [22] proposed to encode spatio-temporal and motion features jointly using 2D and 1D CNNs. The proposed MDN module also decomposes the 3D convolution operation, but our approach does not employ 1D convolutions. Instead, we use 2D convolutions and two functions without trainable parameters to capture features at multiple ranges.

3 The Proposed Maximization and Differentiation Network

The face of a person suffering from depression exhibits specific spatio-temporal patterns of variation. The goal of automatic depression detection from videos is to encode the facial expression variations that carry the most relevant discriminative information. In this context, our proposed approach captures the spatial and temporal information without using 3D filters, allowing to limit model complexity. We propose a maximization block to summarize spatio-temporal information, and a difference block to encode the details of the spatio-temporal variations. These blocks are combined into the MDN module.

3.1 Maximization Block

The idea of the maximization block is to model global spatial and temporal variations. Using a function that summarizes such variations in a cascade with 2D convolutional layers allows the module to extract relevant spatio-temporal features, which can improve the performance of a depression detection model. As the block is based on the max function, it has the potential to capture smooth facial variations. Given that for an input feature map the semantic

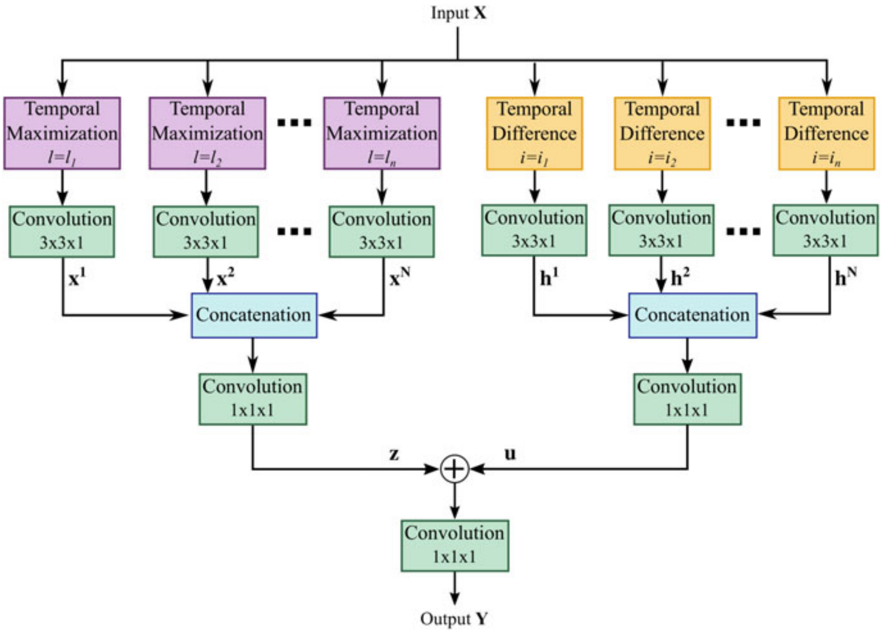

Input X
Temporal Maximization $l=l_1$
Temporal Maximization $l=l_2$
Temporal Maximization $l=l_n$
Temporal Difference $i=i_1$
Temporal Difference $i=i_2$
Temporal Difference $i=i_n$
Convolution 3x3x1
$\mathbf{x}^1$
$\mathbf{x}^2$
$\mathbf{x}^N$
$\mathbf{h}^1$
$\mathbf{h}^2$
$\mathbf{h}^N$
Concatenation
Convolution 1x1x1
z
u
Output Y

Fig. 1. Architecture of the MDN module. It is composed of maximization blocks which capture smooth facial expression variations, and difference blocks which explore sudden spatial temporal transitions. A linear combination is performed at the last stage of the module.

information is redundant along the temporal depth, we claim that such information can be summarized employing a simple operation without the using of trainable parameters. Let $\mathbf{X} \in \mathbb{R}^{N\times T\times H\times W\times C}$ represents an input feature map, where N, T, H, W and C are the batch size, temporal depth, height, width, and the number of channels, respectively. We formally define the operation as

$$\mathbf{V}_{t,h,w} = \max\{\mathbf{X}_{t:t+l,h,w}\}, \tag{1}$$

where $\mathbf{V}$ is the spatio-temporal representation, l is the length of the sliding window used to perform max pool along depth axis, and t, h, w denote the depth and the spatial dimensions, respectively. Note that this representation employs the same dimensions of the input feature map.

Instead of exploring spatio-temporal variations with structures that employ fixed temporal depths, our proposed block uses different ranges of dynamics, contributing to capture supplementary information for depression representation. As shown in Fig. 1, the maximization block is composed of N branches, each one can operating in a distinct range, i.e., $l_1, l_2, \ldots, l_N$. It is important to note that a higher number of branches increases the number of parameters, which in turn increases the model training times. On the other hand, a small number of branches may decrease the capabilities of the model. Let $\mathbf{x}^i$ denote the output of branch i, then the block's output can be expressed by

$$\mathbf{z} = \mathcal{H}\Big\{ \bigcup\nolimits_{n=1}^{N} \mathbf{x}^n \Big\}, \tag{2}$$

where $\mathbf{z}$ represents the final feature map, $\mathcal{H}\{\}$ is a fusion function carried out by a $1\times 1\times 1$ convolutional layer, $\bigcup$ is the operation that concatenates the output of each branch, and N refers to the number of branches.

This procedure encodes a set of spatio-temporal information in a single map, which can convey in its texture information about movement, favoring the exploitation of the dynamics by a set of 2D filters. Moreover, as our approach is based on structures with variable temporal depth, the use of 2D filters rather than 3D filters, avoids an exponential increase in the number of parameters and decreases the risk of overfitting.

3.2 Difference Block

To generate a robust representation of facial variations, it is important to encode sudden transitions of facial structures. These transitions can, for example, assist the model to analyze segments of a video with similar facial expression variations. Motivated by this, we propose a structure called difference block that explores the velocity of facial expression variations.

Let $\mathbf{X} \in \mathbb{R}^{N\times T\times H\times W\times C}$ define the input feature maps, the first step of the difference block is to compute the absolute value of the difference between the feature maps. This operation is defined by

$$\mathbf{H}_t = |\mathbf{X}_t - \mathbf{X}_{t-i}|, \tag{3}$$

where $\mathbf{H}_t$ is the output of the operation, t is the temporal depth, and i represents ith order difference. Similar to maximization block, the difference block is formed by N branches which obtain velocity of the spatio-temporal variations by performing difference of order $i_1, i_2, \ldots, i_N$. In our implementation, we keep the depth size of the output equal to the input feature map adding zeros to the input features when carrying out the operation.

As the difference block is designed to explore short variations, lower order differences should be employed, such as 1, 2 and 3. The difference block with high order is useful to explore long-term variations. As we can see in Fig. 1, 2D filters explore the spatial dependencies in the feature maps generated in this process. Finally, the block's output is generated using the following equation:

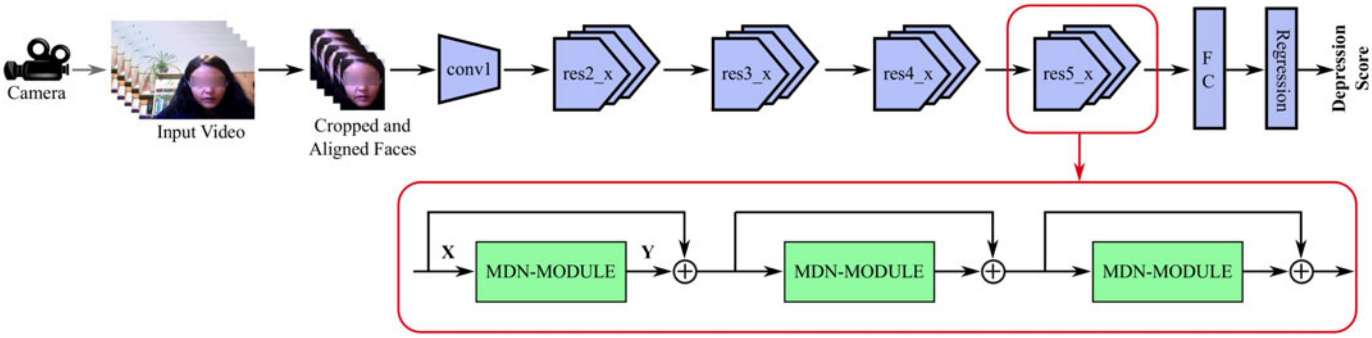


Fig. 2. Overall architecture of the MDN. The first stage is responsible for cropping and alignment of the faces. The feature extractor is based on ResNet architectures where we replace the residual blocks with MDN module, with the exception of the first layer.

$$\mathbf{u} = \mathcal{H}\Big\{ \bigcup_{n=1}^{N} \mathbf{h}^n \Big\}, \tag{4}$$

where $\mathbf{h}^n$ is the output of the nth branch, N is the number of branches, $\bigcup$ is the concatenation operator, and $\mathcal{H}$ is the fusion function. We perform the concatenation along the channels' axis in both difference and maximization blocks.

3.3 MDN Module

The combination of the maximization with difference blocks generates the MDN module. Observe that the maximization block and difference block can explore distinct spatio-temporal information and can also operate in different temporal ranges. With that, our MDN module has the potential to encode spatial and temporal information from smooth and sudden facial variations. Such ability can significantly boost the performance of a model for automatic depression detection.

As shown in Fig. 1, the outputs of the maximization and difference blocks are merged using a linear combination, which fuses the features of the two blocks by addition. For that, it is necessary to make sure that the dimensions of the output feature maps of the blocks are the same. We use the fusion function $\mathcal{H}$ in the blocks to adjust the feature maps. The advantage of this approach is to reinforce the complementary behavior of the blocks. Then, an additional $1 \times 1 \times 1$ convolutional layer is employed to adjust the number of channels to match the input feature maps, since we employ our MDN module inside structures with residual-like connections which additionally fuse the features, as illustrated in Fig. 2. Given this later convolutional layer, we consider our MDN module with two layers.

4 Experimental Methodology

4.1 Datasets

To evaluate the performance of our proposed MDN, we conduct extensive experiments on two publicly available benchmark datasets, namely the Audio-Visual Emotion Challenge 2013 and 2014 (AVEC2013 [32] and AVEC2014 [33]) depression sub-challenge datasets. These datasets were employed in the AVEC sub-challenge, where the goal was to estimate the score of individuals on the Beck Depression Inventory (BDI-II). According to the BDI-II score, the severity of depression can be classified in four levels: minimal $(0 - 13)$, mild $(14 - 19)$, moderate $(20 - 28)$, and severe $(29 - 63)$.

Although it is possible to find other publicly available datasets for depression assessment, such as AVEC 2016 [46], these datasets only provide feature sets of the individuals. Our proposed architecture is designed to explore spatio-temporal dependencies directly from facial videos. To the best of our knowledge, AVEC2013 and AVEC2014 datasets are the only ones that currently provide with raw facial video data. For this reason, and following the state-of-the-art, we benchmark our experiments in these two datasets.

The AVEC2013 dataset is derived from a subset of the audio-visual depressive language corpus (AViD-Corpus). The subjects were recorded during an interaction with a computer performing diverse tasks, such as counting from 1 to 10. In total, the dataset contains 150 video clips allocated into three different partitions: training, development and test sets. Each set consists of 50 videos which have a label related to depression score of subjects. The videos have duration ranging between 20 and 50 minutes with an average video length of 25 minutes.

The AVEC2014 dataset is also a subset of AViD-Corpus. For this dataset, two tasks named Freeform and Northwind are performed while the subjects of the videos are recorded. In the Freeform task, the subjects answer questions such as discuss a sad childhood memory. In the Northwind task, subjects read audibly an excerpt from a fable. In both tasks, the videos are allocated into three partitions: training, development, and test sets. Each set contains 50 videos with a ground truth numerical label for every video. The dataset is formed of 300 videos that range between 6 and 248 seconds. For both datasets, the frame rate of the videos is 30 frames per second (fps).

4.2 Experimental Setup

Since our MDN module is designed to be embedded in structures with identity shortcut connections, the proposed architecture is based on 3D residual networks [26], although the MDN module could be also employed in other different 3D networks (e.g., I3D or C3D [24], [25]). In this subsection, we describe the resulting deep network architecture.

The MDN architecture is a convolutional network which explores spatio-temporal variations using maximization and difference structures. Employing our MDN module, five networks are built with sizes of 18, 34, 50, 100 and 152 layers. The details of the networks are presented in Table 1. All the networks have the first layer (conv1) with one block

TABLE 1
The Proposed Networks

Layers	Output Size	Filter	F	Num. of Blocks (18,34,50,100,152)	Depth	Order
conv1	$56 \times 56 \times 16$	$7 \times 7 \times 7$	64	1	-	-
Pooling	$28 \times 28 \times 8$	$3 \times 3 \times 3$ max pool, stride 2				
res2_x	$28 \times 28 \times 8$	$3 \times 3 \times 1$ $1 \times 1 \times 1$	32	2,3,3,3,3	$l_1 = 2, l_2 = 3, l_3 = 4$	$i_1 = 1, i_2 = 2$
res3_x	$14 \times 14 \times 4$	$3 \times 3 \times 1$ $1 \times 1 \times 1$	64	2,4,4,14,24	$l_1 = 1, l_2 = 2, l_3 = 3$	$i_1 = 1, i_2 = 2$
res4_x	$7 \times 7 \times 2$	$3 \times 3 \times 1$ $1 \times 1 \times 1$	128	2,6,14,29,45	$l_1 = 1, l_2 = 1, l_3 = 2$	$i_1 = 1, i_2 = 2$
res5_x	$4 \times 4 \times 1$	$3 \times 3 \times 1$ $1 \times 1 \times 1$	256	2,3,3,3,3	$l_1 = 1, l_2 = 1, l_3 = 1$	$i_1 = 1, i_2 = 2$
Pooling	$1 \times 1 \times 1$	$4 \times 4 \times 1$ avg pool				
Fully connected	1×256	256D fully connected				
Regression	1	linear regression				

F represents the number of feature channels. Depth is the length of spatio-temporal variation that is explored by the maximization block while Order indicates the difference used. The res3_1, res4_1 and res5_1 layers, perform spatial temporal downsampling with stride 2.

and the others with different number of blocks. Only conv1 uses typical 3D convolution because it employs a different temporal kernel depth. Moreover, we employ a different temporal depth for each channel of the maximization block, considering the input size, to benefit the exploitation of the features. However, when the temporal information of the input is equal to 1, we set the depth to 1. Observe that the networks employ MDN module with a maximization block composed by 3 branches whereas difference block have 2. In the next section, we analyze the effect of changing the number of branches and the temporal range that is explored by the blocks.

After the sequence of convolutions, the average pooling layer with kernel size $4 \times 4 \times 1$ produces a 256-dimensional feature vector which is fed to the last layer. As the depression detection from facial videos can be considered as a regression problem, our last layer is composed by one fully connected layer and a linear regression function that we implemented using an additional fully connected layer with one neuron.

Training. Due to the fact that there is a limited amount of training data in the AVEC2013 and AVEC2014 datasets to train a deep architecture from scratch, the proposed MDN networks are initially trained on face recognition. The networks are pre-trained on the VGGFace2 dataset that includes 3.31 million images of 9,131 identities [47]. In this process, an image is selected from the dataset and replicated 16 times in order to make a clip that is fed into the model. We employ Stochastic Gradient Descent (SGD) with momentum of 0.9, weight decay 0.0001, and an initial learning rate of 0.01. The learning rate is multiplied by 0.1 after every 10 epochs. At this stage, the input values are per channel subtracted by the average value of VGGFace2. In this face recognition pre-training, the last layer of the models is a classification layer that is removed in the next stage.

For the fine-tuning stage, the ADAM optimization algorithm is adopted with an initial learning rate of 0.001, and a weight decay of 0.00001, and this rate is multiplied by 0.1 after each epoch where the limit is set to 0.00001. To build one training sample, a frame inside the video is randomly chosen and the subsequent frames are collected, where the frame sampling is empirically set to 7.5 fps. We pre-process the input by using the Multi-task Cascaded Convolutional Network (MTCNN) [48] for face detection and alignment in each frame of the video that are subsequently resized. This results in samples of 112 pixels ×112 pixels ×16 frames. For data augmentation, each sample is horizontally flipped with 25 percent probability, randomly rotated to 10 degrees with 25 percent probability, and turned upside down with 25 percent probability. The training samples created in this process are labelled using the same depression score as their original videos.

Testing. In the testing, we analyze the facial video by dividing it using a sliding window, with non-overlapping clips of 16 frames each. The final estimated depression score for an individual in a sample (input video) is obtained by simply averaging the predicted depression scores of all clips that compose the test video.

Evaluation Measures. For performance evaluation of the proposed architecture and a fair comparison with the state-of-the-art methods, two metrics are employed: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

5 Results and Discussion

In this section, we show the efficiency of the proposed approach in exploring spatial temporal dependencies from facial dynamics. First, we present the pre-training strategy and an analysis of different configurations of the MDN module. In the sequence, we provide different networks using our proposed module and compare them with standard 3D ResNet, other 3D schemes, and the state-of-the-art methods. Next, we perform cross database and error analysis and visualize the features generated by our architecture and the activation maps. Finally, we evaluate our method for pain estimation.

5.1 Pre-Training of MDNs

Properly initialized weights for fine-tuning towards depression detection can significantly improve the performance of deep networks. Table 2 reports results of MDN-50 pre-trained on ImageNet [52], and VGGFace2, as well as without pre-training. The results clearly indicate that MDN-50 achieves significantly better performance when pre-trained on large datasets. The model achieves its best performance when pretrained on VGGFace2, although the results are

TABLE 2
Analysis of Performance Using Different Datasets to Pre-Train the MDN

Pre-training	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
None	9.40	7.56	9.09	7.39
ImageNet	8.62	6.72	8.26	6.45
VGGFace2	**8.13**	**6.39**	**8.16**	**6.45**

very competitive on AVEC2014. This is expected since VGGFace2, AVEC2013, and AVEC2014 are face datasets. Therefore, once the MDN is pre-trained on VGGFace2, the MDN develops the ability to explore facial structures which can be considered as basis to encode the spatio-temporal variations in faces.

5.2 MDN Module Branch Number Analysis

Table 1 shows the definitions of the networks that consider an MDN module with 5 branches, using 3 branches to explore smooth information and 2 branches to capture the sudden temporal variations. However, the MDN module can be configured with a different number of branches and orders in both maximization and difference blocks. In this section, we study the effect of changing the number of branches in the maximization and difference blocks as well as the value of the temporal range that is analyzed. We conduct the study considering several configurations of the MDN module for MDN-50, i.e., MDN model with 50 layers. Since both datasets, AVEC2014 and AVEC2013, contain similar face videos and the analysis requires the training of several models and a long training process, we performed this analysis solely on the AVEC2014 dataset.

Table 3 reports the performance of the MDN-50 employing various configurations for MDN module. Specifically, we analyze the models for depth in range of $1 \leq l \leq 4$ where the temporal depth of input features is considered to define the values of depth in each layer of the model. Regarding order values, we define $i_n = n$ where n is the nth branch. The first model employs MDN module without difference block whereas the second one uses the module without maximization block. Results in Table 3 indicate that the models achieve similar results. Observe that the third model, which employs both blocks with the same configuration, achieves better results than both models, indicating the importance of exploring smooth and sudden information. Moreover, the networks with MDN module using an order equal to one, and one branch for exploring smooth information, achieve a better performance by using the sequence of 4, 3, 2 and 1 as depth values. Applying this sequence in the maximization block normally contributes to improve the performance of the model. In general, increasing the number of branches also improves the results of the model. However, the value of depth and order should be carefully chosen. For instance, the model using MDN module which captures temporal variations with values of depth equal to 1 and 2, and the sequence of 4, 3, 2 and 1 as depth values, outperformed all the models with just two branches, one for maximization block and other for difference block. However, this is not true for the other models using 2 branches for difference block and 1 for maximization block. Comparing the results when the MDN module is formed by using two maximization blocks and one difference block with this one composed by one maximization block and two difference blocks, we can see that the performance is competitive. Similar findings can be observed when we employ three branches for the maximization (or difference) block, and two for the difference (or maximization) block.

As can be seen from Table 3, the performance of the models using the MDN module with two branches for the maximization/difference block and three branches for the difference/maximization block is very similar when compared with the ones using four branches, two for each block. Moreover, the model with MDN module employing order equal to 1 and 2 combined with a maximization block that uses three branches achieves the best result in terms of RMSE. Based on these results, in the subsequent experiments, we decided to specify our MDN module using two branches in the difference block (Order = [1,2]) and three branches in the maximization block (see the last entry of Table 3).

TABLE 3
Evaluation of the MDN-50 With Different Configurations for the MDN Module

Layer								**AVEC2014**	
res2_x		**res3_x**		**res4_x**		**res5_x**			
Depth	Order	Depth	Order	Depth	Order	Depth	Order	RMSE	MAE
[4]	-	[3]	-	[2]	-	[1]	-	9.44	7.85
-	[1]	-	[1]	-	[1]	-	[1]	9.52	7.96
[4]	[1]	[3]	[1]	[2]	[1]	[1]	[1]	8.98	7.10
[3]	[1]	[2]	[1]	[1]	[1]	[1]	[1]	9.32	7.48
[2]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	9.64	7.10
[4]	[1, 2]	[3]	[1, 2]	[2]	[1, 2]	[1]	[1, 2]	8.64	6.70
[3]	[1, 2]	[2]	[1, 2]	[1]	[1, 2]	[1]	[1, 2]	9.36	7.37
[2]	[1, 2]	[1]	[1, 2]	[1]	[1, 2]	[1]	[1, 2]	9.20	7.08
[3, 4]	[1]	[2, 3]	[1]	[1, 2]	[1]	[1, 1]	[1]	9.00	6.92
[2, 3]	[1]	[1, 2]	[1]	[1, 1]	[1]	[1, 1]	[1]	8.75	6.71
[2, 3]	[1, 2]	[1, 2]	[1, 2]	[1, 1]	[1, 2]	[1, 1]	[1, 2]	8.40	6.53
[3, 4]	[1, 2]	[2, 3]	[1, 2]	[1, 2]	[1, 2]	[1, 1]	[1, 2]	8.37	6.58
[2, 3]	[1, 2, 3]	[1, 2]	[1, 2, 3]	[1, 1]	[1, 2, 3]	[1, 1]	[1, 2, 3]	8.35	**6.41**
[2, 3, 4]	[1, 2]	[1, 2, 3]	[1, 2]	[1, 1, 2]	[1, 2]	[1, 1, 1]	[1, 2]	**8.16**	6.45

Depth is related to maximization block, and Order refers to the order of the difference block. Values of Depth and Order are detailed for each layer of the MDN-50. The number of values in Depth or Order indicates the number of branches. For example, an Order = [1,2] means that there are two branches in the difference block.

TABLE 4
Error Rates of the Proposed MDN and 3D ResNet for Depression Detection, and Analysis of Their Time and Memory Complexity

Network	AVEC2013		AVEC2014		P.	F.
	RMSE	MAE	RMSE	MAE		
ResNet-18	9.24	7.06	9.14	6.92	33	8.38
MDN-18	8.96	7.21	8.82	6.77	7	5.66
ResNet-34	8.63	6.82	8.56	6.47	64	12.85
MDN-34	8.26	6.82	8.42	6.77	14	6.73
ResNet-50	8.81	6.92	8.40	6.79	63	12.22
MDN-50	8.13	6.39	8.16	6.45	21	7.40
ResNet-101	8.51	6.79	8.20	6.57	121	17.80
MDN-100	7.62	**6.14**	7.92	6.21	36	10.34
ResNet-152	8.30	6.58	8.01	6.30	168	24.7
MDN-152	**7.55**	6.24	**7.65**	**6.06**	52	13.36

***P.** and **F.** represent parameters ($\times 10^6$) and FLOPs ($\times 10^9$), respectively.*

5.3 Comparison With 3D Models

In order to show the efficiency of our approach, we present results of the proposed architecture and other 3D models. We begin by comparing our method with 3D ResNet in terms of RMSE, MAE and computational complexity. In addition, we also compare our architecture with Inflated 3D ConvNet (I3D) and Temporal 3D ConvNet (T3D) models. All 3D ResNet, I3D, and T3D models are trained following the same procedure as our proposed method – we first pretrained the model on VGGFace2 dataset, and then finetune it on either AVEC2014 or AVEC2013 datasets.

5.3.1 Analyses on AVEC2013

Table 4 reports the results for several MDN configurations and 3D ResNets on AVEC2013. As can be seen, the performance of the MDN improves with the increase of the network depth, except for MDN-152 in terms of MAE, where MDN-100 achieves better results. As the difference of performance between MDN-152 and MDN-100 is low, we understand that the MDN-152 could already be starting to overfit. It is also possible to observe that MDN-100 and MDN-152 achieve a considerable improvement of performance when compared with the smaller MDN-18.

Table 4 also shows that the MDN outperforms 3D ResNet for depression detection in terms of RMSE. Considering MAE, the results achieved by 3D ResNet-18 are slightly better than MDN-18, while 3D ResNet-34 and MDN-34 obtain the same results. However, for the deeper models, the MDN consistently outperforms the 3D ResNet approaches by a large margin both in terms of RMSE and MAE. For instance, the MDN-100 significantly reduces the MAE by 0.63 compared to 3D ResNet-101. From these results, we argue that MDN models are a better option for depression detection than their 3D ResNet architecture counterparts.

5.3.2 Analyses on AVEC2014

In Table 4, we show the results for MDN networks and 3D ResNet on AVEC2014. As it can be seen, the results of the MDN models improve again with the increase of the network depth, excluding the MDN-18 and MDN-34, that, in terms of MAE, achieve the same results. For this dataset, the MDN-152 achieve the best results, reducing the RMSE by 1.43 compared to MDN-18. Therefore, we might conclude that MDN-152 does not seem to overfit for this dataset.

TABLE 5
Results and Analysis of Complexity of MDN, I3D, and T3D Architectures

Network	AVEC2013		AVEC2014		P.	F.
	RMSE	MAE	RMSE	MAE		
I3D	8.66	6.64	8.55	6.36	13	6.99
T3D	8.75	6.76	8.55	6.54	68	51.64
MDN-50	8.13	6.39	8.16	6.45	21	7.40
MDN-100	7.62	**6.14**	7.92	6.21	36	10.34
MDN-152	**7.55**	6.24	**7.65**	**6.06**	52	13.36

***P.** and **F.** represent parameters ($\times 10^6$) and FLOPs ($\times 10^9$), respectively.*

We also show in Table 4 that the MDN outperforms 3D ResNet in terms of RMSE. Analyzing MAE, the results achieved by 3D ResNet-34 are better than MDN-34, but for the other models, the MDN outperforms the 3D ResNet approaches in terms of RMSE and MAE. For example, the MDN-152 significantly reduces the RMSE by 0.30 compared to 3D ResNet-152. The results in Table 4 indicate that the MDN architecture could be overcoming problems such as ambiguity and overfitting more accurately than 3D ResNet, especially for models with larger network size.

5.3.3 Computational Complexity

Table 4 presents the computational complexity comparison between the proposed MDN and 3D ResNet architectures. The number of parameters of MDN models is considerably less than 3D ResNet. The MDN-18 and MDN-34 have almost 5 times less parameters than 3D ResNet-18 and 3D ResNet-34 whereas the deeper MDN models (with 100 and 152 layers) have almost 3.3 times less parameters than the deeper 3D ResNet models. We also show the number of Floating Point Operations (FLOP) of the architectures as a measure of computational cost. MDN models present a considerable smaller number of FLOPs than 3D ResNet models. E.g., in the case of models with 34 layers, the FLOP value of MDN decreases approximately 2 times when compared with 3D ResNet.

5.3.4 Comparison With Other 3D Methods

We compare our method with other well-known 3D models, I3D and T3D. The I3D model [24] is composed of a basic structure called inception module, which is obtained by inflating 2D filters and pooling kernels of a 2D version of the module. The T3D model [49] contains structures called temporal transition layers which are responsible for capturing temporal information in different ranges. These two architectures have been successfully employed in action recognition, and the comparison with such models is important to measure the capabilities of the proposed MDN architecture.

Table 5 shows a direct comparison between the performance of I3D [24], T3D [49], and our three best models (MDN-50, MDN-100, and MDN-152). When compared with T3D, the I3D has competitive results with a smaller number of parameters. On the other hand, our proposed networks outperform the I3D model on both datasets (except for

TABLE 6
Error Rates of Methods for Predicting the Depression Scores on the AVEC2013 Dataset

Method	RMSE	MAE
Baseline [32]	13.61	10.88
LPQ + SVR (Kächele *et al.* [34])	10.82	8.97
MHH + LBP (Meng *et al.* [35])	11.19	9.14
PHOG (Cumins *et al.* [36])	10.45	N/A
LPQ-TOP + MFA (Wen *et al.* [37])	10.27	8.22
CCA (Kaya *et al.* [38])	9.72	7.86
Two CNN (Zhu *et al.* [15])	9.82	7.58
Two C3D (Jazaery *et al.* [18])	9.28	7.37
Classifier + Regressor (Ma *et al.* [50])	8.91	7.26
Two C3D (Melo *et al.* [19])	8.26	6.40
Four ResNet-50 (Zhou *et al.* [16])	8.28	6.20
ResNet-50 (Melo *et al.* [20])	8.25	6.30
Behavior signals (Song *et al.* [45])	8.10	6.16
Two ResNet-50 (Melo *et al.* [51])	7.97	**5.96**
MDN-50 (ours)	8.13	6.39
MDN-100 (ours)	7.62	6.14
MDN-152 (ours)	**7.55**	6.24

TABLE 7
Error Rates of Methods for Predicting the Depression Scores on the AVEC2014 Dataset

Method	RMSE	MAE
Baseline [33]	10.86	8.86
MHH + PLS (Jan *et al.* [39])	10.50	8.44
LGBP-TOP + LPQ (Kaya *et al.* [40])	10.27	8.20
DTL (Kang *et al.* [43])	9.43	7.74
Two CNN (Zhu *et al.* [15])	9.55	7.47
Two C3D (Jazaery *et al.* [18])	9.20	7.22
Two C3D (Melo *et al.* [19])	8.31	6.59
VGG + FDHH (Jan *et al.* [21])	8.04	6.68
ResNet-50 + Pool (Zhou *et al.* [17])	8.43	6.37
Four ResNet-50 (Zhou *et al.* [16])	8.39	6.21
ResNet-50 (Melo *et al.* [20])	8.23	6.15
Two ResNet-50 (Melo *et al.* [51])	7.94	6.20
MDN-50 (ours)	8.16	6.45
MDN-100 (ours)	7.92	6.21
MDN-152 (ours)	**7.65**	**6.06**

MDN-50 on AVEC2014 in terms of MAE). We can observe that the difference of performance is higher in terms of RMSE. Regarding the T3D model, our models achieve better results where the difference in terms of RMSE on AVEC2013 is 1.13, considering the MDN-100 model.

In Table 5, we also present the computational complexity of I3D, T3D and MDN architectures. Compared to T3D, the MDN-50, MDN-100 and MDN-152 models use fewer parameters and require a smaller number of FLOP computations. The I3D model employs even less parameters when compared with our MDN models, and requires around 6.99 FLOPs, at the cost of a worse performance. These results are expected since our model is designed to explore both sudden and smooth temporal information.

In summary, the structures of the MDN module based on complementary functions and diverse depths demonstrated good potential to capture spatio-temporal variations in facial expressions. The proposed architecture can learn how to obtain a rich representation of the facial expression variations even with limited training data. The results of the proposed models indicate good performance to explore appearance and dynamics of facial videos for depression detection.

5.4 Comparison With State-of-the-Art

We compare the performance of our three networks, MDN-50, MDN-100, and MDN-152, with the state-of-the-art methods for depression detection on AVEC2013 and AVEC2014 datasets.

5.4.1 Comparisons on AVEC2013

Table 6 shows the performance of our proposed method compared with baselines and state-of-the-art methods on AVEC2013 dataset. The methods based on hand-engineered representations are [32], [35], [36], [37], [38], [50]. All these methods are outperformed by our MDN networks. Zhu *et al.* [15] proposed a method based on two-stream networks which uses RGB frames and optical flow as input. The proposed models achieve better results than this method, indicating that having structures with capability of capturing multiple ranges of information is effective for depression detection. In [18] and [19], the authors explore different facial regions using two C3D models. The MDN models outperform both methods, demonstrating the power of the model in exploring diverse facial regions. When compared with the models that employ one or more ResNet-50, MDN-50 achieves very competitive results, although MDN-50 employs fewer parameters.

MDN-100 and MDN-152 outperform the method in [20] that is based on distribution learning. In [16], the authors employ four ResNet-50 to explore facial areas, MDN-100 outperforms such method whereas MDN-152 obtains better results in terms of RMSE, and competitive performance in terms of MAE. We believe that such results confirm the importance of capturing directly spatio-temporal information with the MDN module rather than only appearance information. Song *et al.* [45] explore multiple behavior signals using Fourier transforms and a CNN. It can be observed that the performance of MDN-100 and MDN-152 surpasses this model, although, in terms of MAE, for MDN-152, the results are competitive. Finally, the authors in [51] employ a two-stream network where a temporal pooling method captures dynamic information into an image map. As we can see, MDN-100 and MDN-152 achieve better results in terms of RMSE, and competitive results in terms of MAE when compared with the method in [51].

5.4.2 Comparisons on AVEC2014

Table 7 reports the comparative results of our proposed models and the state-of-the-art on AVEC2014 dataset. Our methods outperform the schemes based on hand-crafted features that are [33], [39], [40]. The authors in [43] apply Deep Transformation Learning (DTL) to encode deep features. The MDN models yield lower values of RMSE and MAE than this method. We can observe that MDN-50 achieves good results where, in terms of RMSE, it is only outperformed by the methods in [21], [51] which employ many more parameters. MDN-100 and MDN-152 achieve

TABLE 8
Comparisons Considering Single Task and Fusion of Tasks on AVEC2014

Method	Task	RMSE	MAE
Behavior signals [45]	Freeform	8.30	6.78
MDN-152	Freeform	**7.67**	**5.95**
Behavior signals [45]	Northwind	-	-
MDN-152	Northwind	7.57	6.12
Behavior signals [45]	Fusion	7.15	5.95
MDN-152	Fusion	**7.10**	**5.74**

better results than the method in [17] which explore facial features with attention mechanisms. MDN-152 outperforms the methods in [15], [16], [18], [19], [20], [51]. The authors in [21] employ VGG network to explore facial images and use Feature Dynamic History Histogram (FDHH) to map the changes in the features. MDN-100 and MDN-152 achieve better results compared to this approach. Observe that our models outperform methods with different pooling schemes for facial static features, providing a good alternative approach to these methods. From the results in Tables 6 and 7, we can claim that our architecture is an efficient option to capture spatio-temporal information related to depressive behaviors from facial videos.

5.4.3 Task-Based Comparisons on AVEC2014

In Table 8, we present the performance of the proposed model for each task and for the combination of them as well as the results presented in [45]. As mentioned previously, the AVEC2014 have two tasks: Freeform and Northwind which are considered in the analysis. The results of our architecture for each task are very similar, indicating that our method keeps a good performance for exploring spatio-temporal variations regardless of the task. The combination of the tasks is carried out by a simple score fusion scheme considering all the values generated in both videos. It is important to note that each participant has one sample in each task (Freeform/Northwind) with the same depression score. Therefore, our method generates predictions by analyzing both samples. As we can see, the performance of the method improves with the fusion of the tasks, since the score fusion act as a regularizer that minimizes the effect of outliers. When compared against the method in [45], we observe that our architecture outperforms such method in task-based or combination of tasks approach. The results suggest that our architecture can produce more discriminative features using facial videos than models based on the analysis of high-level behavior features.

5.5 Cross-Database Analysis

To assess the generalization capability of MDN-152, we perform cross-database validation on AVEC2013 and AVEC2014 datasets. In this procedure, the model is trained on the source database and tested on the target database. Table 9 presents the results of this experiment. As can be seen, when the source is AVEC2013 dataset, the performance of the model degrades slightly when compared with AVEC2014 as the source database. However, in both cases, the results are competitive with the ones shown in Table 4. The representations learned by our proposed model provide good generalization ability.

TABLE 9
Performance of the Proposed Method in Cross-Dataset Setting

Training set	Test set	RMSE	MAE
AVEC2013	AVEC2014	8.04	6.40
AVEC2014	AVEC2013	7.90	6.19

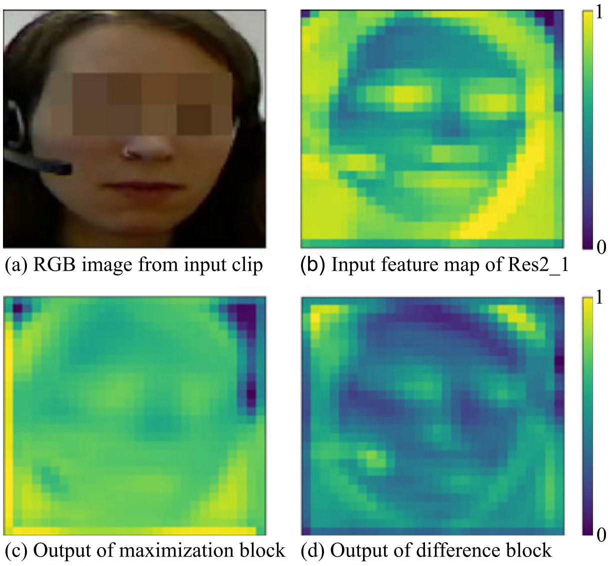
5.6 Qualitative Results

5.6.1 MDN Module Feature Visualization

Fig. 3 shows an analysis of our MDN based on depression feature maps generated by maximization and difference blocks. To facilitate the analysis and visualization, we consider the MDN module employed in the res2_1 layer. Fig. 3a presents a frame from the RGB input clip which is being analyzed, while the output of the max pooling layer, the one after the conv1 layer, that is fed into the res2_1 layer is shown in Fig. 3b. This provides insight into the input type for the MDN module. From the output of the maximization block, it can be noticed that the scheme spreads more energy along the face of the subject compared to the original input features. It demonstrates that the block is paying attention to global spatio-temporal information which increases the potential to explore smooth facial variations. With respect to the difference block, it can be seen that it captures the motion of feature maps and, since the block is based on first and second-order differences, this allows the module to explore sudden spatio-temporal variations. The complementary characteristics of the blocks builds a module with potential to explore rich spatio-temporal variations.

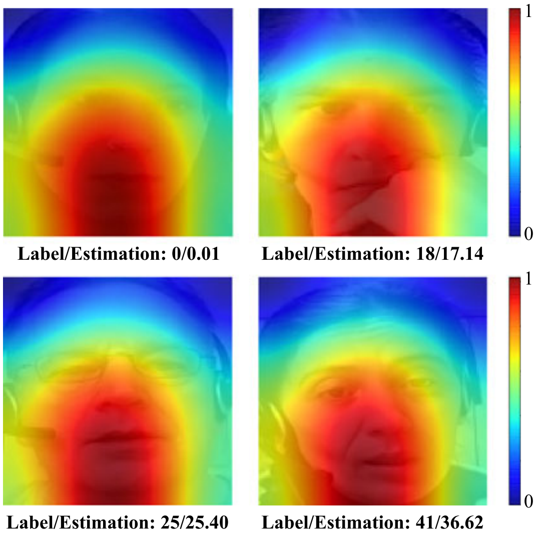
5.6.2 Visualization of Activation Maps

In order to interpret how the MDN architecture predicts depression scores from faces, we visualize class activation maps produced using the Grad-CAM method [53]. Fig. 4



(a) RGB image from input clip
(b) Input feature map of Res2_1
(c) Output of maximization block
(d) Output of difference block

Fig. 3. Feature visualization of MDN module.



Label/Estimation: 0/0.01
Label/Estimation: 18/17.14
Label/Estimation: 25/25.40
Label/Estimation: 41/36.62

Fig. 4. Visualization of activation maps for inputs with different depression levels.

shows an example of class activations maps produced for 4 distinct depression levels, that were interpolated and overlaid onto the corresponding facial images. The facial areas that most contribute to the prediction are represented by lighter colors. As shown in the figure, our model pays high attention to an area from the eyes to the chin. Interestingly, the most active area for all cases is the region that covers the mouth. It is important to observe that manifestations of depression include slow speech, fewer smiles, mouth shape, etc., which are characteristics that the model may explore. These visualizations show that MDN presents a different behavior when compared to models like in [16] which changes the most important facial area in accordance with the depression level. This indicates that MDN may rely on more optimal facial regions to explore spatio-temporal variations.

5.7 Error Analysis

In order to provide more information about the capabilities of the proposed architecture in a way that can be translated into clinical practice, we present the error for each sample (videos in the testing set) of the AVEC2013 and AVEC2014 datasets. We depict the errors in Fig. 5, ordered from the video presenting the smallest error to the one presenting the largest error. By observing the figure, we can conclude that the probability of error is approximately equally distributed from 0 to 12.5, with only a few outliers over that value. More concretely, the model achieves error less than 6.0 for more than 60 percent of the samples for both AVEC2013 and AVEC2014. It is worth noting that the error around 6.0 indicates a misclassification of the depression severity only between adjacent categories and for scores at the border of the class (e.g., the predicted level is 18, mild severity level, and the actual score is 12 which is the minimal level). The worst case is when a subject with minimal level of depression is classified with a severe level or the other way around. This is the case when the error is greater than 16. Our proposed model produced error greater than 16 only on 1 and 5 videos on AVEC2013 and AVEC2014, respectively. These results show that our architecture generalizes well, and the probability of grave misclassification is small.

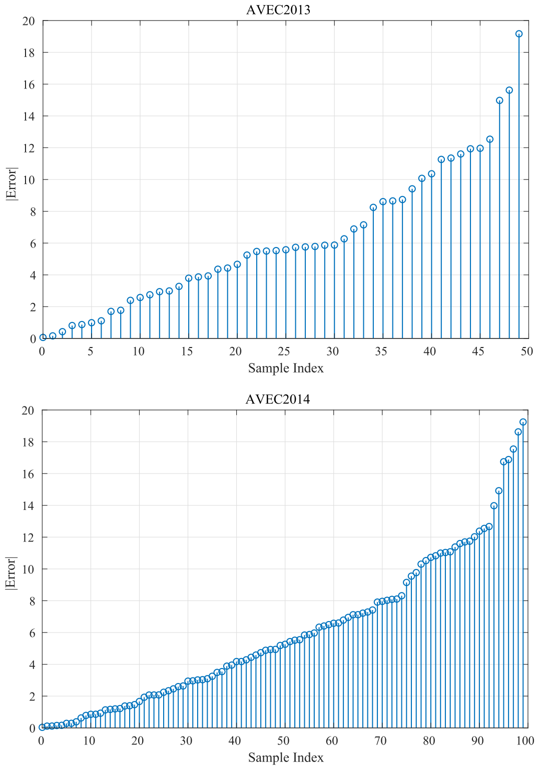


Fig. 5. Visualization of the error per sample using MDN-152 on AVEC2013 (above) and AVEC2014 (below). To facilitate the visualization, we present the absolute value of the error.

5.8 Pain Estimation

To further validate the ability of our proposed MDN to capture and leverage spatio-temporal information, additional experiments are conducted for pain intensity estimation. The dataset employed is the well-known UNBC-McMaster Shoulder Pain Expression Archive Database [54]. It includes 200 face videos of 25 subjects, each one annotated using PSPI score at frame-level in range of $0 - 15$.

For fair comparison with the state-of-the-art schemes, we report the performance of our approach in terms of Mean Squared Error (MSE) and MAE, where leave-one-subject-out cross-validation strategy is adopted. As the input of the MDN is a clip (16 frames), we define the ground truth as the mean of pain intensity of each frame inside the clip. In Table 10, we show the results of MDN-152 compared with six methods presented in the literature. As we can see, MDN outperforms five methods. For instance, our method obtains better results than the method in [59], where such method uses around 138 million parameters whereas MDN

TABLE 10
Comparison of Methods for Predicting the Intensity of Pain on the UNBC-McMaster Dataset

Method	MSE	MAE
Shape + DCT + LBP (Kaltwang *et al.* [55])	1.39	–
HoT (Florea *et al.* [56])	1.21	–
OSVR (Zhao *et al.* [57])	–	0.81
RCNN (Zhou *et al.* [58])	1.54	–
VGG16 + LSTM (Rodriguez *et al.* [59])	0.74	0.5
SCN (Tavakolian *et al.* [60])	**0.32**	–
MDN (ours)	0.68	**0.42**

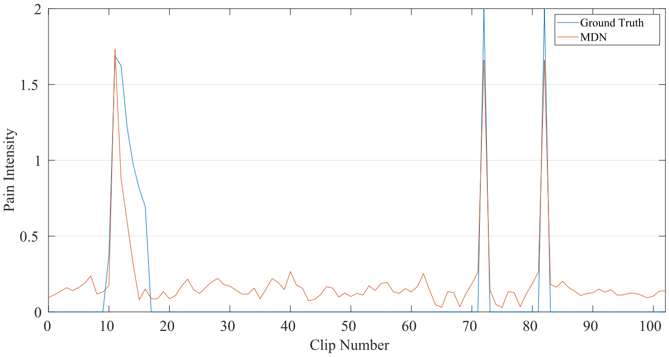


Fig. 6. Visualization of pain estimation using MDN-152 on a subject in UNBC dataset.

employs only 52 million parameters. Our approach achieves competitive results when compared with the method in [60], but that method uses 586.8 million parameters, which means more than 11 times the number of parameters of MDN, demonstrating the efficiency of MDN to explore the spatio-temporal information in other related problems.

In order to show the capabilities of our architecture in estimating different intensities of pain, Fig. 6 shows the ground truth and the predictions of MDN on a video of a subject. As shown, MDN detects the intensities of pain in a satisfactory way, and follows the different transitions of levels of pain. These results indicate that MDN has good potential to explore face expression variations related to pain.

6 Conclusion

We presented the Maximization and Differentiation Network (MDN) for encoding spatio-temporal variations of face videos for automatic depression detection. The proposed method is composed of a maximization block to model smooth facial expression variations and a difference block to encode sudden facial variations. The combination of these blocks forms the MDN module which explore multiple temporal information without 3D convolutions. We incorporated our MDN module in 3D ResNet-type architectures to generate our novel MDN architecture. We evaluated the performance of the proposed method on the two benchmark datasets for depression detection from facial videos, namely, AVEC2013 and AVEC2014. The experiments demonstrated the improvement in performance against 3D ResNet as well as T3D and I3D models. Our architecture also outperformed the state-of-the-art approaches for depression detection. As a future work, we intend to investigate other complementary modalities (e.g., audio and video-based biosignals), integrating these signals in our proposed architecture in order to further improve the performance of the model.

References

[1] J. Thevenot, M. B. López, and A. Hadid, "A survey on computer vision for assistive medical diagnosis from faces," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1497–1511, Sep. 2018.

[2] G. Giannakakis *et al.*, "Stress and anxiety detection using facial cues from videos," *Biomed. Signal Process. Control*, vol. 31, pp. 89–101, 2017.

[3] 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, Oct.–Dec. 2019.

[4] M. Murphy and M. Peterson, "Sleep disturbances in depression," *Sleep Med. Clin.*, vol. 10, pp. 17–23, 2015.

[5] S. Borgman, I. Ericsson, E. K. Clausson, and P. Garmy, "The relationship between reported pain and depressive symptoms among adolescents," *J. Sch. Nurs.*, vol. 36, pp. 87–93, 2018.

[6] J. P. Lépine and M. Briley, "The epidemiology of pain in depression," *Hum. Psychopharmacol.*, vol. 19, no. S1, pp. S3–S7, 2004.

[7] M. Jansson-Fr öjmark and K. Lindblom, "A bidirectional relationship between anxiety and depression, and insomnia? A prospective study in the general population," *J. Psychosomatic Res.*, vol. 64, pp. 443–449, 2008.

[8] Z. A. E. Sarhan, H. A. E. Shinnawy, M. E. Eltawil, Y. Elnawawy, W. Rashad, and M. S. Mohammed, "Global functioning and suicide risk in patients with depression and comorbid borderline personality disorder," *Neurol., Psychiatry Brain Res.*, vol. 31, pp. 37–42, 2019.

[9] J. L. Sotelo and C. B. Nemeroff, "Depression as a systemic disease," *Personalized Med. Psychiatry*, vol. 1-2, pp. 11–25, 2017.

[10] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*., Washington, DC, USA: American Psychiatric Publishing, 2013.

[11] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: A meta-analysis," *Lancet*, vol. 374, pp. 609–619, 2009.

[12] M. J. Bostwick, "Recognizing mimics of depression: The '8 Ds'," *Curr. Psychiatry*, vol. 11, pp. 31–36, 2012.

[13] W. Gaebel and W. Wölwer, "Facial expressivity in the course of schizophrenia and depression," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 254, pp. 335–342, 2004.

[14] J. Joshi, R. Goecke, G. Parker, and M. Breakspear, "Can body expressions contribute to automatic depression analysis?," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit*, 2013, pp. 1—7.

[15] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Trans. Affec. Comput.*, vol. 9, no. 4, pp. 578–584, Oct.–Dec. 2018.

[16] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 542–552, Jul.–Sep. 2020.

[17] X. Zhou, P. Huang, H. Liu, and S. Niu, "Learning content-adaptive feature pooling for facial depression recognition in videos," *Electron. Lett.*, vol. 55, no. 11, pp. 648–650, 2019.

[18] M. Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 262–268, Jan.–Mar. 2021.

[19] W. C. de Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3d networks for detecting depression from facial expressions," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2019, pp. 1–8.

[20] W. C. de Melo, E. Granger, A. Hadid, "Depression detection based on deep distribution learning," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 4544–4548.

[21] A. Jan, H. Meng, Y. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 3, pp. 668–680, Sep. 2018.

[22] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: Spatiotemporal and motion encoding for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2000–2009.

[23] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6201–6210.

[24] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4724–4733.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[26] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D-CNNs retrace the history of 2D CNNs and ImageNet?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6546—6555.

[27] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

[28] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 305–321.

[29] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5534–5542.

[30] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, and F. Meriaudeau, "Designing a framework for assisting depression severity assessment from facial image analysis," in *Proc. IEEE Int. Conf. Signal Image Process. Appl.*, 2015, pp. 578–583.

[31] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an affective interface for assessment of psychological distress," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 539–545.

[32] M. F. Valstar *et al.*, "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emot. Challenge*, 2013, pp. 3–10.

[33] M. Valstar *et al.*, "AVEC 2014: 3D dimensional affect and depression recognition challenge", in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 3–10.

[34] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression", in *Proc. 3rd Int. Conf. Pattern Recognit. Appl. Methods*, 2014, pp. 671–678.

[35] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. 3rd ACM Int. Workshop Audio/visual Emot. Challenge*, 2013, pp. 21–30.

[36] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: A multimodal approach," in *Proc. 3rd ACM Int. Workshop Audio/visual Emot. Challenge*, 2013, pp. 11–20.

[37] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 7, pp. 1432–1441, Jul. 2015.

[38] H. Kaya and A. A. Salah, "Eyes whisper depression: A CCA based multimodal approach," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 961–964.

[39] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 73–80.

[40] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, 2014, pp. 19–26.

[41] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 255–259.

[42] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proc. 4th Int. Workshop Audio/Visual Emot. Challenge*, 2014, pp. 87–91.

[43] Y. Kang, X. Jiang, Y. Yin, Y. Shang, and X. Zhou, "Deep transformation learning for depression diagnosis from facial images," in *Proc. Chinese Conf. Biometric Recognit.*, 2017, pp. 13–22.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[45] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.2970712.

[46] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emot. Challenge*, 2016, pp. 3–10.

[47] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 67–74.

[48] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, pp. 1499–1503, Oct. 2016.

[49] A. Diba *et al.*, "Temporal 3D ConvNets: New architecture and transfer learning for video classification," 2017, *arXiv:1711.08200*.

[50] X. Ma, D. Huang, Y. Wang, and Y. Wang, "Cost-sensitive two-stage depression prediction using dynamic visual clues," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 338–351.

[51] W. C. de Melo, E. Granger, and M. B. Lopez, "Encoding temporal information for automatic depression recognition from facial analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2020, pp. 1080–1084.

[52] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[53] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep Networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[54] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, 2011, pp. 57–64.

[55] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *Proc. Int. Symp. Visual Comput.*, 2012, pp. 368–377.

[56] C. Florea, L. Florea, and C. Vertan, "Learning pain from emotion: Transferred HoT data representation for pain intensity estimation," in *Comput. Vis.-ECCV Workshops*, 2014, pp. 778–790.

[57] R. Zhao, Q. Gan, S. Wang, and Q. Ji, "Facial expression intensity estimation using ordinal information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3466–3474.

[58] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 1535–1543.

[59] P. Rodriguez *et al.*, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2017.2662199.

[60] M. Tavakolian and A. Hadid, "A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics," *Int. J. Comput. Vis.*, vol. 127, pp. 1413–1425, 2019.

Wheidima Carneiro de Melo (Student Member, IEEE) was born in Manaus, AM, Brazil, in 1983. He received the BSc degree from the Federal University of Amazonas (UFAM), and the MSc degree in electrical engineering from UFAM in 2014. He is currently working toward the PhD degree in computer science and engineering with the University of Oulu. Since 2013, he has been with the Superior School of Technology, Amazonas State University, as a lecturer. His research interests include affective computing, computer vision, machine learning, and digital signal processing.

Eric Granger (Member, IEEE) received the PhD degree in EE from Ecole Polytechnique de Montréal in 2001. From 1999 to 2001, he was a defense scientist with DRDC, Ottawa, and from 2001 to 2004, with R&D with Mitel Networks. In 2004, he joined the Deptartment of Systems Engineering, École de technologie supérieure, Université du Québec, Montreal, Canada, where he is currently a full professor and the director of LIVIA, a research laboratory focused on computer vision and artificial intelligence. His research interests include pattern recognition, machine learning, computer vision, and computational intelligence, with applications in affective computing, biometrics, face recognition, medical image analysis, and video surveillance.

Miguel Bordallo López He received the master"s and PhD degrees from the University of Oulu, in 2010 and 2014, respectively. He studied telecommunication engineering with the Technical University of Madrid, Spain. In 2006, he joined the Center for Machine Vision and Signal Analysis, University of Oulu, where he currently holds the title of docent in real-time imaging-based sensing. He is currently a senior researcher with the VTT Data-Driven Life Group. He has authored more than 35 scientific publications. His current research interests include embedded and healthcare AI. He is an associate editor for the *Journal of Real-Time Image Processing.* He was the 2017 Nokia Bell Labs Prize Finalist.