# A Deep Multiscale Spatiotemporal Network for Assessing Depression from Facial Dynamics

Wheidima Carneiro de Melo, *Student Member, IEEE,* Eric Granger, *Member, IEEE,* and Abdenour Hadid, *Senior Member, IEEE*

**Abstract**—Recently, deep learning models have been successfully employed in many video-based affective computing applications (e.g., detecting pain, stress, and Alzheimer's disease). One key application is automatic depression recognition – recognition of facial expressions associated with depressive behaviour. State-of-the-art deep learning algorithms to recognize depression typically explore spatial and temporal information individually, by using 2D convolutional neural networks (CNNs) to analyze appearance information, and then by either mapping facial feature variations or averaging the depression level over video frames. This approach has limitations in terms of its ability to represent dynamic information that can help to accurately discriminate between depression levels. In contrast, models based on 3D CNNs allow to directly encode the spatio-temporal relationships, although these models rely on temporal information with fixed range and single receptive field. This approach limits the ability to capture variations of facial expression with diverse ranges, and the exploitation of diverse facial areas. In this paper, a novel 3D CNN architecture – the Multiscale Spatiotemporal Network (MSN) – is introduced to effectively represent facial information related to depressive behaviours from videos. The basic structure of the model is composed of parallel convolutional layers with different temporal depths and sizes of receptive field, which allows the MSN to explore a wide range of spatio-temporal variations in facial expressions. Experimental results on two benchmark datasets show that our MSN architecture is effective, outperforming state-of-the-art methods in automatic depression recognition.

**Index Terms**—Affective Computing, Depression Detection, Deep Learning, 3D Convolution Neural Network, Face Analysis, Spatiotemporal Expression Recognition, Multiscale Processing.

## 1 Introduction

The use of computer vision and machine learning techniques for automatic diagnosis is an emerging area in healthcare and medicine fields, since such techniques can provide an unobtrusive and objective information about a patient's state. Indeed, technologies that can accurately recognize the affective state of an individual using contact-free sensors can represent a powerful tool, providing personalized diagnostics and therapeutic treatment plans. These techniques commonly recognize facial patterns, which act as a mirror of health condition, since certain medical states change appearance and/or expression of the face [1]. For example, schizophrenia symptoms can be predicted using a Deep Neural Network (DNN) for analysis of facial expressions [2], and Amyotrophic Lateral Sclerosis (ALS) can be detected using hand-crafted features by analyzing the facial movements [3].

Among the various applications of automatic medical diagnosis from faces, detection of Major Depressive Disorder (MDD) has received attention from the scientific community because such disorder is one of the most common and costly mental disorders. MDD, also referred to as depression, is interpreted as negative condition of mind that remains for a long period. The symptoms related to depression include pessimism, sadness, irritability, diminution of pleasure, fatigue, insomnia, weight problems, lack of concentration, feeling hopeless, feeling worthless, anxiety, low self-esteem, and, in more grave cases, depression leads to suicide and substance abuse [4], [5], [6]. Moreover, depression may increase the risks of acquiring and sometimes contribute to advance of severe clinical states, such as diabetes, cardiovascular disease, and cancer [7].

Normally, pharmacologic, cognitive behavioral therapy, and interpersonal therapy are effective treatments for MDD. However, there are frequent errors in clinical evaluation of MDD. Indeed, depression evaluation is based on Diagnostic and Statistical Manual of Mental Disorders (DSM-5) specifications [8], which are identified during structured clinical interview. The intensity of depression can be verified by using self-report inventory, such as Beck Depression Inventory (BDI), or an inventory like Hamilton Depression Rating (HAM-D), which is administered by a clinician experienced in treating psychiatric patients. A high number of false-positive rate has been presented by some studies, with latent serious consequences, including death of patient [9], [10]. Furthermore, the clinical evaluations normally are time-consuming and require considerable physician experience.

Given the challenges in the diagnosis of depression, the computer vision community has been investigating methods for accurate estimation of patient's level of depression based on patterns of nonverbal behaviour. Studies have shown a series of nonverbal manifestations, such as psychomotor delay, which can convey information about the level of depression [11], [12], [13]. In fact, subjects with depression present gloomy and sad facial expression [14],

- *W.C. de Melo and A. Hadid are with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. E-mail: wheidima.melo@oulu.fi, hadid.abdenour@oulu.fi*
- *E. Granger is with the Laboratoire d'imagerie, de vision, et d'intelligence artificielle, Dept. of Systems Engineering, École de technologie supérieure, Université du Québec, Montreal, Canada. E-mail: eric.granger@etsmtl.ca*

*Manuscript received August 28, 2019; revised April 19, 2020.*

and evidence low levels of social behaviors, such as less facial movements, few body and hand gestures, limited eye contact, and absence of smiles [4]. The automated diagnosis of depression using facial information explores these spatial or spatio-temporal features captured in images or videos.

This paper focuses on techniques for accurate assessment of depression levels based on facial expression captured in videos. This is considered to be a challenging recognition problem and continues to drive much academic research. A key challenge in real-world scenarios is the significant variations over time of the facial expressions for different persons, sensors, computing devices and operational environments. Despite the recent advancement of various facial and speech-based technologies, developing an efficient system for expression recognition remains a challenging task [15].

Most of previous work on automatic depression assessment focuses on extracting discriminant features from facial regions captured in video frames to assess the intensity of depression. State-of-the-art machine learning models exploit spatial and temporal information separately, using a 2D Convolutional Neural Network (CNNs) for feature extraction, and some scheme for map the variation of the features, or for averaging the level of depression in each face frame [16], [17], [18], [19], [20]. These approaches represent facial regions in video frames based on spatial features, which limit their ability in encoding rich dynamic information required for depression level estimation. To improve estimation accuracy, some authors have employed 3D CNNs, like the Convolutional 3D (C3D) [21] to leverage spatio-temporal information [22], [23]. However, these methods exploit temporal information from video in single range, and the facial expression variations occur in wide range. Moreover, this approach employs structures with fixed receptive field which may impair the exploitation of different facial areas.

In this paper, we address these challenges by introducing a new 3D convolutional architecture for accurate depression detection. The proposed model – called Multiscale Spatiotemporal Network (MSN) – can directly leverage the spatio-temporal dependencies and dynamics of facial structures. As the manifestations of depression indicate facial dynamics that are comprised of short to long range temporal information, our model explores different temporal ranges in order to efficiently capture the facial dynamics related to depression. In addition, our model employs several receptive fields in order to maximize the exploitation of distinct spatial areas, since different areas of the face convey diverse information about depression levels [18], [23]. These characteristics allow our proposed model to explore multiscale spatio-temporal features within an end-to-end learning strategy. To validate the proposed approach, we compared the MSN to various state-of-the-art (conventional and deep learning) models for automatic depression recognition in terms of accuracy and computational complexity, using videos from the Audio Visual Emotion Challenge (AVEC 2013 and 2014) datasets. Code is available at https://github.com/wheidima/MSN.

The remainder of this paper is organized as follows. Section 2 provides some background on models for depression detection. In Section 3, the proposed MSN is described. The experimental methodology is defined in Section 4, and the results and analysis are presented in Section 5.

## 2 Related Work

In the recent years, there has been a growing interest in automatic depression assessment from facial information. The Audio-Visual Emotion Challenge and Workshop in the years of 2013 [24] and 2014 [25] (AVEC2013 and AVEC2014) has contributed notably for researching on depression detection. These events had as part of competition the task that required participants to predict the level of self-reported depression in each video. The datasets used by the participants are called AVEC2013 and AVEC2014 datasets and are made available for research purposes. The mentioned datasets are one of few datasets that provide raw data (video and audio) information, whereas other datasets only make available features of subjects [4].

In the AVEC2013 challenge [24], the competition provided baseline system to process visual and audio data. The visual features are obtained by using a popular local descriptor namely Local Phase Quantisation (LPQ) [26], and Support Vector Regressor (SVR) [27] is employed to estimate the depression levels. In [28], Meng *et al.* employed Motion History Histogram [29] to capture motion information of facial expressions. Cummins *et al.* [30] investigated Space-Time Interest Points (STIP) [31] and Pyramid of Histogram of Gradients (PHOG) [32] descriptors. Wen *et al.* [33] proposed to extract dynamic features based on LPQ from Three Orthogonal Planes (LPQ-TOP). In the AVEC2014 challenge [25], the baseline visual features are obtained by employing Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [34] which combines dynamic and spatial texture analysis with Gabor filtering. In [35], the authors calculated variations of eye and face positions, combined with motion information, then employed SVR method. Jan *et al.* [36] extracted three distinct texture feature representations, and predicted the depression levels using partial least square [37] and linear regression technique. Finally, the authors in [38] calculated canonical correlation analysis on LPQ and baseline features to estimate a continuous depression levels.

The traditional depression detection schemes described previously have primarily been focused on hand-engineered representations. More recently, deep learning techniques have been employed to model depressive patterns. Such techniques have produced discriminant feature representations, achieving state-of-the-art results in depression recognition. In one of the first works using deep learning, Zhu *et al.* [16] proposed a two-stream CNN to capture facial appearance and dynamics, with one channel inputs facial areas, and the second one inputs facial flows. Two fully connected layers perform the fusion of the features and estimate the depression level. Jan *et al.* [17] extracted visual features using Visual Geometry Group (VGG) architecture [39] from facial images. In order to model the temporal movement on the visual feature space, the authors employed Feature Dynamic History Histogram (FDHH). In [18], Zhou *et al.* employed deep learning model with Global Average Pooling (GAP) to explore various facial areas with a scheme to combine the response from distinct facial areas. This

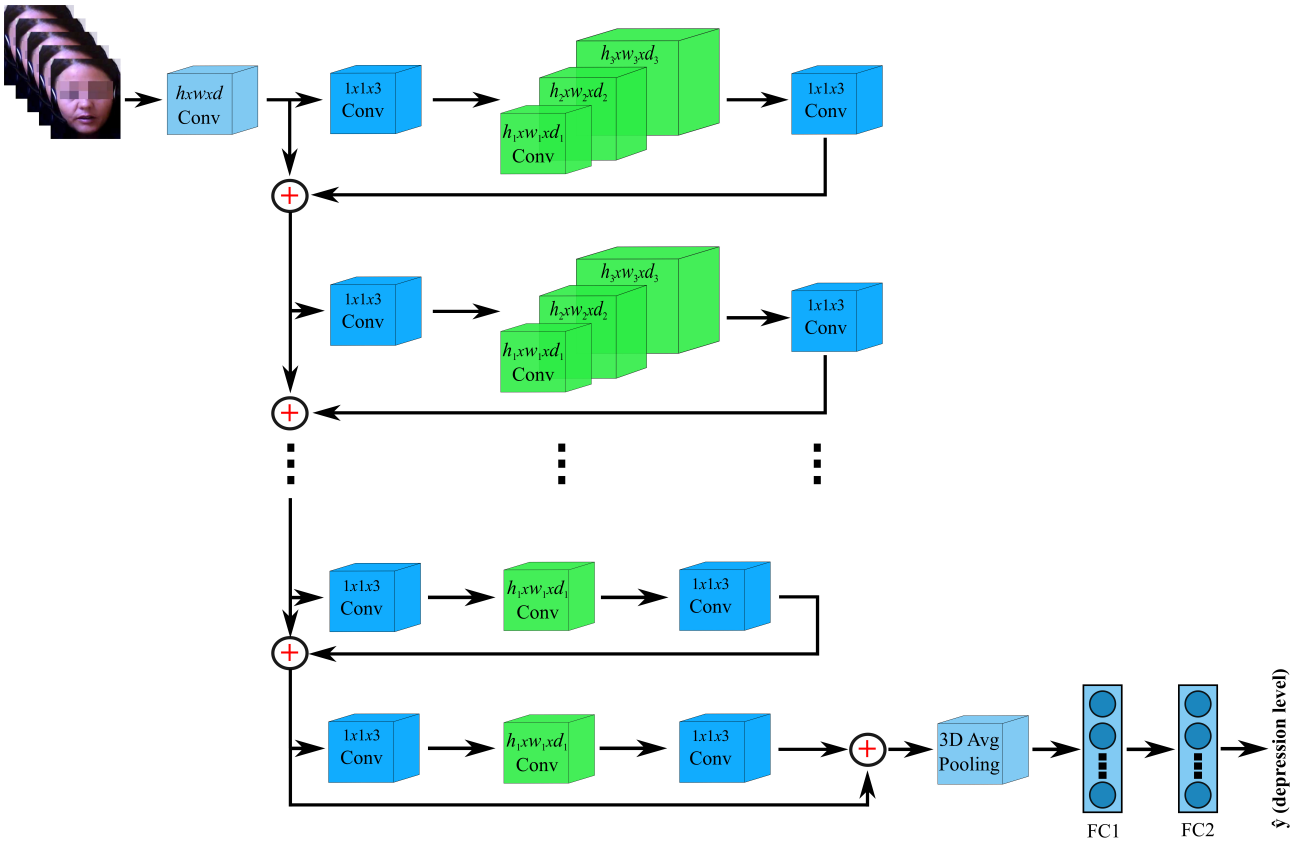



Fig. 1. Block diagram of the proposed Multiscale Spatiotemporal Network (MSN) for automatic depression assessment.

model showed that some facial regions are more important than others for depression analysis. Finally, Melo *et al.* [19] presented a deep learning method for estimation of depression levels from facial frames through distribution learning, using a new expectation loss function.

All these deep learning methods consider the temporal information by integrating, in different ways, the visual features extracted from video frames. Such an approach generates difficulties in describing significant dynamic information frequently necessary for robust depression recognition. Some attempts have been made to explore directly the spatial and temporal information using 3D convolutional neural network. Jazaery *et al.* [22] proposed to employ C3D method to produce spatio-temporal features from facial videos at two different scales, and a Recurrent Neural Network (RNN) to model transitions of the features. In [23], the authors proposed to extract spatio-temporal features from global and local facial regions. The local region refers to a coarse eye region whereas global region is full-face region. The depression level is defined by the fusion of predictions from a C3D trained on the global region and C3D trained on local region. Moreover, the models obtain good results exploring a certain local region, but the performance remains approximately the same after the models combine the response of other facial regions, demonstrating difficulties to explore spatial structures in different facial regions.

Some authors propose methods to estimate depression level from features or human behaviours. Such schemes may analyze facial landmarks, head pose, gaze direction, action units, hand-crafted features and deep learning representations [40], [41]. In [42], the authors proposed two frameworks for depression detection from human behaviour primitives. The first explores statistics of this information while the second uses 2D CNN to leverage spectral representation of the behaviour signals. Haque *et al.* [43] employ Causal Convolutional Neural Network (C-CNN) to analyze 3D videos of facial landmarks. In [44], Du *et al.* employed Temporal Convolutional Network (TCN) [45] and atrous convolutions [46] to learn long-term representation for depression detection from visual behaviours. In contrast to these methods, our proposed approach directly explores the spatio-temporal dependencies, without further method such as human behaviour detector, which may benefit the method since it is not limited to model features or behaviour primitives. Moreover, our method captures the spatio-temporal information at various scales in order to improve the depression representation whereas generally these methods model just the temporal information using different ranges.

Besides C3D network, some 3D CNNs have been presented to map spatial and temporal information. Such architectures primarily employ convolutional layers with fixed temporal depth. Carreira *et al.* proposed the Inflated 3D Convolutional Network (I3D) [47]. The method is based on Inception-v1 model [48]. The 2D convolutional filters and pooling kernels employed by Inception-v1 model are inflated into 3D framework. In [49], the authors employed ResNet expended to 3D structures. Tran *et al.* [50] proposed to decompose 3D convolutional filters into distinct spatial and temporal filters. In summary, these architectures mostly

employ 3D kernels and pooling layers with fixed depth (dimension related to temporal information) and receptive field for all layers of model. Using such methodology, the models decrease the potential of capturing spatio-temporal information with different sizes, regarding to depressive behaviours. We address these limitations by including diverse spatio-temporal kernels in the proposed architecture for depression recognition.

## 3 Multiscale Spatiotemporal Network

A patient suffering from depression exhibits spatio-temporal alterations in his/her facial information. The spatial information is related to facial expressions, texture and structures. The movement over time constitutes spatial changes and deformations which encompasses facial dynamics of the subject. From this perspective, the aim of this paper is to capture the facial dynamics that incorporate plenty of the substantial information for automatic depression detection.

Deep learning architectures based on 3D CNNs have the potential to encode and leverage spatio-temporal information from facial videos. Such networks are comprised of 3D filters and pooling layers which are trained to learn spatio-temporal features. The depth of the filters determines the range of temporal information that can be explored, and the spatial size defines the area of input that will be analyzed. In this work, we develop a Multiscale Spatiotemporal Network (MSN) which incorporates various 3D convolutional filters with different temporal depths and spatial size in the basic building block. In Figure 1, an overview of our proposed method is presented.

The basic building block employs identity shortcut connections to connect the input of each 3D block to its output features. The residual connection is adopted as it enables the training of very deep networks at the same time that decreases problems of overfitting [51]. Two 3D convolutional layers with fixed spatial size and temporal depth, and various 3D convolutional layers with different spatial size and temporal depths constitute the basic block. Figure 2 illustrates the basic building block. The 3D convolutional filters have the depth in the range of $d \in \{d_1, d_2, d_3\}$, and spatial size with dimensions equal to $h \times w$, where $h \in \{h_1, h_2, h_3\}$ and $w \in \{w_1, w_2, w_3\}$. In this way, the proposed basic block has the ability to capture an extensive spatio-temporal information that encode depressive behaviours.

The predicted output of the basic building block can be defined as:

$$\bar{y} = \sigma(BN(\mathcal{H}(x, \{H_i\}_{i=1}^{M})) + x) \quad (1)$$

where $x$ is the input of the basic block, $\mathcal{H}(.)$ is the function that learns the residual mapping, BN stands for Batch Normalization, $\sigma$ is the activation function called Rectified Linear Unit (ReLU), $\{H_i\}_{i=1}^{M}$ denotes the parameters of the convolutional layers, and $M$ is the total of convolutional layers. The function $\mathcal{H}(.)$ is determined by (see also Figure 2):

$$\mathcal{H}(x) = H_5 \left( \bigcup_{i=2}^{M-1} \sigma(BN(H_i \sigma(BN(H_1 x)))) \right) \quad (2)$$

where $\bigcup$ represents the operation of concatenation, $H_1$ is the weight parameters of the first convolutional layer, whereas $H_5$ represents parameters of the last convolutional layer, hence $M$ is equal to 5.

The first convolutional layer ($H_1$) of the basic building block is defined with convolution kernel of $1 \times 1 \times 3$. This layer receives features map ($x$) of previous layer and generates output which is fed into next three parallel convolutional layers. The convolution kernels of $h_1 \times w_1 \times d_1$, $h_2 \times w_2 \times d_2$, and $h_3 \times w_3 \times d_3$ constitute the parallel network. We define $h_1 = w_1 = d_1$, $h_2 = w_2 = d_2$, and $h_3 = w_3 = d_3$. In the next stage, the outputs of parallel convolutional layers are concatenated, increasing the depth of resulting features. For this reason, we apply the convolution kernel of $1 \times 1 \times 3$ of the last convolutional layer ($H_5$) to control the depth of the features (this is also the motivation to employ the same kernel size in the first convolutional layer). Moreover, ReLU activation and batch normalization are employed after every convolutional layer, with exception of $H_5$ layer which only applies batch normalization.

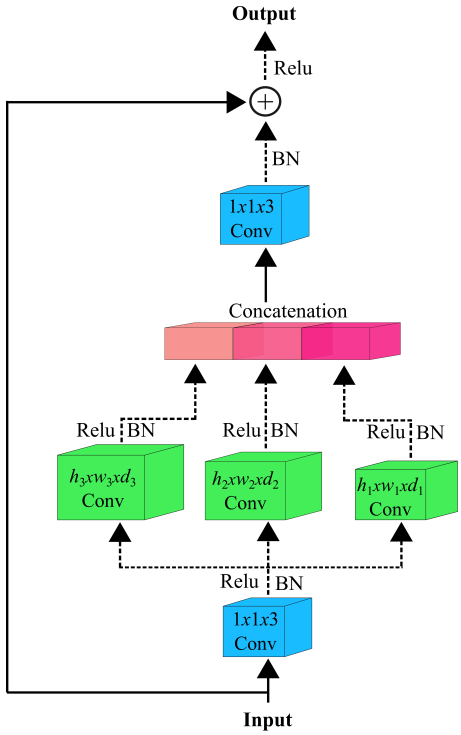



Fig. 2. A diagram of our proposed basic building block. The dashed line indicated an additional operation. BN refers to Batch Normalization.

In order to sum the resulting features of $H_5$ convolutional layer with $x$, the channels of such features should be equal to channels of $x$. The number of channels of the feature maps is defined by the number of filters in the convolutional layer, consequently the number of channels of $x$ and $H_5$ output can be different, which impedes the sum operation being performed. To prevent this problem, we insert a convolutional layer with kernel size of $1 \times 1 \times 3$, and batch normalization in the residual connection. After the sum operation, ReLU activation is performed to generate the final feature map ($y$) of the basic building block.

A complete description of each layer of the proposed MSN method is presented in Table 1. In the first layer

TABLE 1
A description of the proposed MSN architecture. Channels refer to the number of filters employed in the layer.

| Layer | Spatial Output | Kernel Size | Channels | Number of Layers | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 27 | 36 | 51 | 69 | 99 |
| Conv1 | $56 \times 56$ | $7 \times 7 \times 7$ | 64 | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| 3D Max Pooling | $28 \times 28$ | $3 \times 3 \times 3$ | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Conv2 | $28 \times 28$ | $1 \times 1 \times 3$<br>$h_1 = 3, w_1 = 3, d_1 = 3$<br>$h_2 = 5, w_2 = 5, d_2 = 5$<br>$h_3 = 7, w_3 = 7, d_3 = 7$<br>$1 \times 1 \times 3$ | 32<br>32<br>32<br>32<br>128 | $\times 2$ | $\times 3$ | $\times 3$ | $\times 3$ | $\times 3$ |
| Conv3 | $28 \times 28$ | $1 \times 1 \times 3$<br>$h_1 = 3, w_1 = 3, d_1 = 3$<br>$h_2 = 5, w_2 = 5, d_3 = 5$<br>$h_3 = 7, w_3 = 7, d_3 = 7$<br>$1 \times 1 \times 3$ | 64<br>64<br>64<br>64<br>256 | $\times 2$ | $\times 3$ | $\times 4$ | $\times 4$ | $\times 4$ |
| 3D Max Pooling | $14 \times 14$ | $2 \times 2 \times 2$ | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Conv4 | $14 \times 14$ | $1 \times 1 \times 3$<br>$h_1 = 3, w_1 = 3, d_1 = 3$<br>$1 \times 1 \times 3$ | 256<br>256<br>512 | $\times 2$ | $\times 3$ | $\times 6$ | $\times 12$ | $\times 22$ |
| 3D Max Pooling | $7 \times 7$ | $2 \times 2 \times 2$ | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Conv5 | $7 \times 7$ | $1 \times 1 \times 3$<br>$h_1 = 3, w_1 = 3, d_1 = 3$<br>$1 \times 1 \times 3$ | 512<br>512<br>1024 | $\times 2$ | $\times 3$ | $\times 3$ | $\times 3$ | $\times 3$ |
| 3D Avg Pooling | $1 \times 1$ | $7 \times 7 \times 2$ | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Fully Connected | $1 \times 512$ | - | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Dropout (50%) | $1 \times 512$ | - | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Fully Connected | $1 \times 512$ | - | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Dropout (50%) | $1 \times 512$ | - | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |
| Regression | $1 \times 1$ | - | - | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ | $\times 1$ |

(Conv1), the convolution is performed with spatial stride 2 and temporal stride 1, downsampling spatially the input of the network by factor of two. Max-pooling layers with kernel size $2 \times 2 \times 2$ and spatio-temporal stride 2 are applied between Conv3, Conv4 and Conv5 which means the size of output feature is reduced spatio-temporally by a factor of 2 in comparison with the layer input feature. A Max-pooling layer with kernel size $3 \times 3 \times 3$ is likewise placed before Conv2 to perform spatio-temporal downsampling. Since along the model the spatio-temporal information is reduced, we change the structure of basic building block, removing the convolutional layers with kernel size $h_2 \times w_2 \times d_2$ and $h_3 \times w_3 \times d_3$. In this way, the model can efficiently explore the feature maps related to these layers and it also contributes to control the number of parameters of the architecture. Finally, the stage of classification is composed by 3D average pooling layer, two fully connected layers with 512 neurons and a regression output layer which generates depression level scores.

For depression estimation, the model should have the ability to predict a continuous value. Thus, the regression loss function of the proposed MSN is Mean Squared Error (MSE). For a given training sample (clip) $n$, the MSE is determined by computing the Euclidean distance between the estimated output prediction $\hat{y}_n$ and ground truth value $y_n$. According to this distance, the loss function of the proposed method is given by:

$$E = \frac{1}{2N} \sum_{n=0}^{N-1} (\hat{y}_n - y_n)^2 \tag{3}$$

where $N$ is the number of samples.

## 4 Experimental Setup

### 4.1 Datasets:

In order to evaluate and compare the ability of our proposed MSN to predict the depression levels of subjects, extensive experiments are conducted on two publicly-available benchmark datasets namely Audio-Visual Emotion Challenge 2013 and 2014 (AVEC2013 and AVEC2014) depression sub-challenge datasets. As our proposed method is designed to explore spatial and temporal information, we decided to use these two datasets because they are the only ones to provide the raw video information.

The datasets were used in the AVEC sub-challenge in which the goal was to predict score of subjects on the Beck Depression Inventory (BDI). Twenty one questions compose the BDI-II, and a scale between 0 and 3 is used to score every question. The BDI scores can be classified into four severity levels: minimal $(0-13)$, mild $(14-19)$, moderate $(20-28)$, and severe $(29-63)$.

**AVEC2013 Dataset**. The AVEC2013 depression dataset is a subset of the audio-visual depressive language corpus (AViD-Corpus), containing 150 videos from 82 individuals. During a Human-Computer Interaction task, the individual is recorded using a webcam and microphone. The dataset is organized into three distinct partitions: training, development and test sets. Each partition contains 50 videos which have a label corresponding the depression level of a subject. The longest video reaches 50 minutes in duration, and the shortest lasts 20 minutes. The average depression level is 15.1 and 14.8 for training and development sets, respectively.

**AVEC2014 Dataset**. The AVEC2014 depression dataset also uses a subset of AViD-Corpus. The individuals are recorded using a webcam and microphone. During acquisition of

Fig. 3. Samples from AVEC2013 and AVEC2014 datasets.

the videos, the subjects perform two tasks: Freeform and Northwind tasks. In the first, the subjects respond to questions like discuss a sad childhood memory. In the second one, subjects read audibly an excerpt from a fable. In both activities, the recordings are segmented into three partitions: training, development and test set. Each partition contains $50$ videos which have a ground truth label. In total, there are $300$ videos ranged in duration between $6$ and $248$ seconds. We perform experiments employing training and development sets from both tasks as training data, and the test sets are used to measure the performance of the model. Some samples from both datasets are shown in Figure 3. For privacy concerns, all the samples shown in this paper are blurred.

### 4.2 Settings:

The proposed method explores appearance and temporal information from facial videos. As the videos in the datasets contain more information than only facial expressions, the first step of our proposed method is face pre-processing. This step performs detection and alignment of the faces captured in videos, providing frontal face regions. Following the methods in [19], [23], the Multi-Task Cascade Convolutional Network (MTCNN) [52] is chosen to simultaneously detect and align the faces. The MTCNN includes the proposal network (P-Net), refinement network (R-Net) and output network (O-Net). The P-Net and R-Net produce and examine candidate windows as well as remove non-face windows. The O-Net defines the bounding box and five facial landmarks which are employed to face detection and alignment. The facial images that are generated in this procedure have dimensions of $112 \times 112$. This process is performed for all video frames of the AVEC2013 and AVEC2014 datasets.

Due to substantial number of frames in the samples and redundant temporal information, the video samples in AVEC2013 and AVEC2014 are usually downsampled [16], [18], [19], [22], [23]. In our approach, we temporally subsampling both datasets by a factor of $8$. With that, we explore the same spatio-temporal distribution in both datasets.

**Training the model**. To train the proposed model, we randomly chose a frame inside the video and collect the subsequent frames to make a training clip. If the selected temporal position does not allow defining a clip, we loop the video. We empirically define the size of clip for the proposed method.

As it is well known, data augmentation is very important for learning deep neural networks. In the process of training, the frames are horizontally flipped with $50\%$ probability, randomly rotated with $30$ degrees, and top-to-bottom rotated. All produced training samples retain the identical depression level as their original videos.

Because of the limited data in AVEC2013 and AVEC2014 datasets to train a deep model from scratch, the proposed MSN architecture is initially trained on face recognition. The model is pre-trained on VGGFace2 dataset that contains $3.31$ million images of $9,131$ identities [53]. An input is formed by replicating an image in accordance with the number of frames in a clip. In the training stage, Stochastic Gradient Descent (SGD) with momentum of $0.9$, weight decay $0.0001$, batch size of $24$, and initial learning rate of $0.01$ is adopted. The learning rate is multiplied by $0.1$ after every $10$ epochs. Additionally, the input values are subtracted by the average value of VGGFace2. The fine-tuning process is employed in the MSN architecture using ADAM optimization algorithm with decaying learning rate, where the initial learning rate is $0.0001$. The proposed method is implemented using Keras framework [54] in Nvidia Tesla T4 GPU.

**Testing the model**. For the testing stage, the input clips are generated by using sliding window, where the video inputs are segmented into non-overlapped clips. The final depression level for a subject present in a video input is determined by averaging the estimated depression scores for all the clips which constitute the video.

### 4.3 Performance Measures:

The performance of the proposed and baseline models are assessed on AVEC2013 and AVEC2014 datasets in terms of two evaluation metrics – Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The MAE is computed using:

$$MAE = \frac{1}{M}\sum_{i=0}^{M-1}|x_i - \hat{x}_i| \tag{4}$$

where $x_i$ is the ground truth for $i$th input video, $\hat{x}_i$ denotes the predicted value, and $M$ is the number of video samples. The RMSE is defined by:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=0}^{M-1}(x_i - \hat{x}_i)^2} \tag{5}$$

with identical definitions.

## 5 Results and Discussion

### 5.1 Analysis of the configuration:

To determine the optimal configuration of the proposed MSN, we start by analyzing the network size. In the sequence, we investigate different structures in the basic building block. Employing the basic building block, we develop five networks with sizes of $27$, $36$, $51$, $69$ and $99$. The details of each network are presented in Table

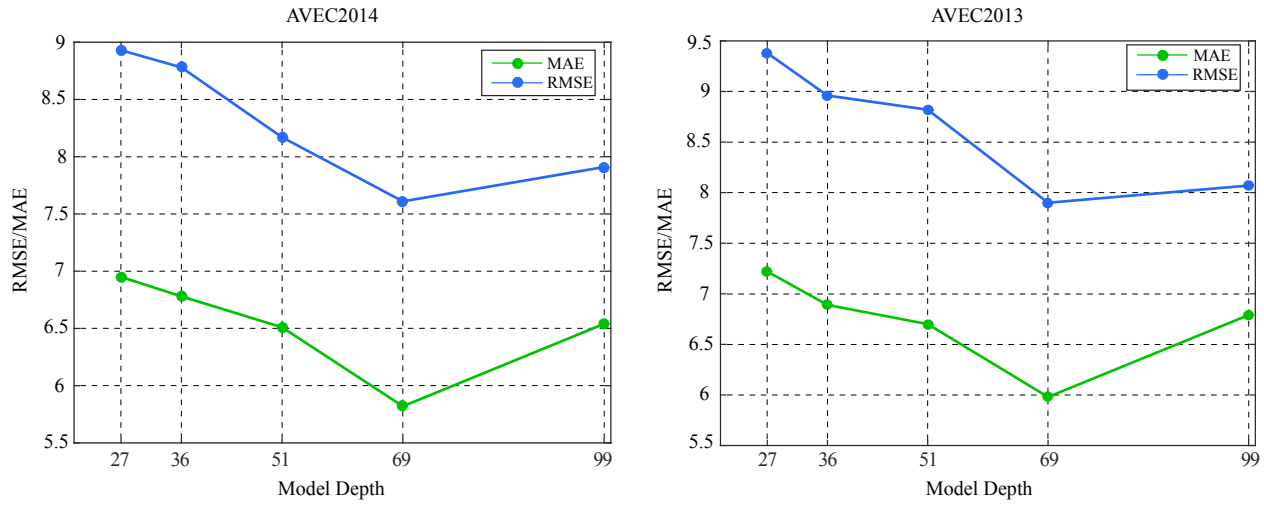



Fig. 4. The performance of the MSN architecture in terms of RMSE and MAE for different network sizes on AVEC2013 (right side) and AVEC2014 (left side) datasets.

1. As can be observed, the 3D convolutional kernels are defined by $h_1 = w_1 = d_1 = 3$, $h_2 = w_2 = d_2 = 5$, and $h_3 = w_3 = d_3 = 7$. Figure 4 shows the results for this configuration using the five networks, where we can see that the performance of the model improves, *i.e.* the value of RMSE and MAE decrease, with the increase of the network size until achieves 69 layers. For the size of 99, there is an increment of RMSE and MAE. Therefore, increasing the number of layers does not imply in improvement of the performance. The reason is a meaningful number of parameters and a limited amount of depression data.

Given that the depressive behaviours consist of wide range of spatio-temporal variations, the temporal depth, the receptive field and the number of parallel 3D structures in the basic building block are very important in capturing these variations. We investigate these three components, considering $h_1 = w_1 = d_1$, $h_2 = w_2 = d_2$, $h_3 = w_3 = d_3$ and $h_4 = w_4 = d_4$, *i.e.* the temporal depth and receptive field of 3D convolutional kernels are concurrently changed with equal values. In Table 2, the results of the proposed MSN are presented using different 3D convolutional kernels and from 1 to 4 parallel convolutional layers. The results indicate that the increment of layers for a maximum of 3 parallel kernels may contribute to improve the performance of the method. However, it is necessary to select the dimensions of the kernel carefully. For instance, the network with basic building block with three parallel layers $h_1 = 3$, $h_2 = 5$ and $h_3 = 9$ outperforms the one without parallel layers $h_1 = 3$, whereas the network using $h_1 = 3$, $h_2 = 7$ and $h_3 = 11$ increases the error in relation to network with $h_1 = 3$. Moreover, for a same number of parallel layers, it is possible to observe that structures which explore short, mid and long range obtain better results. The comparison between the network with $h_1 = 3$, $h_2 = 5$ and $h_3 = 9$ and the one with $h_1 = 3$, $h_2 = 7$ and $h_3 = 11$ is an example of this fact. The best results are achieved by using three parallel layers with $h_1 = 3$, $h_2 = 5$ and $h_3 = 7$. These results confirm that structures with different kernels may contribute to capture wide range spatio-temporal variations.

The previous discussion shows that the best results are obtained by using 69 layers and the basic building block with three parallel 3D convolutional kernels which are $h_1 = w_1 = d_1 = 3$, $h_2 = w_2 = d_2 = 5$ and $h_3 = w_3 = d_3 = 7$. For the following analysis, we use this configuration for the proposed MSN.

TABLE 2
Analysis of different 3D convolutional kernels. As $h = w = d$, we omit the terms $w$ and $d$.

| 3D Convolutional Kernels | AVEC2013 | | AVEC2014 | |
|---|---|---|---|---|
| | RMSE | MAE | RMSE | MAE |
| $h_1 = 3$ | 8.79 | 6.92 | 8.36 | 6.50 |
| $h_1 = 3\ h_2 = 5$ | 8.68 | 6.82 | 8.14 | 6.40 |
| $h_1 = 3\ h_2 = 7$ | 8.71 | 6.55 | 8.45 | 6.37 |
| $h_1 = 3\ h_2 = 5\ h_3 = 7$ | **7.90** | **5.98** | **7.61** | **5.82** |
| $h_1 = 3\ h_2 = 5\ h_3 = 9$ | 8.18 | 6.29 | 7.78 | 6.04 |
| $h_1 = 3\ h_2 = 5\ h_3 = 11$ | 8.35 | 6.63 | 7.74 | 6.16 |
| $h_1 = 3\ h_2 = 7\ h_3 = 9$ | 8.52 | 6.96 | 8.60 | 6.95 |
| $h_1 = 3\ h_2 = 7\ h_3 = 11$ | 8.82 | 6.92 | 8.72 | 6.76 |
| $h_1 = 3\ h_2 = 5\ h_3 = 7\ h_4 = 9$ | 8.31 | 6.51 | 8.03 | 6.45 |

### 5.2 Comparison with State-Of-The-Art Methods:

The performance of the proposed method is compared with other spatio-temporal methods (e.g. C3D and I3D) and with the recent state-of-the-art schemes for automatic depression recognition on the AVEC2013 and AVEC2014 datasets.

With the objective of conducting a direct and fair comparison with the state-of-the-art methods, we present results in terms of MAE and RMSE. Moreover, in order to better gain insight into the potential of MSN model, we also generate results by employing Inflated 3D Convolutional Network (I3D) [47] and Convolutional 3D (C3D). It is important to mention that I3D has been successfully applied in action recognition, demonstrating potential to capture efficiently spatio-temporal features.

**Analysis on AVEC2013**. Table 3 shows results obtained using the proposed method, I3D, and C3D architectures on AVEC2013 dataset. The I3D network produces more discriminant features than C3D, achieving better results. This difference of performance over C3D can be justified

TABLE 3
Performance of Proposed Method on AVEC2013.

| Proposed Methods | RMSE | MAE |
|---|---|---|
| C3D | 9.31 | 7.25 |
| I3D | 8.66 | 6.64 |
| MSN | **7.90** | **5.98** |

by the fact that the I3D model has a deeper network and an efficient structure called Inception module. Our MSN architecture also outperforms the C3D model, reducing MAE and RMSE by margin larger than $1.00$. This indicates that the basic building block of MSN is more efficient than the structure that employs a fixed kernel size of $3 \times 3 \times 3$. In the comparison between MSN and I3D, the MSN architecture seems to capture spatio-temporal features more efficiently than I3D model. Such result confirms that exploring the spatio-temporal dependencies using distinct ranges is very important for encoding facial expression modifications for depression detection from facial information.

TABLE 4
Performance of Proposed Method on AVEC2014.

| Proposed Methods | RMSE | MAE |
|---|---|---|
| C3D | 8.99 | 7.20 |
| I3D | 8.55 | 6.36 |
| MSN | **7.61** | **5.82** |

**Analysis on AVEC2014**. The results using I3D, MSN and C3D architectures on AVEC2014 dataset are presented in Table 4. The performance of C3D model is again the worst, showing that architectures with fixed kernel size are not efficient to capture spatio-temporal information with different sizes. Moreover, these results confirm that MSN architecture can represent a short, middle and long facial variations related to depression more efficiently than I3D. It is worth noting that I3D model employs one $7 \times 7 \times 7$ 3D convolutional layer and the other layers use basically $3 \times 3 \times 3$ kernels while MSN employs several layers with multiscale kernels that means various 3D convolution operations with distinct spatio-temporal sizes which increase in potential the exploitation of depressive facial variations.

TABLE 5
Complexity study of MSN, C3D and I3D architectures.

| Methods | Parameters ($\times 10^6$) | Time (seconds) | FLOP ($\times 10^6$) |
|---|---|---|---|
| C3D | 32.1 | 0.040 | 8.9 |
| I3D | 13.0 | 0.030 | 26.2 |
| MSN | 77.7 | 0.074 | 164.9 |

**Space and time complexity of the MSN**. In spite of the fact that deep learning algorithms are able to produce discriminative representations, the 3D convolutions tend to be computationally expensive and memory intensive. Table 5 shows the computational complexity of the proposed method MSN in comparison with C3D and I3D models. Compared to I3D, C3D has approximately $2.5$ times more parameters. It also implies that C3D needs more disk space than I3D. Despite this result, I3D method has better results in terms of RMSE and MAE when compared to C3D due to its efficient architecture. Our proposed MSN employs around $2.5$ and $6$ more parameters in comparison with C3D and I3D, respectively. This result is expected since our method has structures to explore wide range spatio-temporal information whereas C3D and I3D basically capture the spatial and temporal facial expression variations within a fixed range.

In Table 5, we also report the computation cost in terms of (1) the number of Floating Point Operations (FLOPs), and (2) the time required to predict the depression level given an input clip. We evaluated the performance of the models on Nvidia Tesla T4 GPU. When compared to I3D, the number of FLOP of the C3D is less. The reason is that the Inception module is more complex than the basic block of C3D, and I3D is a deeper network than C3D. However, I3D requires less time to estimate the output for an input clip in contrast with C3D. We understand that the Inception module has 3D convolution layers that are parallel which allows the multiprocessing systems to compute the output of each layer simultaneously, decreasing the inference time. Our proposed MSN increases the FLOP values by approximately $19$ and $7$ compared to C3D and I3D, respectively. The MSN requires $0.074$ seconds to estimate the depression level of subjects in a clip which means $1.85$ times more than C3D and $2.46$ times more than I3D. Therefore, the parallel layers of basic building block of MSN allow to generate features simultaneously like I3D, producing reasonable inference time.

TABLE 6
Comparison of Schemes for Predicting The Level of Depression on AVEC2013 Dataset.

| Proposed Methods | RMSE | MAE |
|---|---|---|
| Baseline [24] | 13.61 | 10.88 |
| LPQ + SVR (Kächele *et al.* [56]) | 10.82 | 8.97 |
| MHH + LBP (Meng *et al.* [28]) | 11.19 | 9.14 |
| PHOG (Cummins *et al.* [30]) | 10.45 | - |
| LPQ-TOP + MFA (Wen *et al.* [33]) | 10.27 | 8.22 |
| LPQ + Geo (Kaya *et al.* [58]) | 9.72 | 7.86 |
| Two DCNN (Zhu *et al.* [16]) | 9.82 | 7.58 |
| C3D (Jazaery *et al.* [22]) | 9.28 | 7.37 |
| HOG + HOF + MBH + FV (Ma *et al.* [20]) | 8.91 | 7.26 |
| C3D (Melo *et al.* [23]) | 8.26 | 6.40 |
| Four DCNN (Zhou *et al.* [18]) | 8.28 | 6.20 |
| ResNet-50 (Melo *et al.* [19]) | 8.25 | 6.30 |
| MSN (proposed) | **7.90** | **5.98** |

**Comparison on AVEC2013**. In Table 6, we show the performance of our proposed method compared with baseline and state-of-the-art methods on AVEC2013 dataset. The models in [20], [24], [28], [30], [33], [56], [58] are based on hand-engineered representations. For example, Kächele *et al.* [56] employ Local Phase Quantization (LPQ) and Support Vector Regression (SVR) to predict depression levels. The results of MSN architecture outperform all other methods. In [16], the authors proposed a method that explores temporal information (optical flow) and appearance separately. The MSN obtains better results than this method, indicating that exploring spatio-temporal information directly is more suitable for depression prediction. The authors in [22] and [23] employ two C3D models to explore different face regions. The MSN outperforms both methods, showing the power of the model in capturing spatio-temporal infor-

mation from diverse facial regions. MSN outperforms the methods in [18], [19] which only explore spatial information. Observe that the method in [18] is more efficient than C3D and I3D models (see Table 2). These results show the importance of capturing the spatio-temporal information rather than only the spatial information, but the structures of the model should have the ability to explore spatio-temporal information in different ranges.

TABLE 7
Comparison of Schemes for Predicting The Level of Depression on AVEC2014 Dataset.

| Methods | RMSE | MAE |
|---|---|---|
| Baseline [25] | 10.86 | 8.86 |
| MHH + PLS (Jan *et al.* [36]) | 10.50 | 8.44 |
| LGBP-TOP + LPQ (Kaya *et al.* [38]) | 10.27 | 8.20 |
| MHI + MSI (Espinosa *it al.* [35]) | 9.84 | 8.46 |
| DTL (Kang *et al.* [57]) | 9.43 | 7.74 |
| Two DCNN (Zhu *et al.* [16]) | 9.55 | 7.47 |
| C3D (Jazaery *et al.* [22]) | 9.20 | 7.22 |
| C3D (Melo *et al.* [23]) | 8.31 | 6.59 |
| VGG + FDHH (Jan *et al.* [17]) | 8.04 | 6.68 |
| Four DCNN (Zhou *et al.* [18]) | 8.39 | 6.21 |
| ResNet-50 (Melo *et al.* [19]) | 8.23 | 6.15 |
| MSN (proposed) | **7.61** | **5.82** |

**Comparison on AVEC2014**. Table 7 compares the results of our proposed method against the state-of-the-art on AVEC2014 dataset. The obtained results by MSN architecture generate values of MAE and RMSE lower than conventional schemes such as Local Gabor Binary Patterns from Three Orthogonal Planes with Local Phase Quantisation (LBGP-TOP + LPQ) [38]. Deep Transformation Learning (DTL) is employed in [57] to project deep features from face recognition task into new subspace - our proposed method obtains better results than this method. The MSN also outperforms the deep models in [16]. In [17], the authors present a deep learning method to explore spatial information and employ Feature Dynamic History Histogram (FDHH) to capture changes in the features. The MSN outperforms such method by significant margin. It shows that encoding jointly spatial and temporal information is a better approach to capture different ranges of facial dynamics related to depressive behaviours. The MSN also outperforms the deep models that only explore spatial information in [18], [19] and the model based on C3D in [23]. The results in Tables 4 and 5 confirm our assumption that spatial and temporal information captured in a multiscale approach is very important to encode facial expression variations for depression detection.

In Figure 5, we provide the predictions for all videos in the test sets of AVEC2013 and AVEC2014. When patients with a minimal level of depression, it can be seen that some samples were misclassified as mild level. These are the worst cases, which means that the proposed model can avoid misclassifications of patients with minimal level as severe level of depression. For patients with a severe level of depression, the model misclassified the samples as having the minimal level in 2 and 5 cases on AVEC2013 and AVEC2014, respectively. Moreover, we present the correlation between predictions and actual value by using Pearson Correlation Coefficient (PCC). As can be seen in Figure 5, the proposed model achieves PCC values of $0.727$ and $0.750$ on AVEC2013 and AVEC2014 datasets, respectively. These results indicate that the model provides a good level of accuracy, and a low likelihood of a very serious misclassification.

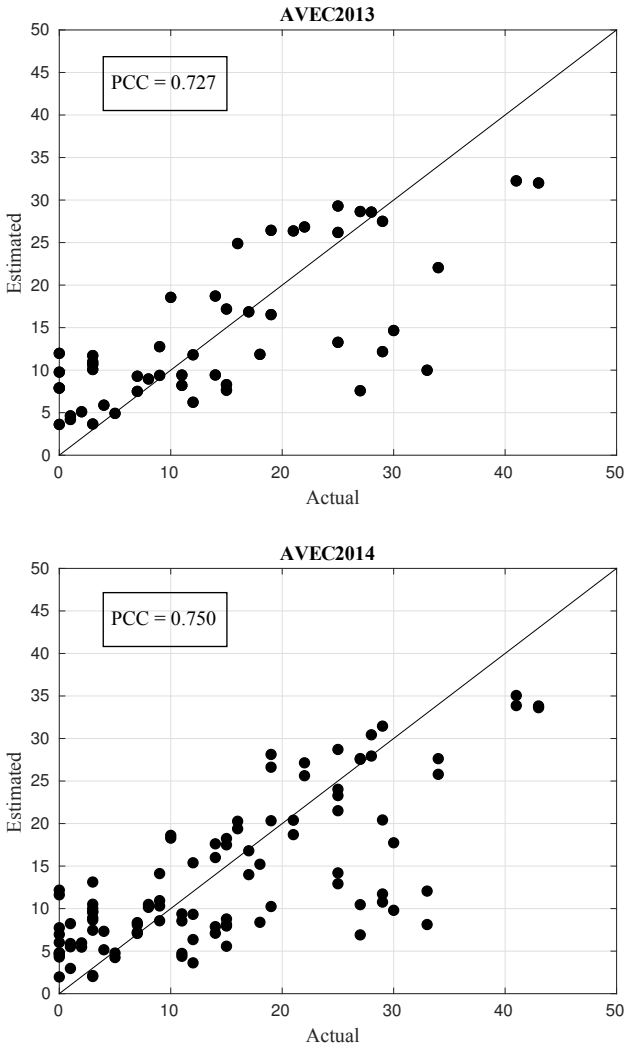



Fig. 5. Estimations of the proposed method on AVEC2013 (top) and AVEC2014 (bottom).

### 5.3 Visualization of activation mapping

Figure 6 presents the visualization attention over input clips produced by our MSN architecture. We visualize the activation maps by using Grad-CAM [55]. In our analysis, we considered two cases: the network with basic building block using single kernel and the proposed MSN. We show two images from an input clip (16 frames) – first and last frames. We consider all severity level of depression in order to analyze spatial and temporal regions that most favors to depression recognition. It is possible to observe that both networks focus attention in almost the whole facial area, *i.e.* the models capture facial expression variations from diverse facial areas. As the appearance and the motion information are employed to explore facial dynamics, the models increase, along the time, the region where it is paid more attention toward eye and mouth area in all depression levels which also indicates that such areas are important for generating a rich depression representation.

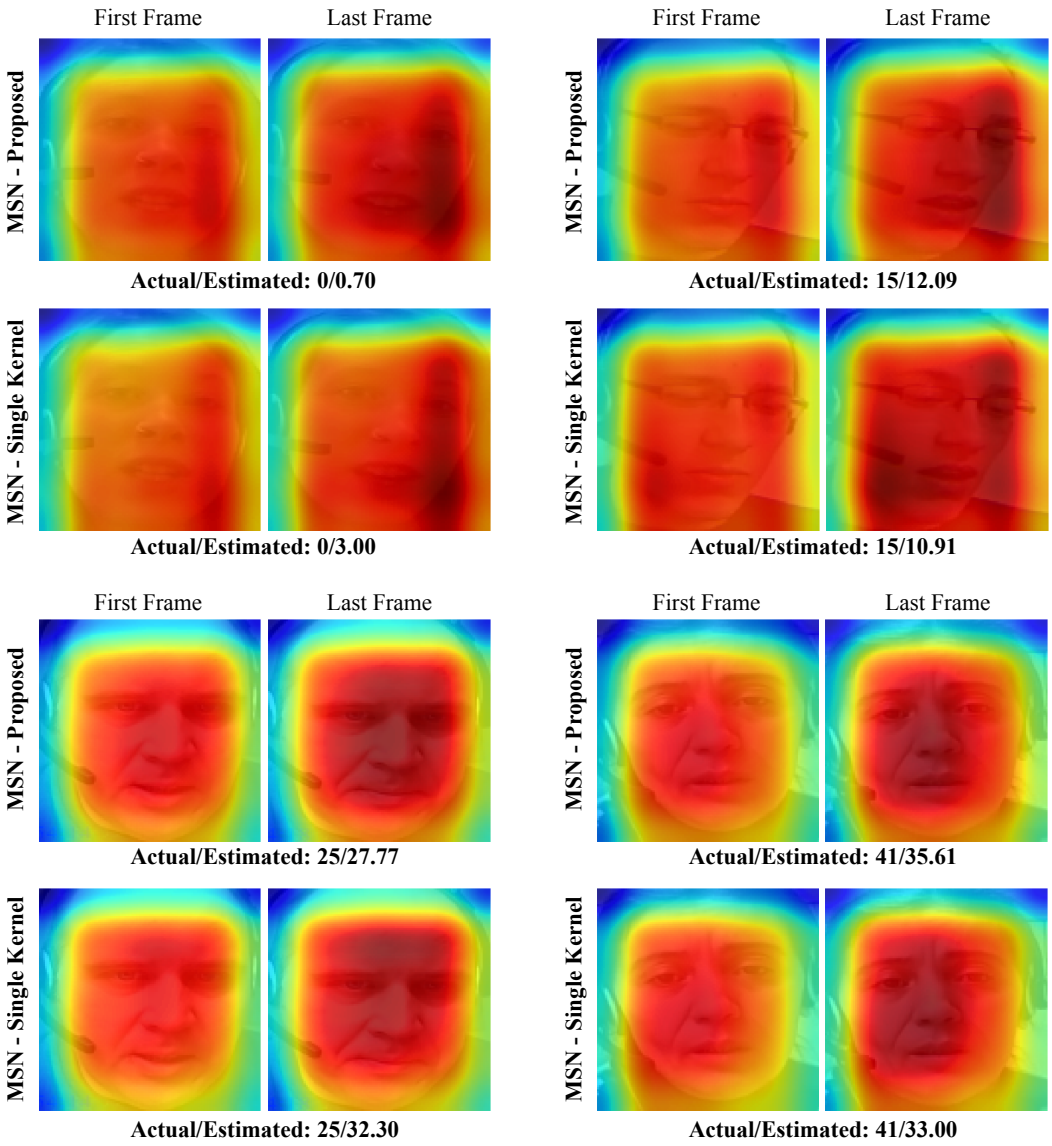



Fig. 6. Depression attention maps for input images with different depression levels. MSN-Single kernel means the network that employs the basic building block with only one 3D convolutional layer, $h_1 = w_1 = d_1 = 3$. MSN-Proposed refers to the network that uses the basic building block with three parallel 3D convolutional layers – $h_1 = w_1 = d_1 = 3$, $h_2 = w_2 = d_2 = 5$ and $h_3 = w_3 = d_3 = 7$. To generate the visualization, we employ the algorithm in [55].

In Figure 6, we can also observe the effect of the multi-scale ability of the MSN architecture. Analyzing spatio-temporal variations of mouth area, we can see that the proposed method explores more information from such area. For instance, examining the heatmap of the patient with minimal level of depression, it can be noted that the proposed model explores intensively nearly the whole mouth whereas the model with single kernel has more difficulties to analyze such area. It is also observed similar result on the patient with moderate level of depression. Such results are due to the capacity of the proposed MSN to investigate longer range of spatio-temporal variations when compared to the model with single kernel. Furthermore, the proposed MSN seeks to pay attention to the most relevant facial areas. In the example of the patient with mild level of depression, the model with single kernel pay high attention to the corner of the face in the first frame. Then, the exploration of this corner expands as we can see in the last frame of the clip. The proposed MSN does not pay high attention to this corner. Instead of that, the MSN focuses mainly on facial area that involves roughly eyes and mouth. For the patient with severe level of depression, the proposed MSN pays high attention to an area encompassing eyes and mouth which is slightly smaller than the one explored to the model with single kernel. Based on these observations, we can claim that the proposed MSN explores more efficiently the spatio-temporal information when compared with the model with single kernel.

## 6 Conclusion

In this paper, we explored the importance of spatial and temporal information for automatic depression assessment. We conducted this study by introducing a novel framework to represent the facial expression alterations called Multi-scale Spatiotemporal Network (MSN). The architecture has

the potential to encode rich spatio-temporal information of modifications in facial expressions using 3D convolutional layers with various kernel sizes, which allow the method to capture appearance and dynamics in different ranges. Such ability is important for modeling depressive behaviours from facial expression variations. In the experiments carried out with benchmark AVEC2013 and AVEC2014 depression datasets, the proposed MSN demonstrated to be more effective than I3D and C3D architectures in exploring spatio-temporal information. Moreover, MSN achieved good results and outperformed state-of-the-art methods, showing its effectiveness for depression detection. We believe that the results of this work can contribute to the progress of automatic medical diagnosis based on face analysis. The basic building block of MSN has the potential to capture rich spatio-temporal features and can be explored for detecting other abnormalities reflective of diseases in person's facial expressions. As a future work, we intend to employ our MSN model in another health care application based on facial information.

## ACKNOWLEDGMENTS

This research was partially supported by the Academy of Finland and the Natural Sciences and Engineering Research Council of Canada.

## REFERENCES

[1] J. Thevenot, M.B. López, and A. Hadid, "A Survey on Computer Vision for Assistive Medical Diagnosis From Faces," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 1497–1511, 2018.

[2] M. Bishay, P. Palasek, S. Priebe, and I. Patras, "SchiNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis," *IEEE Trans. on Affective Computing*, 2019.

[3] A. Bandini, J.R. Green, B. Taati, S. Orlandi, L. Zinman, and Y. Yunusova, "Automatic Detection of Amyotrophic Lateral Sclerosis (ALS) from Video-Based Analysis of Facial Movements: Speech and Non-Speech Tasks," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2018, pp. 150–157.

[4] 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*, 2017.

[5] 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," *Neurology, Psychiatry and Brain Research*, vol. 31, pp. 37–42, 2019.

[6] A. Bozorgmehr, F. Alizadeh, S.N. Ofogh, M.R.A. Hamzekalayi, S. Herati, A. Moradkhani, A. Shahbazi and M. Ghadirivasfi, "What Do the Genetic Association Data Say About the High Risk of Suicide in People with Depression? A Novel Network-Based Approach to Find Common Molecular Basis for Depression and Suicidal Behavior and Related Therapeutic Targets," *Journal of Affective Disorders*, vol. 229, pp. 463–468, 2018.

[7] J.L. Sotelo, and C.B. Nemeroff, "Depression as a Systemic Disease," *Personalized Medicine in Psychiatry*, vol. 1–2, pp. 11–25, 2017.

[8] American Psychiatric Association, "Diagnostic and Statistical Manual of Mental Disorders," *American Psychiatric Publishing*, 2013.

[9] A.J. Mitchell, A. Vaze, and S. Rao, "Clinical Diagnosis of Depression in Primary Care: A Meta-Analysis", *The Lancet*, vol. 374, pp. 609–619, 2009.

[10] M.J. Bostwick, "Recognizing Mimics of Depression: The '8 Ds'," *Current Psychiatry*, vol. 11, pp. 31–36, 2012.

[11] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, and F.D. la Torre, "Detecting Depression from Facial Actions and Vocal Prosody," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[12] J. Michalak, N.F. Troje, J. Fischer, P. Vollmar, T. Heidenreich, and D. Schulte, "Embodiment of Sadness and Depression-Gait Patterns Associated with Dysphoric Mood," *Psychosomatic Medicine*, vol. 71, pp. 580–587, 2009.

[13] J.Z. Canales, J.T. Fiquer, R.N. Campos, M.G. Soeiro-de-Souza, and R.A. Moreno, "Investigation of Associations Between Recurrence of Major Depressive Disorder and Spinal Posture Alignment: A Quantitative Cross-Sectional Study," *Gait Posture*, vol. 52, pp. 258–264, 2017.

[14] 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 *Proceedings of International Conference on Signal and Image Processing Applications*, 2015, pp. 578–583.

[15] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *arXiv:1804.08348*, 2018.

[16] 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*, pp. 1–8, 2017.

[17] A. Jan, H. Meng, Y.F.B.A. Gaus, and Fan 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.

[18] 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.

[19] W.C. de Melo, E. Granger, and A. Hadid, "Depression Detection Based on Deep Distribution Learning," in *Proceedings of IEEE International Conference on Image Processing*, 2019, pp. 4544–4548.

[20] X. Ma, D. Huang, Y. Wang, and Y. Wang, "Cost-Sensitive Two-Stage Depression Prediction Using Dynamic Visual Clues," in *Proceedings of Asian Conference on Computer Vision*, 2017, pp. 338–351.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features With 3D Convolutional Networks", in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[22] M.A. Jazaery, and G. Guo, "Video-Based Depression Level Analysis by Encoding Deep Spatiotemporal Features," *IEEE Trans. on Affective Computing*, pp. 1–8, 2018.

[23] W.C. de Melo, E. Granger, and A. Hadid, "Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2019.

[24] 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," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 3–10.

[25] 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," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 3–10.

[26] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, 2002.

[27] D. Basak, S. Pal, D.C. Patranabis, "Support Vector Regression," *Neural Information Processing-Letters and Reviews*, pp. 203-224, 2007.

[28] 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 *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 21–30.

[29] H. Meng, and N. Pears, "Descriptive Temporal Template Features for Visual Motion Recognition," *Pattern Recognition Letters*, vol. 30, pp. 1049–1058, 2009.

[30] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of Depression by Behavioural Signals: A Multimodal Approach," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 11–20.

[31] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[32] A. Bosch, A. Zisserman, and X. Munoz, "Representing Shape with a Spatial Pyramid Kernel," in *Proceedings of International Conference on Image and Video Retrieval*, 2007, pp. 401–408.

[33] 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.

[34] T.R. Almaev, and M.F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition," in *Proceedings of Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 356–361.

[35] H.P. Espinosa, H.J. Escalante, L. Villaseor-Pineda, M. Montes-y-Gmez, D. Pinto-Avedao, and V. Reyez-Meza, "Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 49–55.

[36] A. Jan, H. Meng, Y.F.A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic Depression Scale Prediction Using Facial Expression Dynamics and Regression," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 73–80.

[37] S. De Jong, "Simpls: An Alternative Approach to Partial Least Squares Regression," *Chemometrics and intelligent laboratory systems*, vol. 18, pp. 251–263, 1993.

[38] H. Kaya, F. illi, and A.A. Salah, "Ensemble CCA for Continuous Emotion Prediction," in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 19–26.

[39] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," in *Proceedings of International Conference on Learning Representations*, 2014.

[40] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge,"in *Proceedings of ACM International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[41] H. Dibekliolu, Z. Hammal, and J.F. Cohn, "Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 525–536, 2018.

[42] S. Song, L. Shen, and M. Valstar, "Human Behaviour-Based Automatic Depression Analysis Using Hand-Crafted Statistics and Deep Learned Spectral Features," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2018, pp. 158–165.

[43] A. Haque, M. Guo, A.S. Miner, and L. Fei-Fei, "Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions," in *arXiv preprint arXiv:1811.08592*, 2018.

[44] Z. Du, W. Li, D. Huang, and Y. Wang, "Encoding Visual Behaviors with Attentive Temporal Convolution for Depression Prediction" in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2019, pp. 1–7.

[45] C. Lea, M.D. Flynn, R. Vidal, A. Reiter, and G.D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[46] F. Yu and V. Koltun, "Multi-scale Context Aggregation by Dilated Convolutions," in *Proceedings of International Conference on Learning Representations*, 2016.

[47] J. Carreira, and A. Zisserman, "Quo vadis, Action Recognition? A New Model and The Kinetics Dataset," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4724–4733.

[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[49] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[50] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[52] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 2016.

[53] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Face Across Pose and Age," in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2018, pp. 67–74.

[54] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[55] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

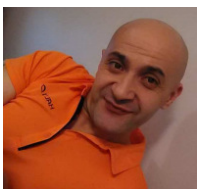
[56] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of Audio-visual Features Using Hierarchical Classifier Systems for the Recognition of Affective States and the State of Depression," in *Proceedings of International Conference on Pattern Recognition Applications and Methods*, 2014, pp. 671–678.

[57] Y. Kang, X. Jiang, Y. Yin, Y. Shang, X. Zhou, "Deep Transformation Learning for Depression Diagnosis from Facial Images," in *Proceedings of Chinese Conference on Biometric Recognition*, 2017, pp. 13–22.

[58] H. Kaya, and A.A. Salah, "Eyes Whisper Depression: A Cca Based Multimodal Approach," in *Proceedings of ACM International Conference on Multimedia*, 2014, pp. 961–964.

**Wheidima Carneiro de Melo** was born in Manaus, AM, Brazil, in 1983. He received his B. Sc. degree in Federal University of Amazonas (UFAM), and the M. Sc. degree in Electrical Engineering, in 2014, from Federal University of Amazonas (UFAM). Since 2013, he has been with the Superior School of Technology (EST) of Amazonas State University (UEA), as professor. He is currently pursuing a Ph.D. degree in Computer Science and Engineering at the University of Oulu. His research interests are in the fields of computer vision, machine learning, and affective computing.

**Eric Granger** received the Ph.D. degree in EE from Eécole Polytechnique de Montréal in 2001. He was a Defense Scientist with DRDC, Ottawa, from 1999 to 2001, and in R&D with Mitel Networks from 2001 to 2004. He joined École de technologie supérieure, Université du Québec, Montreal, Canada, in 2004, 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 adaptive pattern recognition, machine learning, computer vision, and computational intelligence, with applications in biometrics, face recognition and analysis, video surveillance, and computer/network security.



**Abdenour Hadid** received his Doctor of Science in Technology degree in electrical and information engineering from the University of Oulu, Finland, in 2005. Now, he is an Associate Professor at Center for Machine Vision and Signal Analysis, University of Oulu, Finland. His research interests include biometrics and facial image analysis, machine learning and artificial intelligence, human-machine interaction, personalized healthcare and mobile applications. He has authored more than 160 papers in international conferences and journals, and served as a reviewer for many international conferences and journals. His research works have been well referenced by the research community with more than 13000 citations so far according to Google Scholar. Prof. Hadid is currently a senior member of IEEE.