Combining Global and Local Convolutional 3D Networks for Detecting Depression from Facial Expressions

Wheidima Carneiro de Melo[1], Eric Granger[2] and Abdenour Hadid[1]
[1] Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
[2] Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), ETS Montréal, Canada

***Abstract*— Deep learning architectures have been successfully applied in video-based health monitoring, to recognize distinctive variations in the facial appearance of subjects. To detect patterns of variation linked to depressive behavior, deep neural networks (NNs) typically exploit spatial and temporal information separately by, e.g., cascading a 2D convolutional NN (CNN) with a recurrent NN (RNN), although the intrinsic spatio-temporal relationships can deteriorate. With the recent advent of 3D CNNs like the convolutional 3D (C3D) network, these spatio-temporal relationships can be modeled to improve performance. However, the accuracy of C3D networks remain an issue when applied to depression detection. In this paper, the fusion of diverse C3D predictions are proposed to improve accuracy, where spatio-temporal features are extracted from global (full-face) and local (eyes) regions of subject. This allows to increasingly focus on a local facial region that is highly relevant for analyzing depression. Additionally, the proposed network integrates 3D Global Average Pooling in order to efficiently summarize spatio-temporal features without using fully-connected layers, and thereby reduce the number of model parameters and potential over-fitting. Experimental results on the Audio Visual Emotion Challenge (AVEC 2013 and AVEC 2014) depression datasets indicates that combining the responses of global and local C3D networks achieves a higher level of accuracy than state-of-the-art systems.**

I. INTRODUCTION

The idea recognizing persons affective state for health diagnosis and monitoring is promising for the future of health care. It would enable long-term health monitoring, which is important for the treatment and management of a wide range of chronic illnesses, neurological disorders, and mental health issues, such as diabetes, hypertension, asthma, autism spectrum disorder, fatigue, depression, drug addiction, etc.

Major depressive disorder represents a leading cause of disability worldwide and a significant cost to health care systems. This psychiatric disorder is defined as negative state of mind that last over a long period [1]. The detection and treatment of depression has long been recognized as a major priority for improving the health and well-being of millions of individuals. It has an adverse impact on patient feelings, thoughts, behavior and body, and in more severe cases, depression is considered one of the leading causes of suicide and substance abuse [2]. Normally, medications and psychotherapy are effective treatments for depression, however errors in clinical assessment of depression are frequent [3]. Indeed, depression diagnosis is based on the Diagnostic Statistical Manual of mental disorders, and is diagnosed using the Structured Clinical Interview for DSM-IV. The severity of depression is based on a score obtained by answering Hamilton Rating Scale or Beck Depression Inventory-II. Studies have shown an alarming rate of false detection in depression diagnosis by clinicians, with potentially severe consequences [4], [5].

The challenges of diagnosis have driven the scientific community to investigate sensor-based monitoring for accurate prediction of a subject's level of depression based on patterns of verbal and nonverbal behaviour. Sources of audio and visual information can provide clinicians with valuable information for diagnosis. Some studies have shown that properties of voice and speech change when a person is depressed [6], [7]. It is also possible to obtain information about the level of depression by analyzing visual information from face and body, such as facial expressions and head pose [8]. Moreover, recent advances in machine learning and behavioral signal processing have led to promising techniques for automated diagnosis of depression using a range of modalities. For instance, Pampouchidou *et al.* [4] propose a low-cost algorithm to detect depression from faces captured in videos by using curvelet transform and Local Binary Patterns (LBPs). Ma *et al.* [9] cascade a CNN and Long Short-Term Memory (LSTM) for classification of depression levels using audio recordings.

This paper will focus on techniques for accurate detection of depression levels over time based on facial expression captured in videos. Although spatial information is essential, the dynamics of facial behavior is also very important to interpret depression [10], [11], [12]. Alghowinem *et al.* [13] have identified behavioural cues related to depressed individuals, such as slower head movements and avoidance of eye contact. Therefore, exploiting spatio-temporal information in videos is typically considered to improve detection accuracy. However, some key challenges for detection in real-world scenarios are the significant variations over time in the facial expressions according to the specific person, sensors, computing device and operational environment. In addition, there is often a limited amount of labeled data to design the predictive models. It is difficult and costly to create large-scale datasets that are reliably annotated for the detection of depression levels. Finally, it is difficult to encode common and discriminant spatio-temporal features of depression while suppressing subject-specific facial shape variations.

Deep learning architectures, and in particular CNNs, provide state-of-the-art performance in many visual recognition

applications, such as image classification [14] and object detection [15], as well as assisted medical diagnosis [16]. In depression detection, deep learning architectures that process on videos typically exploit spatial and temporal information separately (e.g., by cascading a 2D CNN and then a recurrent NN), which deteriorate the modeling of spatio-temporal relationships [11], [17]. A deep two-stream architecture has also been proposed to exploit facial appearance and facial optical flow [10]. Some 3D CNNs – like the C3D network [19] – have recently been proposed to leverage spatio-temporal relationships, and improve detection accuracy. However, the accuracy and complexity of these models remain an issue when applied to depression detection.

In this paper, C3D networks are combined to explore spatio-temporal dependencies in the global and local facial regions captured in a video. The proposed approach leverages the spatio-temporal information that corresponds local facial regions in order to predict depression levels. This allows to increasingly focus attention on local facial regions that display highly relevant facial variations for analyzing depression, and to increase the value of spatio-temporal information [41]. The local region used in this paper is a coarse eye region. Predictions from a C3D trained on the global full-face region are combined with a C3D trained on local eye region, and their diverse and complementary features allow to increase detection accuracy. In addition, 3D Global Average Pooling (3D-GAP) is integrated into C3D networks to summarize high level spatio-temporal features from the last convolutional layer. With 3D-GAP, the fully connected layers are removed, and the number of model parameters is decreased, leading to a reduced time complexity and risk of over-fitting. The performance of the proposed approach is compared to several state-of the-art (conventional and deep learning) techniques for depression detection on the Audio Visual Emotion Challenge (AVEC 2013 and 2014) datasets.

The rest of this paper is divided as follows. Section II provides a summary of related research on depression recognition based on variations in facial appearance captured in videos. Section III presents the proposed approach that combines the responses of global and local C3D networks and 3D-GAP. Finally, Sections IV and V present the experimental methodology used for proof-of concept validation, and the main results with our discussion, respectively.

II. RELATED WORK

Automatic depression recognition using facial cues is a prominent field which has recently attracted some attention from the computer vision and machine learning communities. Cohn *et al.* [20] showed the viability of automatic depression detection by comparing clinical diagnosis with automatic measured facial actions using an Active Appearance Model (AAM). In [1], the authors also employ an AAM to extract features from eye movement in videos for classifying subjects as depressed or non-depressed. Joshi *et al.* [21] proposed a multimodal approach based on audio and visual information, where visual features are extracted to analyze the movements of body parts.

The AVEC competitions held in 2013 [22] and 2014 [23] had, as sub-challenge, the task that required participants to estimate the level of self-reported depression in each video. The dataset used by the participants is publicly-available for scientific community, and contains audio and visual information. Although a multimodal approach can improve performance [12], this paper focuses on exploring the spatio-temporal information extracted from faces captures in videos using 3D CNNs. The following describes some relevant state-of-the-art methods that apply AVEC2013 and AVEC2014 datasets.

In AVEC 2013, the competition made publicly available baseline system to process for visual and audio data. Regarding the visual features, face detection and alignment are performed for every video frame, then it is used Local Phase Quantisation (LPQ) as dense local appearance descriptor. The LPQ descriptor [24] employs a Short-Term Fourier Transform to obtain local phase over a facial area. Finally, the feature vectors obtained by concatenating histograms of the descriptions are fed to a Support Vector Regressor (SVR) trained to estimate the depression levels.

Meng *et al.* [25] proposed a conventional method based on Motion History Histogram [26] to capture motion information of facial expressions. Additionally, temporal details are highlighted by using Edge Orientation Histogram [27] and LBP [28] descriptors, where features are concatenated to generate the final representation. Lastly, the mapping from feature representation space to level of depression is learned by applying a Partial Least Square Regressor [29]. Cummins *et al.* [30] investigated the use of two different descriptors which are called Space-Time Interest Points (STIP) [31] and Pyramid of Histogram of Gradients (PHOG) [32], with PHOG demonstrating better accuracy. Lingyun Wen *et al.* [33] proposed extracting dynamic feature descriptors based on LPQ from Three Orthogonal Planes (LPQ-TOP). Sparse coding is then applied to organize the feature descriptors, and SVR allows predicting the levels of depression level.

In the AVEC2014 competition, the baseline visual features are extracted by using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [34] which combines dynamic and spatial texture analysis with Gabor filtering. The method convolves consecutive video frames with a number of Gabor filters, and applies LBP in order to extract features from orthogonal XY, XT and YT slices. Finally, the resulting feature representation obtained by concatenating histograms is input to an SVR trained to predict depression levels.

Prez *et al.* [35] compute differences of eye and face positions, combine these values with motion history image, motion static image and motion average image from the segment of video that contains the facial regions, followed by SVR predictions. In [36], Jan *et al.* extract three distinct texture feature representations – using LBP, EOH and LPQ methods – mapping their variations into a feature vector by using 1-D MHH. PLS and Linear regression schemes are used to predict depression levels. Lastly, Kaya *et al.* [37] compute Canonical Correlation Analysis on baseline and LPQ features, exploring eye and mouth areas of the face, and

then aggregating the resulting features to predict a continuous depression levels.

The conventional depression recognition methods described above are based on hand-crafted features. However, deep learning techniques that learn discriminant feature representations from training data are considered to be state-of-the-art in depression recognition. Deep learning models for video-based depression detection often cascade a 2D CNN with an RNN [17]. Most notably, Jan *et al.* [38] propose deep learning techniques to extract features from facial frames and employ feature dynamic history histogram (FDHH) to capture variations in the features. In addition, other authors [10] proposed a two-channel CNN where one channel inputs full facial regions, whereas the second inputs facial flows, with two fully connected layers performing the fusion of the features. However, exploiting spatial and temporal information separately can deteriorate the modeling of spatio-temporal relationships.

In this paper, we focus on improving accuracy by leveraging spatio-temporal information learned with C3D networks. Recently, Mohamad *et al.* [44] proposed to extract spatio-temporal features from facial videos at two different scales, using a C3D network to capture spatio-temporal features and a RNN to model transitions of the features. However, spatio-temporal modeling using 3D deep learning techniques have important challenges in real-world applications. In particular, it is difficult to learn discriminant spatio-temporal features for depression behaviours that are robust to variability of different subjects and capture conditions and devices. The proposed approach explores the use of C3D networks specialized for local and global facial regions and 3D-GAP to improve performance.

III. A GLOBAL-LOCAL C3D MODEL FOR DEPRESSION DETECTION

Video signals are normally described as 3D data that integrates spatial and temporal information. For our purposes, the spatio-temporal information corresponds to the variation in appearances of facial regions extracted from consecutive frames. The system proposed in this paper detects levels of depression using deep 3D CNNs that are trained to model spatio-temporal information.

This section presents a system that combines C3D networks that model the spatio-temporal dependencies in global and local facial regions captured in a video. The global C3D network models spatio-temporal information from the entire face, while the local C3D network allows to focus attention on the eye region that is considered to be highly relevant for depression. Additionally, the 3D Global Average Pooling (3D-GAP) is proposed to summarize the spatio-temporal features of the last convolutional layer. The last two fully connected layers are thereby replaced, decreasing the number of C3D model parameters, and leading to better generalization. Finally, the generated features are applied to regression layer in order to estimate the level of depression.

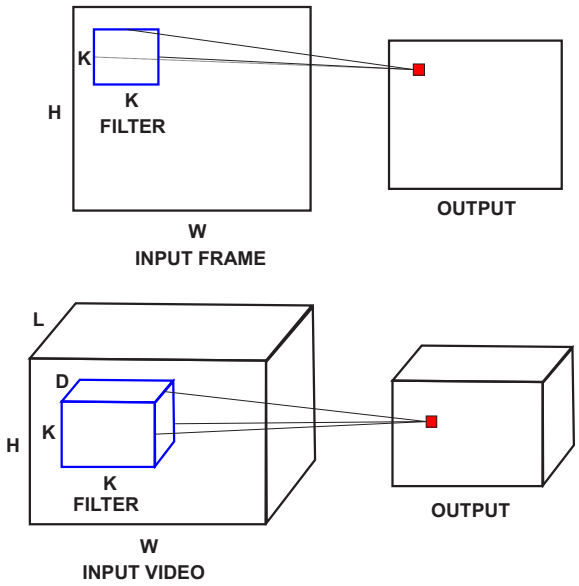

Fig. 1. Illustration of 2D and 3D convolution operations. (Top) Performing 2D convolution using a K × K filter on an image of size H × W generates an output image. (Bottom) Performing 3D convolution using a K × K × D filter on a video of L frames (of size H × W) generates an output volume.

A. Convolutional 3D (C3D) Network:

A system based on 2D CNNs normally generates spatial features using convolution and pooling operations over 2D data. In contrast, 3D CNNs have use 3D convolution and pooling layers to learn spatio-temporal features because the operations are performed on volumes of video data. Figure 1 illustrates the difference between both operations. Note that 3D convolution operations output a volume, whereas 2D convolution outputs an image. In the 2D convolution, a 2D filter of size K × K is slid along the width and height of the input image. In order to compute the 3D convolution, both input and filter have depth dimension D, and the 3D filter also slides in the depth direction. The output of a 3D convolution operation can be computed by using the following equation:

$$Y_{z,x,y} = \sum_{l=0}^{L-1} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} I_{l,h,w} F_{z+l,x+h,y+w}, \qquad (1)$$

where $I_{l,h,w}$ represents the input frame l of the video, and $F_{z+l,x+h,y+w}$ is the filter.

With our proposed method, the deep C3D network is adopted to model spatio-temporal information. This network has been successfully applied in action, scene and object recognition [19]. One reasons for its recent popularity and success is the availability of C3D networks that have been pre-trained on large-scale datasets, and can efficiently capture spatio-temporal features. In this paper, the C3D network is first pre-trained on Sports-1M [39] and UCF101 [40] datasets for action recognition. A pre-trained C3D network can then be fine-tuned for a similar task on application data, through transfer learning. This approach can be efficient since the features are appropriate for both tasks. The public AVEC 2013 and AVEC 2014 depression detection datasets are used

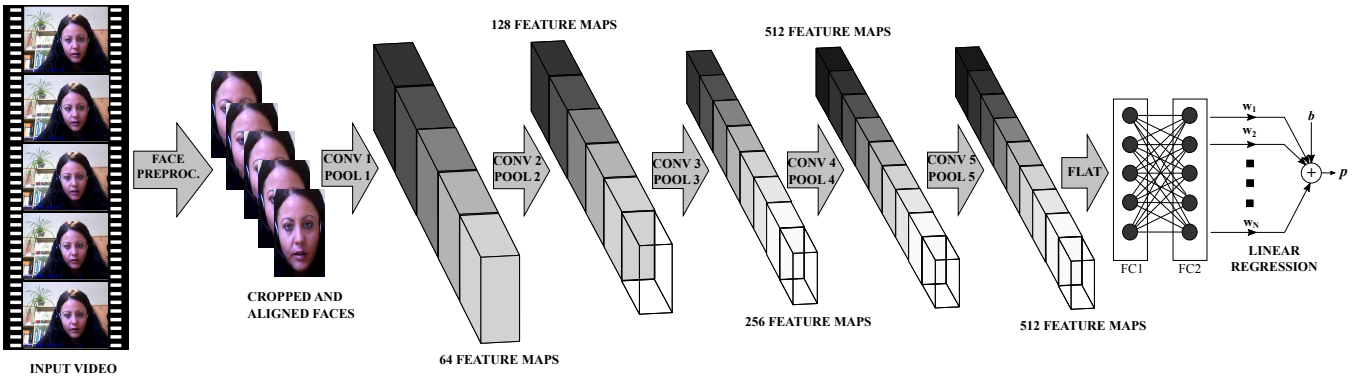

Fig. 2. Illustration of the proposed architecture for performing video-based depression analysis.

for fine-tuning. After the transfer learning process the C3D network can capture spatio-temporal features from complex activities, making them suitable for detection of depression from faces captured in videos.

The original C3D architecture is composed by 8 convolutional layers and 5 pooling layers for feature learning, then two fully connected layers and softmax output layer for classification. The 3D convolution filters have the same dimensions which are $k \times k \times d$, where d is temporal depth and k is spatial size, with $d = 3$ and $k = 3$. The first 3D pooling layer has size of $2 \times 2 \times 1$ and stride $2 \times 2 \times 1$, whereas the other ones has size of $2 \times 2 \times 2$ with stride $1 \times 1 \times 1$. Finally, the fully connected layers have 4096 neurons. During the transfer learning process, the fully connected layers are retrained using 512 neurons and the softmax layer is substituted for regression output layer in order to generate depression level scores. In testing or operational mode, the output depression score is given by:

$$p = \sum_{i=1}^{N} w_i f_i + b, \tag{2}$$

where p is the predicted value, $\mathbf{w}$ and b are the weight and bias parameters of the regression layer, and $\mathbf{f}$ is the input feature vector of length N. The proposed architecture is illustrated in Figure 2. Pre-processing is employed to crop, normalize and align successive faces of a video.

B. 3D Global Averaging Pooling

In order to minimize the problem of over-fitting, 3D Global Average Pooling (3D-GAP) is proposed to reduce the spatial-temporal dimensions of features which are output of the last convolutional layer. Specifically, there are 512 filters in the last convolutional layer, which generates 512 outputs with three-dimensions. This spatio-temporal information is averaged, producing a feature vector with 512 values that is fed into the linear regression model. At the end, the spatio-temporal features are compressed which contributes to reducing the total number of parameters in the deep 3D model. Indeed, the last two fully connected layers are removed, increasing generalization capacity of the model. Feature vector $\mathbf{f}$ regroups the output set of features defined by:

$$f_n = \frac{1}{Z \times Y \times X} \sum_{z=0}^{Z-1} \sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} H(z, y, x), \tag{3}$$

where H represents the three-dimensional set of features, and $n = 1, 2, ..., N$, and N is the number of filters in the last convolutional layer of the model. The proposed architecture with 3D-GAP is illustrated in Figure 3.

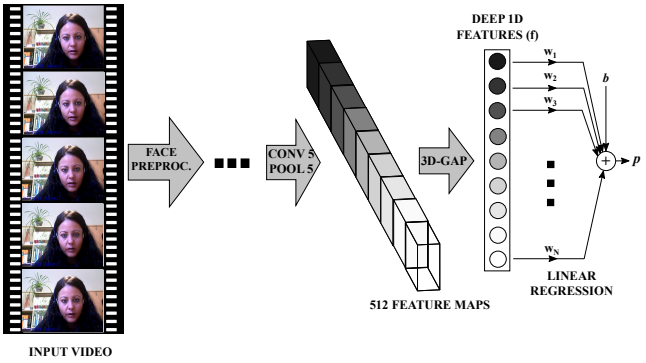

Fig. 3. Illustration of the proposed architecture using 3D-GAP for performing video-based depression analysis.

C. Fusion of Global and Local C3D Networks:

Figure 4 illustrates the proposed models based on global and local C3D networks to estimate depression levels. Initially, the analysis of a facial video involves cropping, normalizing and aligning the consecutive faces extracted from videos. Then, these faces are assigned to overlapping windows or clips before being presented to a C3D network. This procedure allows for mapping the patterns of facial appearance variations over successive frames. However, capturing and encoding discriminant spatio-temporal information to detect depression is a challenging task. The facial structures can be repeated in different mood states, and learning the facial variations of all inter-segment dependencies may require a large amount of labeled data.

The proposed architecture performs late fusion of 2 C3D networks that model the spatio-temporal dependencies in

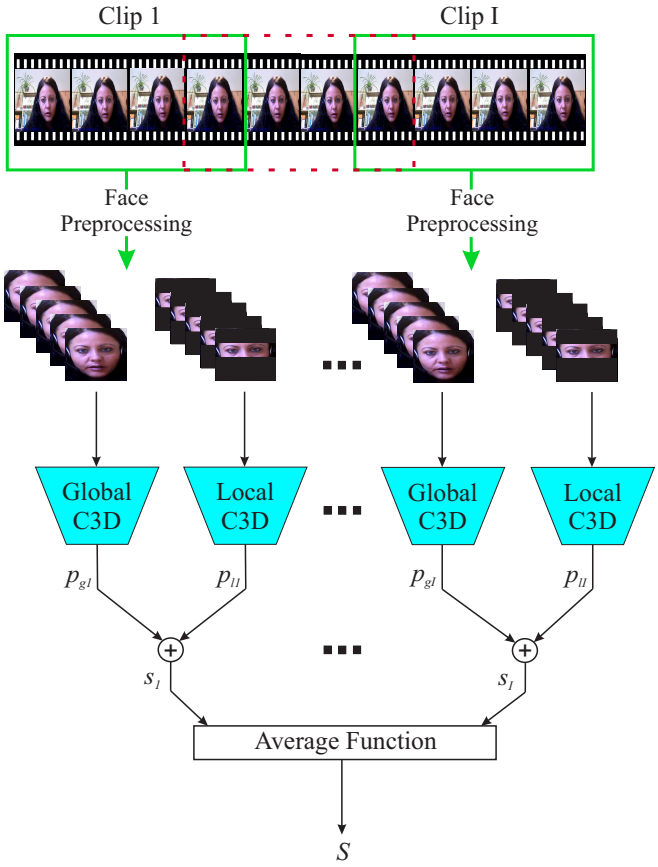

Fig. 4. Proposed approach to estimate depression levels.

global and local facial regions captured. It leverages the spatio-temporal information that corresponds local facial regions. The main rational is to focus more attention on facial areas that display highly relevant facial variations, in order to maximize the value of spatio-temporal information. The local region used in this work is a coarse eye region which represents a highly relevant area for facial expression recognition [41]. Additionally, the spatio-temporal information from the global or full facial region is also combined with local region in the proposed method. With that, the model can explore divers and complementary features which can increase the power of generalization of the proposed model. This architecture is also motivated by trunk-branch ensembles that have been shown to improve performance in video-based face recognition [18].

Consequently, C3D predictions from the local and global clips are combined, and a regression model is trained to predict the overall depressive states. In the prediction stage, an score-level fusion scheme combines the C3D prediction values for both region according to: $s_i = p_{li} + |p_{li} - p_{gi}| \times 0.5$, where s_i is the score for ith clip, p_{li} represents the prediction value for local C3D, p_{gi} is the prediction value for global C3D. To predict the overall depression level, the proposed model outputs a real score value for each video clip, and calculates the average overall output values with:

$$S = \frac{1}{I}\sum_{i=1}^{I} s_i = \frac{1}{I}\sum_{i=1}^{I} p_{li} + |p_{li} - p_{gi}| \times 0.5, \qquad (4)$$

where I is the number of clips in the entire video sequence.

IV. EXPERIMENTAL METHODOLOGY

A. Datasets:

The proposed model is used for automatically predict the depression levels of subjects. In order to evaluate its performance, experiments were performed on two datasets: AVEC2013 and AVEC2014 depression sub-challenge datasets. The objective of the sub-challenge is estimate score of subjects on the Beck Depression Inventory (BDI). The BDI scores range from 0 to 63 with the following classification: 0-13 (none depression), 14-19 (mild depression), 20-28 (moderate depression), and 29-63 (sever depression).

AVEC2013 depression dataset is a subset of the audio-visual depressive language corpus (AViD-Corpus), which comprises 150 videos from 82 subjects. The data video is acquired using a webcam and microphone, while one person performing a Human-Computer Interaction task. The frame rate of the videos is 30 Hz at 640×480, with 24 bits per pixel. In average, the length of videos is about 25 minutes. The recordings are divided into three partitions: training, development and test set of 50 videos, respectively. The experiments are performed considering training and development sets as training data, whereas the test set is used to verify the performance of the proposed method.

The event competition Audio/Visual Emotion Challenge 2014 provided the AVEC2014 depression dataset. This dataset also uses a subset of AViD-Corpus. The subjects are recorded using a webcam and microphone while performing two tasks: Freeform task, participants respond questions such as discuss a sad childhood memory, and Northwind task, participants read audibly an excerpt from a fable. In both tasks, the recordings are divided into three partitions: training, development and test set with 50 videos in each partition. In total, there are 300 videos with average length equals to 2 minutes. In order to perform experiments, training and development sets are used from both tasks as training data, and the test sets are employed to measure the performance of the model. Some samples of both datasets are presented in Figure 5.

Due to the large number of frames in each sample of training datasets, the videos are downsampled, where the frame rate is decreased by a factor of 100 and 10, for AVEC2013 and AVEC2014, respectively. This procedure has been applied for some methods [10], [44] with no loss of performance. Each sample is divided into overlapping 10-frame clips with temporal stride 8.

B. Protocol:

As a first step, face preprocessing is needed to detect and align the faces captured in videos (see Figure 2). The aim is

Fig. 5. Samples from AVEC2014 (top) and AVEC2013 (bottom) databases.

to provide the proposed deep model with frontal face regions, so that it can analyze spatio-temporal information related to these regions. The Multi-Task Cascade Convolutional Network (MTCNN) [43] was adopted to jointly detect and align faces. It is comprised of the proposal network (P-Net), refinement network (R-Net) and output network (O-Net). The P-Net and R-Net generate and analyze candidate windows as well as eliminate non-face windows. The O-Net determines the bounding box and five facial landmarks which are used to face detection and alignment. The resulting facial images have size of 112×112. This procedure is carried out for all frames of the AVEC2013 and AVEC2014 datasets. The subjects are cooperative, and the detected faces are not significantly miss-aligned, so we assume that the alignment process does not impair the spatio-temporal features.

The convolutional layers of the proposed model are initialized with parameters which were pre-trained on Sports-1M dataset and UCF101. Details about this pre-trained model can be found in [19]. The two fully connected layers are randomly initialized. The regression loss function of the proposed method is Mean Squared Error (MSE). The fine-tuning process is applied in the model using gradient descent algorithm with decaying learning rate, where the initial learning rate is 1×10^{-6}. The model is implemented using Tensorflow [42].

C. Performance Measures:

The performance of the proposed and baseline methods for depression detection are evaluated on the test set using two error measures – Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Once the predicted score is obtained using the mean of values estimated of each clip in the video, the MAE is computed using:

$$MAE = \frac{1}{N} \sum_{i=0}^{N-1} |y_i - \hat{y}_i|, \tag{5}$$

where y_i is the ground truth for *i*th input video, $\hat{y}_i$ represents the estimated value, and N is the number of samples. The RMSE is computed using:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}, \tag{6}$$

with the same definitions.

TABLE I
PERFORMANCE OF PROPOSED METHODS ON AVEC2013.

Proposed Methods	RMSE	MAE
C3D Global	9.30	7.20
C3D Global + 3D-GAP	9.24	7.10
C3D Local	8.83	6.79
C3D Local + 3D-GAP	8.37	6.51
C3D Global and Local	8.73	6.69
C3D Global and Local + 3D-GAP	**8.26**	**6.40**

TABLE II
PERFORMANCE OF PROPOSED METHODS ON AVEC2014.

Proposed Methods	RMSE	MAE
C3D Global	8.99	7.23
C3D Global + 3D-GAP	8.97	7.09
C3D Local	8.87	7.02
C3D Local + 3D-GAP	8.55	6.81
C3D Global and Local	8.76	6.90
C3D Global and Local + 3D-GAP	**8.31**	**6.59**

V. RESULTS AND DISCUSSIONS

Tables I and II show the results obtained with the methods proposed for depression detection on AVEC2013 and AVEC2014 datasets, respectively. In these tables, the performance is shown for C3D networks using the global-face, local-face and both facial regions, with and without 3D-GAP.

With the AVEC2013 dataset (Table I), using C3D with the Global-Face representation provides higher error than with the Local-Face representation. When Local-Face and Global-Face predictions are combined at the score level, RMSE is decreased by a margin of 0.54 and 0.10 over Global-Face and Local-Face representations, respectively. The results on AVEC204 dataset (Table II) for Global-Face are 8.99 (RMSE) and 7.23 (MAE), whereas the model obtains RMSE and MAE equals to 8.87 and 7.02, respectively, for Local-Face. Once again, C3D obtains better results when it combines both facial representations. The performance achieved when the C3D network uses a Local-Face (eye region) representation is better than results obtained with a Global-Face representation. These results suggest a certain redundancy of spatio-temporal information in global face regions, and shows the importance of spatial structure and temporal dependencies of the local eye regions for prediction of depression using C3D networks.

Tables I and II show that 3D-GAP also has the potential to improve the accuracy of C3D networks. For the AVEC2013 dataset, using 3D-GAP decreases the MAE obtained with a C3D network using Global-Face and Local-Face representations by margin of 0.10 and 0.28, respectively. 3D-GAP decreases the RMSE of C3D by a margin of 0.47 for fusion of Global-Face and Local-Face predictions. Likewise, the results generated by C3D on AVEC2014 are improved when 3D-GAP is employed. C3D results using the Local-Face representation improve by margin of 0.32, in terms of RMSE. The lowest error is always obtained using a C3D network with 3D-GAP, and when combining Global-Face and Local-Face representations.

It is also possible to compare between C3D and C3D

+ 3D-GAP in terms of the total number of parameters. The C3D network has two fully connected layers with 512 neurons, and the output of the last convolutional layer has size of $7 \times 7 \times 2$. When 3D-GAP method is employed, these two fully connected layers are removed, which translates to a reduction by more than 313k parameters.

Table III shows the performance of proposed methods (C3D network using Global-Face and/or Local-Face representations + 3D-GAP) compared with baseline and state-of-the-art methods on AVEC2013 dataset. Results obtained with the proposed method provides lowest RMSE and MAE values. The baseline method uses hand-crafted features produced with the LPQ descriptor. It can be observed that C3D network using Global-Face and Local-Face + 3D-GAP outperforms this system by margin (e.g., 5.35 in terms of RMSE). Compared to the method presented in [46], it reduces MAE by margin of 1.46. In [10], the system uses two channels to explore the temporal information using facial optical flows (channel 1) and the spatial information (channel 2). As shown, the proposed method achieve better results, which suggests that learning spatio-temporal features directly is more effective for detecting depression.

Table IV presents the performance of proposed methods on AVEC2014 dataset. The errors with the baseline method are 2.27 (MAE) and 2.55 (RSME) higher than with the C3D network using Global-Face and Local-Face + 3D-GAP. The system proposed in [37] based on LGBP-TOP and LPQ descriptors, generates value of MAE 1.61 higher than C3D using Global-Face and Local-Face + 3D-GAP. The proposed method also outperforms the system proposed by Zhu *et al.* [10] for AVEC2014 dataset. Finally, the proposed method achieves competitive performance with the one in [38], where C3D using Global-Face and Local-Face + 3D-GAP obtains better results in terms of MAE.

In Tables III and IV, the method of Mohamad *et al.* [44] employs a C3D network trained on tight facial regions, and a C3D network trained on loose facial regions (cover more than face region). C3D productions from both these face regions are combined with a Recurrent NN. It is important to note that the method using C3D Tight-Face corresponds to the same input as our Global-Face region. Results show that the proposed architecture using C3D networks with Local-Face + 3D-GAP achieves a significantly lower error than C3D for Loose-Face, for Tight-Face, and even for RNN-C3D Combined faces. Moreover, the proposed method (combining C3D networks with Global-Face and Local-Face + 3D-GAP) outperforms all three methods proposed in [44] on AVEC2013 as well as AVEC2014 by considerable margin. From these results, it can be concluded that using the proposed approach (C3D network with 3D-GAP) on AVEC 2013 and AVEC 2014 datasets can allow to efficiently learn spatio-temporal features, and to outperform state-of-the-art methods for depression recognition.

TABLE III
PERFORMANCE OF THE PROPOSED AND BASELINE METHODS FOR DEPRESSION RECOGNITION ON AVEC2013.

Methods	**RMSE**	**MAE**
Baseline [22]	13.61	10.88
LPQ + SVR [45]	10.82	8.97
PHOG [30]	10.45	N/A
MHH + EOH + LBP [25]	11.19	9.14
LPQ-TOP + MFA [33]	10.27	8.22
LPQ + Geo [46]	9.72	7.86
Two DCNN [10]	9.82	7.58
C3D Tight-Face [44]	9.64	7.50
C3D Loose-Face [44]	10.04	8.15
RNN-C3D Combined Faces [44]	9.28	7.37
C3D Global + 3D-GAP (Ours)	9.24	7.10
C3D Local + 3D-GAP (Ours)	8.37	6.51
C3D Global and Local + 3D-GAP (Ours)	**8.26**	**6.40**

TABLE IV
PERFORMANCE OF THE PROPOSED AND BASELINE METHODS FOR DEPRESSION RECOGNITION ON AVEC2014.

Methods	**RMSE**	**MAE**
Baseline [23]	10.86	8.86
MHI + MSI + MAI [35]	9.84	8.46
MHH + PLS [36]	10.50	8.44
LGBP-TOP + LPQ [37]	10.27	8.20
Two DCNN [10]	9.55	7.47
VGG + FDHH [38]	**8.04**	6.68
C3D Tight-Face [44]	9.66	7.48
C3D Loose-Face [44]	9.81	7.73
RNN-C3D Combined Faces [44]	9.20	7.22
C3D Global + 3D-GAP (Ours)	8.97	7.09
C3D Local + 3D-GAP (Ours)	8.55	6.81
C3D Global and Local + 3D-GAP (Ours)	8.31	**6.59**

VI. CONCLUSIONS

In this paper, we proposed an architecture based on 3D CNNs for recognizing a subject's level of depression based on facial regions captured in videos. The proposed method employs Convolutional 3D networks (C3D) that are pre-trained on large-scale datasets in order to learn spatio-temporal features, where a regression model is used to predict the depression level scores. The architecture performs late fusion of scores from diversified C3D networks trained on local (coarse eye) and global (full-face) facial regions in order to analyze the dynamics of facial appearances over consecutive frames. We also propose 3D Global Average Pooling (3D-GAP) to fuse the extracted global and local features and to reduce number of model parameters. Based on experimental results obtained on benchmark AVEC2013 and AVEC2014 depression datasets, the proposed approach can efficiently learn spatio-temporal features, and provide a lower level of error than state-of-the-art methods. Moreover, the local eye region holds valuable spatio-temporal information to discriminate between normal and depressive behaviours using facial videos. In future work, the proposed architecture will be expanded for other health care issues and applications.

ACKNOWLEDGMENTS

The financial support of the Academy of Finland, Infotech Oulu, and the Natural Sciences and Engineering Research Council of Canada is acknowledged. The first author wishes to thank the State University of Amazonas for the support.

REFERENCES

[1] S. Alghowinem, R. Goecke, M. Wagner, G. Parker and M. Breakspear, "Eye movement analysis for depression detection", in *ICIP 2013*.

[2] 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*, 229: 463-468, 2018.

[3] E. Aragons, J.L. Piol and A. Labad, "The Overdiagnosis of Depression in Non-depressed Patients in Primary Care", *Family Practice*, 23: 363-368, 2006.

[4] A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos, F. Yang, G. Lematre and F. Meriaudeau, "Video-Based Depression Detection Using Local Curvelet Binary Patterns in Pairwise Orthogonal Planes", in *EMBC 2016*.

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

[6] L.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber and N.B. Allen, "Detection of Clinical Depression in Adolescents Speech During Family Interactions", *IEEE Trans. Biomedical Engineering*, 58: 574-586, 2011.

[7] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps and T.F. Quatieri, "A Review of Depression and Suicide Risk Assessment Using Speech Analysis", *Speech Communication*, 71: 10-49, 2015.

[8] 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. Affective Computing*, 2017, pp. 1-27.

[9] X. Ma, H. Yang, Q. Chen, D. Huang and Y. Wang, "DepAudioNet: An Efficient Deep Model for Audio Based Depression Classification", in *AVEC 2016*.

[10] Y. Zhu, Y. Shang, Z. Shao and G. Guo, "Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics", *IEEE Trans. Affective Computing*, 2017, pp. 1-8.

[11] S. Song, L. Shen and M. Valstar, "Human Behaviour-Based Automatic Depression Analysis Using Hand-Crafted Statistics and Deep Learned Spectral Features", in *FG 2018*.

[12] L. Chao, J. Tao, M. Yang and Y. Li, "Multi Task Sequence Learning for Depression Scale Prediction from Video", in *ICACII 2015*.

[13] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker and M. Breakspear, "Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors", *IEEE Trans. Affective Computing*, 2016, pp. 1-14.

[14] A. Krizhevsky, I. Sutskever and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *NIPS 2012*.

[15] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", in *CVPR 2016*.

[16] J. Thevenot, M.B. Lpez and A. Hadid, "A Survey on Computer Vision for Assistive Medical Diagnosis From Faces", *IEEE Journal of Biomedical and Health Informatics*, 22: 1497-1511, 2018.

[17] L. Yang, D. Jiang, X. Xia, E. Pei, M.C. Oveneke and H. Sahli, "Multimodal Measurement of Depression Using Deep Learning Models", in *AVEC 2017*.

[18] M. Parchami, S. Bashbaghi, and E. Granger, "Video-based face recognition using ensemble of Haar-like deep convolutional neural networks," in *IJCNN 2017*.

[19] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features With 3D Convolutional Networks", in *ICCV 2015*.

[20] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou and F.D.L. Torre, "Detecting Depression from Facial Actions and Vocal Prosody", in *Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1-7.

[21] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker and M. Breakspear, "Multimodal Assistive Technologies for Depression Diagnosis and Monitoring", *Journal on MultiModal User Interfaces*, 7: 217-228, 2013.

[22] 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 *AVEC 2013*.

[23] 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 *AVEC 2014*.

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

[25] 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 *AVEC 2013*.

[26] H. Meng and N. Pears, "Descriptive Temporal Template Features for Visual Motion Recognition", *Pattern Recognition Letters*, 30: 1049-1058, 2009.

[27] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", in *CVPR 2005*.

[28] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28: 2037-2041, 2006.

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

[30] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke and J. Epps, "Diagnosis of Depression by Behavioural Signals: A Multimodal Approach", in *AVEC 2013*.

[31] I. Laptev, M. Marszaek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies", in *CVPR 2008*.

[32] A. Bosch, A. Zisserman and X. Munoz, "Representing shape with a spatial pyramid kernel", in *ICIVR 2007*.

[33] L. Wen, X. Li, G. Guo and Y. Zhu, "Automated Depression Diagnosis Based on Facial Dynamic Analysis and Sparse Coding", *IEEE Trans. Information Forensics and Security*, 10: 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 *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 *AVEC 2014*.

[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 *AVEC 2014*.

[37] H. Kaya, F. illi and A.A. Salah, "Ensemble CCA for Continuous Emotion Prediction", in *AVEC 2013*.

[38] 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. Cognitive and Developmental Systems*, 10: 668-680, 2018.

[39] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks", in *CVPR 2014*.

[40] K. Soomro, A.R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild", In *CRCV-TR-12-01 2012*.

[41] L.J. Wells, S.M. Gillespie and P. Rotshtein, "Identification of Emotional Facial Expressions: Effects of Expression, Intensity, and Sex on Eye Gaze", *PLoS One*, 10: 1-20, 2016.

[42] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen *et al*., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", *arXiv:1603.04467*, 2015.

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

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

[45] 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 *ICPRAM 2014*.

[46] H. Kaya and A.A. Salah, "Eyes Whisper Depression: A CCA Based Multimodal Approach", in *ICM 2014*.