ENCODING TEMPORAL INFORMATION FOR AUTOMATIC DEPRESSION RECOGNITION FROM FACIAL ANALYSIS

Wheidima Carneiro de Melo[⋆] *Eric Granger*[†] *Miguel Bordallo Lopez*[‡ ⋆]

[⋆] Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland
[†] LIVIA, Dept. of Systems Engineering, École de technologie supérieure, Montreal, Canada
[‡] VTT Technical Research Centre of Finland

ABSTRACT

Depression is a mental illness that may be harmful to an individual's health. Using deep learning models to recognize the facial expressions of individuals captured in videos has shown promising results for automatic depression detection. Typically, depression levels are recognized using 2D-Convolutional Neural Networks (CNNs) that are trained to extract static features from video frames, which impairs the capture of dynamic spatio-temporal relations. As an alternative, 3D-CNNs may be employed to extract spatio-temporal features from short video clips, although the risk of overfitting increases due to the limited availability of labeled depression video data. To address these issues, we propose a novel temporal pooling method to capture and encode the spatio-temporal dynamic of video clips into an image map. This approach allows fine-tuning a pre-trained 2D CNN to model facial variations, and thereby improving the training process and model accuracy. Our proposed method is based on two-stream model that performs late fusion of appearance and dynamic information. Extensive experiments on two benchmark AVEC datasets indicate that the proposed method is efficient and outperforms the state-of-the-art schemes.

Index Terms— Affective Computing, Depression Detection, Expression Recognition, Temporal Pooling, Two-stream Model

1. INTRODUCTION

Major Depressive Disorder (MDD), or simply depression, is a predominant mental disorder characterized by depreciated thoughts, negative feelings, impaired well-being, and behavioral changes [1]. Studies have estimated over 350 million people around the world suffer from depression [2], with more prevalence in females than in males [3]. MDD can elevate the chances of suffering other conditions, such as cancer or heart disease [4]. In the most severe cases, the patient with depression has a high risk committing suicide.

Typical treatments for MDD include tranquilizers, antidepressants and psychotherapy, and they present relatively good efficiency. However, misdiagnosis is frequent in the evaluation of MDD. The reason is that clinical diagnosis is strongly dependent on the subjective analysis of a physician. In fact, the procedure for depression identification is based on a clinical interview that evaluates the patient following the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) specification [5]. The level of depression is determined using a score given by answering questionnaires that are graded following the Hamilton Rating Scale or the Beck Depression Inventory-II [5]. Moreover, several clinical evaluations are normally necessary to infer a consistent conclusion about the level of depression.

Considering the challenges of the evaluation, an objective assessment of MDD from facial information has attracted great interest in the computer vision community. In the clinical literature, it can be found that facial expressions convey evidences of depressive states [6]. The depressed subject has characteristic sad or neutral expressions, and can look tired and worried [6, 5]. The patient can present reduced levels of social behaviors, such as a reduction of facial movements, avoidance of eye contact, and fewer smiles [5]. Depression detection systems can therefore benefit from spatio-temporal information that is captures in facial videos.

Automated Facial Expression Recognition (FER) has been a topic of significant interest for several years. Many techniques have been proposed, often to detect the seven universally recognizable types of emotions – joy, surprise, anger, fear, disgust, sadness and neutral – but also to estimate the level of valence and arousal, or intensity of affective states like fatigue, stress and depression. More recently, spatio-temporal FER techniques have emerged as a promising approach to improve performance, where the expression is estimated from a sequence of consecutive video frames captured during the physical facial expression process of an individual. Deep learning (DL) models, and in particular Convolutional Neural Networks (CNNs), currently achieve state-of-the-art performance in many visual recognition applications like FER [7]. In specific, for automatic depression recognition, the architectures usually predict the level of depression by extracting features from each frame of facial

*The financial support of the Academy of Finland, Infotech Oulu and State University of Amazonas is acknowledged.

videos. In this context, several methods explore spatial and temporal information indirectly, by employing 2D Convolutional filters and a scheme that analyze the alteration of the features, or by pooling the level of depression in every frame. For instance, Jan *et al.* [8] employed a 2D CNN to explore spatial information and Feature Dynamic History Histogram (FDHH) to map variations in the features. Zhou *et al.* [9] used multiple 2D CNNs to investigate different facial regions with a scheme to combine all responses. In [10], the authors proposed to integrate the facial features by using an attention mechanism to fuse features from a video. Such approaches consider the spatial dependencies of extracted features, disregarding the temporal information between the frames, which can impair the exploitation of spatio-temporal relations. As an alternative solution, some authors have used 3D-CNNs to leverage spatio-temporal information [3, 11]. However such models require greater training and inference times, and require optimizing a large number of parameters compared to the size of publicly-available training data, which greatly increases the risk of overfitting.

In our contribution, we elaborating a novel method for depression recognition that encodes the temporal information. In this process, facial temporal variations are encoded by capturing and summarizing the dynamics of videos based on binary codes. The encoded image has the potential to favour the exploitation of spatio-temporal information by an 2D-CNN. Our proposed method is comprised of a two-stream model which explores spatial and temporal information, and combining the predictions of each stream through score-level fusion. In this way, our proposed approach improves the training process, and decreases the need for large amounts of labeled training data. Extensive experiments are conducted on the AVEC2013 and AVEC2014 depression benchmark datasets, and the results indicate the effectiveness of our proposed method.

2. PROPOSED ARCHITECTURE

The deep architectures explore spatial and temporal dependencies to learn discriminative representations from face videos. This process requires a large amount of high-quality labeled training data. For healthcare applications, and specifically depression detection, labeling data is expensive and normally, only limited labeled data is available for training deep models. Consequently, models created using solely these data have reduced potential capture the dynamics of the videos. It is important to note that modeling both facial appearance, structures and dynamics is crucial for facial analysis due to the correlation with depression patterns. In this context, we introduce a new temporal pooling scheme for encoding the dynamics of the video into an image map which is posteriorly employed as input of an 2D CNN model.

The image map can convey information about movement and velocity in the generated texture, which favors the exploitation of dynamics information. Suppose $x[m, n, t]$ denotes the input video or segment, with m and n representing spatial coordinates, and t the temporal coordinate. The first step of proposed method is calculate the difference around the middle point by

$$z[m, n, t] = \sum_{j=0}^{l-1} x[m, n, j]\delta[m, n, j - t] - x[m, n, (l - 1)/2], \tag{1}$$

where $z[m, n, t]$ denotes the output signal, $\delta[m, n, t]$ represents impulse signal, and l is the number of frames in the video clip. As the value in the middle point is zero, we simply remove it. In order to map the direction of the variation, we employ a step activation function defined by

$$z[m, n, t] = \begin{cases} 0, & \text{if } z[m, n, t] \geq 0 \\ 1, & \text{otherwise} \end{cases}, \tag{2}$$

the resulting signal is mapped using two operations. First, sequences of size α are encoded using XOR operation. Then, a binary code is employed to generate the final value. The latter operation can be described by the equation

$$y[m, n] = \sum_{j=0}^{h-1} z[m, n, j] \cdot 2^j, \tag{3}$$

where $y[m, n]$ is the image representation which take advantage of transfer learning when used as the input of pre-trained deep architectures. We define $h = 8$ and the number of frames $l = 2 * \alpha + 1$. In this way, the image map has 8 bits per pixel. Figure 1 illustrates our temporal pooling method.

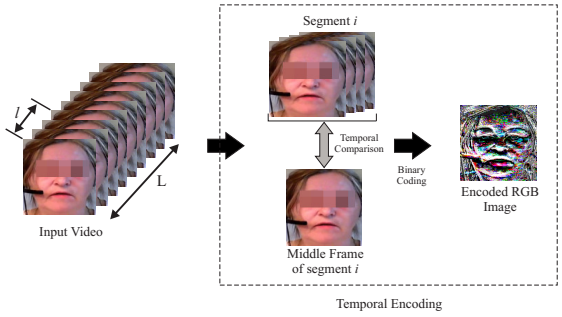

Fig. 1. An overview of the proposed method to encode temporal information. L is the number of frames in an input video.

The image map can be fed to any deep architectures used for images. In this work, ResNet-50 [12] model is employed to explore the dependencies of the encoded image. ResNet architectures use identity shortcut connections which contribute to the training of deeper networks. ResNet-50 is comprised of basic blocks that include a stack of convolutional layers with 1×1, 3×3 and 1×1 kernels. The Global Average Pooling (GAP) summarizes the features produced by the last convolutional layer. The classification stage consists of a fully

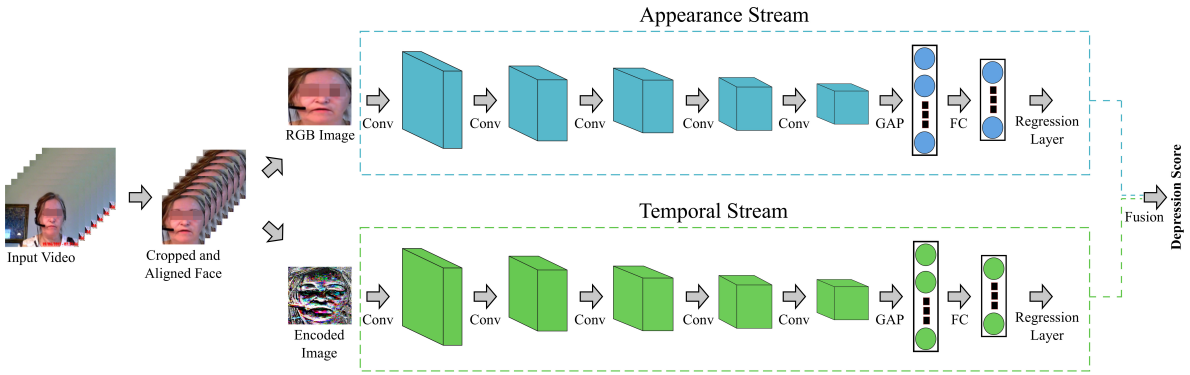

Fig. 2. A representation of the proposed method for automatic depression recognition.

connected layer with 512 neurons and a regression layer that computes the estimated depression level.

As the facial appearance convey valuable static and dynamic information related to depression patterns, our proposed architecture is a two-stream Convolutional Network scheme that includes both temporal and spatial networks. Figure 2 illustrates an overview of our proposed method. The appearance stream explores facial expressions and spatial structures from still images. Both streams employ the same architecture, i.e. ResNet-50, but the networks are trained separately. The mean square error function is employed to solve the regression problem. The error is obtained by calculating the Euclidean distance between the estimated output and the ground truth. A score fusion scheme determines the estimated depression level for the input video clip by simply averaging the scores computed separately by the individual networks belonging to the appearance and temporal streams.

3. EXPERIMENTAL ANALYSIS

Datasets: Our proposed method is employed for automatically estimating the level of depression from facial information. For performance evaluation, extensive experiments were conducted on two benchmarking and publicly available databases, namely the Audio/Visual Emotion Challenge (AVEC) 2013 [13] and 2014 [14] depression sub-challenge datasets. Both datasets are derived from a subset of the audio-visual depressive language corpus (AViD-Corpus). The AVEC2013 dataset is constituted of 150 videos from 82 individuals, and has three partitions: training, development and test set, containing 50 videos each. In AVEC2014 dataset, the individuals are recorded while performing two different tasks: Freeform and Northwind. In both tasks, the recordings are segmented into three partitions: training, development and test set with 50 videos in each partition. Every video is labeled with a depression score with range from 0 to 63.

Protocol and performance measures: The first step of our proposed method is face pre-processing, which is responsible for obtaining the facial regions, as shown in Figure 2. Following the authors in [11], facial cropping and alignment are performed using facial landmarks extracted with Multi-task Cascade Convolutional Networks (MTCNN) [15]. The resulting facial regions are rescaled to 224×224 pixels. To take advantage of transfer learning, the filter weights for both streams of the proposed method are initialized with pre-trained layers on VGG Face dataset [16]. The streams of the model are fine-tuned on AVEC2013 or AVEC2014 employing the ADAM optimizer with learning rate of 0.0001. Data augmentation is used in the training stage: the frames are horizontally flipped with 50% probability and randomly rotated around two axes with $+/-30$ degrees. The training samples generated maintain the same depression level as their original videos.

Table 1. Effect of different values of l for temporal stream.

	Number of Frames			
	$l = 9$	$l = 17$	$l = 25$	$l = 33$
RMSE	8.24	9.20	8.93	9.33
MAE	6.29	6.76	6.79	7.36

In order to evaluate the performance of the proposed architecture and make a fair comparison with the state-of-the-art methods, two metrics are employed: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). For an input video, the overall predicted depression score is obtained by averaging the scores estimated over every video clip.

Experimental results: In the Table 1, we evaluate the performance of the temporal stream using different length of video clips which are encoded by our temporal pooling method. As our approach explore face videos and AVEC2013 has few subjects in common with AVEC2014, we present this analysis on AVEC2014. As can be seen, in terms of MAE, the error increases when the size of video clip increases. The best performance is achieved when the number of frames is 9, mainly for RMSE values. These results show that the spatial structures generated in longer sequences have limited capability to favor the exploitation of various temporal ranges, however our approach is efficient to explore short-term temporal information.

In the Table 2, we present results for depression detection on AVEC2013 and AVEC2014 datasets. The table depicts

Table 2. Comparison between the different streams.

Methods	AVEC2013		AVEC2014	
	RMSE	**MAE**	**RMSE**	**MAE**
Appearance Stream	8.18	6.15	8.08	6.30
Temporal Stream	8.43	6.51	8.24	6.29
Two-Stream model	**7.97**	**5.96**	**7.94**	**6.20**

the performance of both the appearance stream and temporal stream (using $l = 9$) separately as well their combination as a two-stream model. The results obtained by the temporal stream are competitive with the ones from appearance stream. For AVEC2014, the temporal stream achieves better performance in terms of MAE. These results demonstrate that our temporal pooling method can capture and summarize the dynamics of the facial video into an image map in such a way that favors the exploration of dependencies of the image. Moreover, we can see that the two-stream model, which integrates the two depression detection streams, outperforms both individual networks.

Table 3 compares the results of our proposed method against state-of-the-art methods on AVEC2013 dataset. Hand-engineered representations are employed in [13, 17, 18, 19, 20]. Our proposed method outperforms all schemes based on these techniques. A comparison with the method in [21] is very interesting, since such model also employs a two-stream model to explore dynamics and spatial information. However, the authors generate the temporal information using optical flow. Our proposed method outperforms such model by a notable margin, even when considering only our temporal stream. Our proposed method outperforms the models in [11, 3] which explore directly spatio-temporal information using a two-stream 3D CNN. These results confirm our assumption that exploring temporal and spatial information separately using an efficient method to capture the facial dynamics can reduce the overfitting problem. In addition, our proposed method achieves better results than the method in [9] which uses a four-stream model to explore multiple facial regions, showing the importance of exploring the temporal information for depression detection. Finally, the authors in [22] explore spatial information and employ attention mechanism to fuse facial features. Our proposed method outperforms such method, which suggest that exploring temporal information between the frames is a better approach.

Table 4 shows the results of proposed method compared with state-of-the-art methods on AVEC2014 dataset. Our proposed method achieves better results than the schemes in [14, 23, 24] which are based on handcrafted features. The proposed method also outperforms the deep learning schemes proposed in [21, 3, 11, 22, 9] on AVEC2014, confirming the good performance of our model. In [25], the method is based on distribution learning with expectation loss function. The proposed method outperforms, in terms of RMSE, such method. However, using distribution learning seems to be an interesting alternative for Euclidean loss. Observe that our method obtains better results on AVEC2013. A specific deep model is proposed in [26] to extract features from facial frames, while FDHH is used to capture temporal feature variation. Our proposed method outperforms this scheme. In summary, these results show the effectiveness of our two-stream model in determining depression level from facial appearance and dynamics.

Table 3. Comparison of methods for predicting the level of depression on the AVEC2013 dataset.

Methods	RMSE	MAE
Baseline	13.61	10.88
LPQ + SVR (Kächele *et al.*)	10.82	8.97
MHH + LBP (Meng *et al.*)	11.19	9.14
LPQ-TOP + MFA (Wen *et al.*)	10.27	8.22
LPQ + Geo (Kaya *et al.*)	9.72	7.86
Two DCNN (Zhu *et al.*)	9.82	7.58
C3D (Jazaery *et al.*)	9.28	7.37
C3D (Melo *et al.*)	8.26	6.40
ResNet-50 + Pool (Zhou *et al.*)	8.43	6.37
Four DCNN (Zhou *et al.*)	8.28	6.20
ResNet-50 (Melo *et al.*)	8.25	6.30
Ours (two-stream model)	**7.97**	**5.96**

Table 4. Comparison of methods for predicting the level of depression on the AVEC2014 dataset.

Methods	RMSE	MAE
Baseline	10.86	8.86
MHH + PLS (Jan *et al.*)	10.50	8.44
LGBP-TOP + LPQ (Kaya *et al.*)	10.27	8.20
Two DCNN (Zhu *et al.*)	9.55	7.47
C3D (Jazaery *et al.*)	9.20	7.22
C3D (Melo *et al.*)	8.31	6.59
VGG + FDHH (Jan *et al.*)	8.04	6.68
ResNet-50 + Pool (Zhou *et al.*)	8.43	6.37
Four DCNN (Zhou *et al.*)	8.39	6.21
ResNet-50 (Melo *et al.*)	8.23	**6.15**
Ours (two-stream model)	**7.94**	6.20

4. CONCLUSION

This paper introduced a two-stream model for automated depression detection from face videos. A new temporal pooling method is proposed to encode the dynamics of facial expressions into an image map which is employed as input of ResNet-50 architecture, composing stream that encodes temporal information. The encoded image can be used in any existing deep methods and has the potential to favor the exploitation of temporal information by its texture. The spatial information is complementary explored by an appearance stream. A score fusion scheme combine both streams. Extensive experiments on public AVEC2013 and AVEC2014 datasets indicated the efficiency of the proposed method compared to different methods present in the literature.

5. REFERENCES

[1] 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.

[2] T. Vos *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the global burden of disease study 2013," *The Lancet*, vol. 386.

[3] 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.

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

[5] 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.

[6] H. Ellgring, *Non-verbal Communication in Depression*, Cambridge University Press, New York, 2007.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*.

[8] A. Jan, H. Meng, Y.F.B.A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. on Cognitive and Developmental Systems*, vol. 10, pp. 668–680, 2018.

[9] 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.

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

[11] W.C. de Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3d networks for detecting depression from facial expressions," in *FG 2019*.

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

[13] M. Valstar *et al.*, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *AVEC 2013*.

[14] M. Valstar *et al.*, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *AVEC 2014*.

[15] 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.

[16] O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015*.

[17] M. Kächele *et al.*, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," in *ICPRAM 2014*.

[18] H. Meng *et al.*, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *AVEC 2013*.

[19] 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.

[20] H. Kaya and A.A. Salah, "Eyes whisper depression: A cca based multimodal approach," in *ACM MM 2014*.

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

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

[23] 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*.

[24] H. Kaya, F. Çilli, and A.A. Salah, "Ensemble cca for continuous emotion prediction," in *AVEC 2014*.

[25] W.C. de Melo, E. Granger, and A. Hadid, "Depression detection based on deep distribution learning," in *ICIP 2019*.

[26] A. Jan, H. Meng, Y.F.B.A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Trans. on Cognitive and Developmental Systems*, vol. 10, pp. 668–680, 2018.