DEPRESSION DETECTION BASED ON DEEP DISTRIBUTION LEARNING

Wheidima Carneiro de Melo[⋆] *Eric Granger*[†] *Abdenour Hadid*[⋆]

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

ABSTRACT

Major depressive disorder is among the most common and harmful mental health problems. Several deep learning architectures have been proposed for video-based detection of depression based on the facial expressions of subjects. To predict the depression level, these architectures are often modeled for regression with Euclidean loss. Consequently, they do not leverage the data distribution, nor explore the ordinal relationship between facial images and depression levels, and have limited robustness to noisy and uncertain labeling. This paper introduces a deep learning architecture for accurately predicting depression levels through distribution learning. It relies on a new expectation loss function that allows to estimate the underlying data distribution over depression levels, where expected values of the distribution are optimized to approach the ground-truth levels. The proposed approach can produce accurate predictions of depression levels even under label uncertainty. Extensive experiments on the AVEC2013 and AVEC2014 datasets indicate that the proposed architecture represents an effective approach that can outperform state-of-the-art techniques.

***Index Terms*—** Affective Computing, Depression Detection, Deep Distribution Learning, Convolutional Neural Nets.

1. INTRODUCTION

Depression has been recognized as a cause of disability worldwide with a considerable cost to health care systems [1]. It encompasses negative thoughts and feelings, impacts physical well-being and behaviours, and in more severe cases, depression is considered as one of the leading causes of suicide and substance abuse. Although antidepressant medications and psychotherapy are effective treatments for depression, medical errors in clinical evaluation of depression are common [2]. Indeed, assessment of depression 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 evaluated based on a score obtained by answering Hamilton Rating Scale or Beck Depression Inventory-II.

Given the harmful effects of depression on individuals and society, pattern recognition and computer vision communities have proposed methods that are based on verbal and nonverbal information for accurate estimation of a subject's depression level. The patterns of change in audio and visual information have been exploited for automated, contact-free analysis and diagnosis of depressive behaviors. Some authors have exposed that attributes of voice and speech alter in depressed subjects [3]. For the visual information, face and body carry important cues, such as facial expressions, eye blinks, and head pose and movement [4]. For example, Jan *et al.* [5] combined features extracted from facial expressions using a deep learning model, and vocal expressions using regression techniques.

This paper focuses on techniques for accurate detection of depression levels based on faces captured in videos. In this context, designing a system capable of encoding discriminant features for depression detection is challenging because the appearance of faces may vary considerably according to the specific subject, capture conditions (pose, illumination, blur), and sensors. It is difficult to encode common and discriminant spatial features of depression while suppressing these context- and subject-specific facial variations. Moreover, collecting and annotating a large-scale video dataset is challenging, where labels incorporate some noise and uncertainty.

Deep learning architectures currently provide state-of-the-art performance for depression detection. These complex architectures can exploit spatial and temporal information using 2D-CNNs, 3D-CNNs, RNNs, etc. For example, in [6], the authors proposed a two-stream deep model to explore facial expressions from facial images and optical flows. 3D-CNNs are employed by Melo *et al.* [7] and Jazaery *et al.* [8] in order to encode spatio-temporal information, with an RNN is used in [8] to encode the sequential features. Zhou *et al.* [9] employed deep learning model to explore multiple regions of the face with a scheme to fuse the response from different facial regions. Most of these state-of-the-art architecture are based on regression techniques, often using Euclidean loss to penalize the differences between the predicted and ground-truth depression [6, 8, 9]. However, such loss functions are based on labeled facial images, and do not explicitly explore the ordinal relationship between the facial images and depression levels. In addition, in order to improve performance, the architectures tend to employ more than one channel to explore multi-regions of facial frames, which further increases the

978-1-5386-6249-6/19/$31.00 ©2019 IEEE

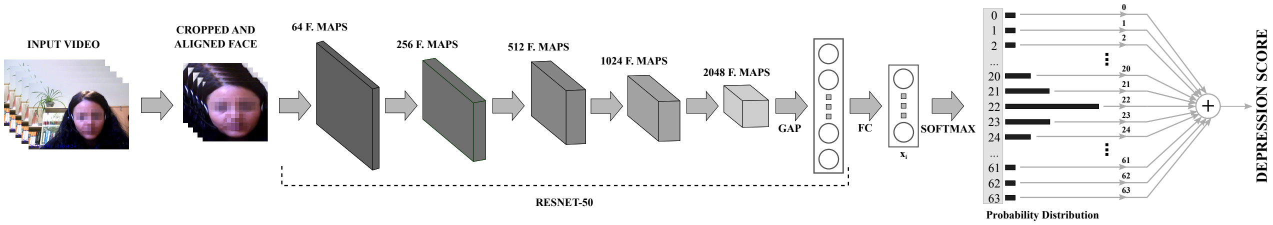


Fig. 1. The proposed method to perform depression analysis. The softmax arrow represents the last FC and softmax function.

model complexity.

In this paper, a deep learning architecture is introduced to accurately predict depression levels, where distribution learning is proposed to model the ordinal relationship between the facial images and depression levels. Our method relies on a new expectation loss to estimate underlying depression distributions, and thereby increases performance of the model without the need to employ multiple channels. In contrast with methods in literature, the proposed loss penalizes the distance between the expected value of a predicted depression distribution and the ground-truth depression levels. Consequently, the proposed approach can address problems with noisy and ambiguous labels since the model explores the relationship between the facial images and depression levels. This leads to effective architectures that model the data distribution over all depression levels, which allow to predict robust depression levels. The performance of our proposed approach was compared favourably against several state-of the-art (conventional and deep learning) techniques for depression detection on two publicly available Audio Visual Emotion Challenge (AVEC) datasets – AVEC2013 and AVEC2014.

2. PROPOSED ARCHITECTURE

Distribution learning allows assigning a label distribution to an object rather than a single or multiple label [10]. When a model learns the distribution related to a label space for a sample, it indicates the level of significance of each label present in this space. Then, during inference, the method can explore this distribution in order to improve its predictive accuracy. Distribution learning has been successfully employed in tasks like emotion recognition [11] and age estimation [12]. For example, Pan *et al.* [12] employed distribution learning to embed mean-variance loss and softmax loss into a deep neural network. However, these methods assume that the mean and variance are available during training, which is not the case in applications of depression detection. Moreover, with these methods, the learning process relies on statistical moments of a distribution and softmax loss. This can impair the learning process for depression detection because statistical moments are used in a complementary way, and softmax does not allow to explore the distribution.

Figure 1 illustrates the deep learning architecture proposed to predict depression levels. Although applicable in a wide range of CNNs, expectation loss is embedded into ResNet-50 [13] for distribution learning. ResNet architectures have shown the efficiency of deeper networks by using identity shortcut connections. In the ResNet-50 model, the basic blocks are comprised of a stack of convolutional layers with 1×1, 3×3 and 1×1 kernels. After the last convolutional layer, the features are summarized by using Global Average Pooling (GAP). The classification stage is composed of a fully connected layer with 512 neurons, softmax layer, and a step to compute the expected values of the depression distribution. Note that the input of the proposed architecture is a video. Consequently, in order to predict the final depression score in testing mode, the average of all predicted values is computed for each frame.

In this paper, expectation loss is proposed to estimate the data distribution over depression levels, allowing robust predictions given the ambiguous depression levels. This loss seeks to penalize the difference between ground truth depression levels and the expected values of a depression distribution. Let $y_i \in \{0, 1, .., K-1\}$ represents the depression level label of input i, $\mathbf{x_i}$ represents the feature representation, and $\mathbf{z} = f(\mathbf{x_i}) \in R^{N \times K}$ denotes the output of the last fully connected layer. In this case, the softmax probability can be obtained using:

$$p_{i,j} = \frac{e^{z_{i,j}}}{\sum\limits_{m=0}^{K-1} e^{z_{i,m}}}, \tag{1}$$

where $p_{i,j}$ represents the probability that input i belongs to class (depression level label) j, $\mathbf{p_i}$ is the depression distribution estimated for sample i, and $z_{i,j}$ denotes component j of $f(\mathbf{x_i})$. The expected value E_i from the depression distribution $\mathbf{p_i}$ of input i is computed using:

$$E_i = \sum_{j=0}^{K-1} j \cdot p_{i,j}, \tag{2}$$

where j represents the depression labels.

Based on Eq. 2, the expectation loss function of the proposed method for distribution learning is defined as:

$$L = \frac{1}{2M} \sum_{i=0}^{M-1} (E_i - y_i)^2, \qquad (3)$$

where M is the batch size. The distance between each expected value and depression level label is computed using L_2 metric. As a result, we expect that during the training, the model progressively learns a distribution whose expected value approaches the depression level label.

3. EXPERIMENTAL ANALYSIS

Datasets: The proposed architecture is employed for automatically predicting the level of depression in subjects. For performance evaluation, experiments are conducted on AVEC2013 and AVEC2014 depression sub-challenge datasets. The objective of the sub-challenge is to estimate depression level of subjects on the Beck Depression Inventory (BDI). The BDI scores range from 0 to 63 with the following definitions: 0-13 (none depression), 14-19 (mild depression), 20-28 (moderate depression), and 29-63 (sever depression).

The AVEC2013 depression dataset is a subset of the audio-visual depressive language corpus (AViD-Corpus), which is composed of 150 videos from 82 subjects. The dataset is divided into three partitions: training, development and test set of 50 videos each. The AVEC2014 depression dataset is also a subset of AViD-Corpus. The subjects are recorded using a webcam and a microphone while performing two tasks: Freeform task, participants respond to 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. Following the approach in [6, 8, 9], we down-sampled each video of datasets, where the frame rate is decreased by a factor of 100 and 10, for AVEC2013 and AVEC2014, respectively. Some samples from both datasets are shown in Figure 2 (we blurred the samples for privacy concerns).

Fig. 2. AVEC2014 (bottom) and AVEC2013 (top) samples.

Protocol and performance measures: In our proposed model, face pre-processing is first performed in order to obtain the facial regions. Multi-task Cascade Convolutional Networks (MTCNN) [14] is employed for detecting facial landmarks. Based on these points, cropping and alignment of the faces are done. The resulting facial images have a size of 224×224 pixels. The convolutional layers of the proposed model are initialized with parameters which were pre-trained on VGG Face dataset [15]. The fine-tuning process is applied to the model using ADAM optimizer with learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1 \times 10^{-8}$. The proposed network is implemented using Keras framework [16].

Table 1. RMSE and MSE for depression predictions obtained by different losses on AVEC2013 and AVEC2014.

Methods	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
Softmax loss	10.22	7.50	10.37	7.89
Euclidean loss	8.48	6.72	8.71	6.37
Expectation loss	**8.25**	**6.30**	**8.23**	**6.15**

The performance of the proposed method for depression detection is evaluated in terms of 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 each frame of the video.

Experimental results: In order to assess the efficiency, we compared the RMSE and MAE performance of our proposed approach with expectation, softmax and Euclidean loss functions. With softmax, the weights of the function are used to compute the depression level just in the testing phase. The evaluation is carried out using ResNet-50 on AVEC2013 and AVEC2014 datasets. As Table 1 shows, the results when using softmax loss are worse than those employing Euclidean loss for depression prediction. Using only softmax during the training phase does not create a sharp probability distributions which impair the prediction. When the expectation loss is employed, the model achieves the best performance. This can be explained by the fact that during the training phase, the model yields a more compact distribution in the direction of ground-truth depression level, and this increases the performance of the model.

In Table 2, the performance of our proposed method is compared with state-of-the-art methods on AVEC2013 dataset. The methods in [17, 18, 19, 20, 21] are based on handcrafted features. For instance, Local Phase Quantization (LPQ) is employed as baseline method in AVEC2013 competition [17]. Results show that the proposed method outperforms all others. The methods proposed by Zhu *et al.* [6], Jazaery *et al.* [8] and Melo *et al.* [7] encode spatial and temporal information. It is worth noting that, even considering only spatial information, our proposed method achieves better results than these methods. Moreover, in [9], the authors employed a four-stream deep network to explore multi-region of face, where each stream is a ResNet-50 model. As we can see, our proposed method yields competi-

Table 2. Comparison of methods for predicting the level of depression on the AVEC2013 dataset.

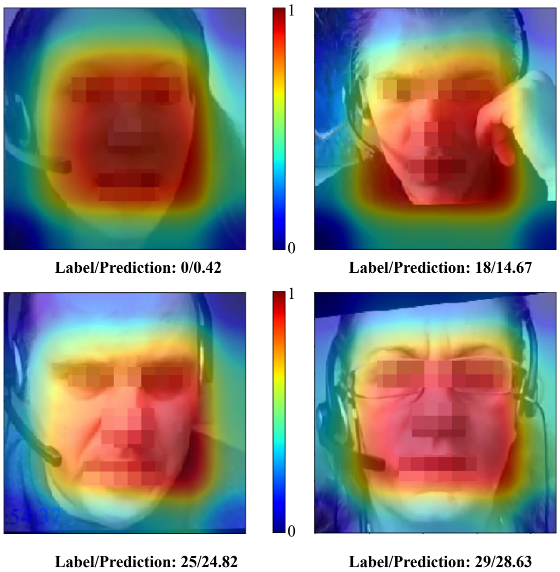
Methods	RMSE	MAE
Baseline [17]	13.61	10.88
LPQ + SVR (Kächele *et al.* [18])	10.82	8.97
MHH + LBP (Meng *et al.* [19])	11.19	9.14
LPQ-TOP + MFA (Wen *et al.* [20])	10.27	8.22
LPQ + Geo (Kaya *et al.* [21])	9.72	7.86
Two DCNN (Zhu *et al.* [6])	9.82	7.58
C3D (Jazaery *et al.* [8])	9.28	7.37
C3D (Melo *et al.* [7])	8.26	6.40
Four DCNN (Zhou *et al.* [9])	8.28	**6.20**
Ours (ResNet-50 + Expectation loss)	**8.25**	6.30

tive results compared to the method in [9], while using only one ResNet-50 model. This suggest the good generalization power of our model.

Table 3. Comparison of methods for predicting the level of depression on the AVEC2014 dataset.

Methods	RMSE	MAE
Baseline [22]	10.86	8.86
MHH + PLS (Jan *et al.* [23])	10.50	8.44
LGBP-TOP + LPQ (Kaya *et al.* [24])	10.27	8.20
Two DCNN (Zhu *et al.* [6])	9.55	7.47
C3D (Jazaery *et al.* [8])	9.20	7.22
C3D (Melo *et al.* [7])	8.31	6.59
VGG + FDHH (Jan *et al.* [5])	**8.04**	6.68
Four DCNN (Zhou *et al.* [9])	8.39	6.21
Ours (ResNet-50 + Expectation loss)	8.23	**6.15**

Table 3 compares the MAE and RMSE accuracy of proposed and related methods on the AVEC2014 dataset. The methods based on handcrafted features are in [22, 23, 24]. We can cite as an example, the baseline method provided by AVEC2014 competition which is based on Local Binary Pattern (LBP) and LPQ. Our proposed method achieves better results than the ones based on these handcrafted descriptors. Once again, the proposed method outperforms the methods proposed by Zhu *et al.* [6], Jazaery *et al.* [8] and Melo *et al.* [7]. In [5], the authors proposed a deep learning model to explore facial frames and employed feature dynamic history histogram (FDHH) to capture variations in the features. Our method achieves better results in terms of MAE than the method in [5], while this later is better in terms of RMSE. This indicates that the exploration of temporal redundancies can also contribute to depression detection. Finally, our method outperforms the method in [9] in terms of both MAE and RMSE. Such a method achieves good results, but it requires large memory and computational cost consumption since it is used with four ResNet-50 models which results in around 94 million parameters for feature extraction whereas our proposed method employs less than one forth of total number of parameters.



Label/Prediction: 0/0.42

Label/Prediction: 18/14.67

Label/Prediction: 25/24.82

Label/Prediction: 29/28.63

Fig. 3. Attention maps for different depression levels.

Figure 3 shows the attention maps generated by the proposed method from samples of each severity level classification. This information allows to gain insight into the regions that most contribute to depression detection. In general, the model focuses attention on the central region of the face, which represents the area between the eyes and the mouth. This is reasonable because the model explores facial expressions that provide important clues about depression. Moreover, some movements, such as when the subjects put their hands on their face, may impair the performance of the model since they make it difficult to analyze facial expressions.

4. CONCLUSION

In this paper, a deep distribution learning architecture is introduced for automated depression detection. Expectation loss is proposed to train the model in order to estimate a data distribution over depression levels, where the expected value of the estimated depression distribution provides the predicted value. This distribution allows to explore the ordinal relationships between facial images and depression levels, and has the potential to improve accuracy of the model without additional streams. Experiments on public AVEC2013 and AVEC2014 datasets indicated that our proposed approach yields in interesting results compared to related works in the literature.

5. ACKNOWLEDGEMENT

The financial support of the Academy of Finland and Infotech Oulu is acknowledged. W. C. de Melo would like to thank the State University of Amazonas for its support.

6. REFERENCES

[1] C. S. Dewa, N. Chau, and S. Dermer, "Examining the comparative incidence and costs of physical and mental health-related disabilities in an employed population," *J. of Occupational and Environmental Medicine*, vol. 52, no. 7, pp. 758–762, 2010.

[2] E. Aragons, J. L. Piol, and A. Labad, "The overdiagnosis of depression in non-depressed patients in primary care," *Family Practice*, vol. 23, pp. 363–368, 2006.

[3] 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. on Biomedical Engineering*, vol. 58, pp. 574–586, 2011.

[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*, pp. 1–27, 2017.

[5] 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.

[6] 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, pp. 578–584, 2018.

[7] 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.

[8] 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.

[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. Geng, C. Yin, and Z. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2401–2412, 2013.

[11] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *ICM 2015.*

[12] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *CVPR 2018.*

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

[14] 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.

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

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

[17] M. Valstar *et al.*, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *ACM Int'l workshop on Audio/visual emotion challenge*, 2013.

[18] 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 *Int'l Conf. on Pattern Recognition Applications and Methods*, 2014.

[19] H. Meng *et al.*, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Int'l W on Audio/visual Emotion Challenge*, 2013.

[20] 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.

[21] H. Kaya and A. A. Salah, "Eyes whisper depression: A cca based multimodal approach," in *ACM Int'l Conference on Multimedia*, 2014.

[22] M. Valstar *et al.*, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Int'l W. on Audio/Visual Emotion Challenge*, 2014.

[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 *Int'l W. on Audio/Visual Emotion Challenge*, 2014.

[24] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Int'l W. on Audio/Visual Emotion Challenge*, 2014.