Pattern Recognition Letters 150 (2021) 115–121

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Sequential fusion of facial appearance and dynamics for depression recognition

Qian Chen[a], Iti Chaturvedi[b], Shaoxiong Ji[c], Erik Cambria[a,*]

[a] Nanyang Technological University, Singapore 639798, Singapore
[b] James Cook University, Townsville 4814, Australia
[c] Aalto University, Espoo 02150, Finland

ARTICLE INFO

Article history:
Received 29 March 2021
Revised 9 July 2021
Accepted 10 July 2021
Available online 21 July 2021

Edited by : xxx

Keywords:
Depression recognition
Facial representation
Convolutional neural network
Multimodal learning
Sequential fusion

ABSTRACT

In mental health assessment, it is validated that nonverbal cues like facial expressions can be indicative of depressive disorders. Recently, the multimodal fusion of facial appearance and dynamics based on convolutional neural networks has demonstrated encouraging performance in depression analysis. However, correlation and complementarity between different visual modalities have not been well studied in prior methods. In this paper, we propose a sequential fusion method for facial depression recognition. For mining the correlated and complementary depression patterns in multimodal learning, a chained-fusion mechanism is introduced to jointly learn facial appearance and dynamics in a unified framework. We show that such sequential fusion can provide a probabilistic perspective of the model correlation and complementarity between two different data modalities for improved depression recognition. Results on a benchmark dataset show the superiority of our method against several state-of-the-art alternatives.

1. Introduction

Major depressive disorder (MDD) is a psychological disorder that exhibits feelings of sadness, loss, or anger that may impact a person's usual social activities. At a global level, over 300 million people of all ages suffer from different levels of depression, equivalent to 4.4% of the world's population [1]. A depressive episode can be classified into a mild, moderate, or severe level, depending on the symptoms. Mild depression may bring difficulties in continuing with ordinary work and social activities. More seriously, the feeling of depression may occur comorbidity with self-mutilation [2], and depressed people are more likely to commit suicide than the general population [3]. Early detection of depressive or other mental disorders provides a possible way for mental intervention [4].

In clinical practice, the diagnosis procedure for MDD can usually be labor-intensive and highly relies on expertise observations. Due to the increasing number of people suffering from MDD around the world, methods for automated depression analysis appear to be urgent for objective and efficient diagnosis. Recently, automated depression diagnosis based on computer vision techniques has drawn increasing attention [5], and the significance of the verbal cues for depression analysis has been demonstrated in various depression detection/recognition tasks [6–11]. Besides, visual cues like facial expression and facial dynamics have also proven to be effective in depression analysis [12–15]. This paper investigates facial depression recognition, aiming to predict the depression level for a given face video based on the BDI-II metric [16].

While encouraging progress has been made over the past few years, automated depression analysis in videos remains challenging due to the following reasons. On the one hand, unlike those large-scale image datasets (e.g., ImageNet [17]) for visual recognition [18], the size of most existing depression datasets (e.g., AVEC 2014 [5]) is relatively small due to the privacy concerns. While representation learning based on convolutional neural network (CNN) has been proved to be more effective than hand-crafted descriptors in visual-based depression recognition [15], the lack of labeled data makes the model training with deep networks prone to over-fitting in practice.

On the other hand, many learning methods in the literature have been devoted to multimodal fusion of audio and/or video features for depression recognition [9,14,15], which have demonstrated boosted recognition performance by exploiting the complementary information encoded in different modalities. However, essential correlation and diversity between different visual modalities have not been well investigated in previous visual-based methods, especially for multimodal fusion of visual cues with deep CNN architecture.

* Corresponding author.
E-mail address: cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.patrec.2021.07.005
0167-8655/

In this paper, we propose a multimodal deep learning approach for facial depression recognition to address these issues. Specifically, a sequential fusion of facial appearance and dynamics is introduced to facilitate such multimodal representation learning. Here, facial appearance and dynamics are adopted as the basis modalities in our multimodal fusion framework, as they have been validated to be effective in visual-based depression diagnosis [15]. Hence, a fusion between the two modalities is first operated on the blocks of a two-stream CNN. By such fusion of mid-level features in the CNN training, an initial interaction is conducted to optimize the complementary patterns. Then, the extracted feature, together with the predicted label from the first stream (e.g., RGB modality), is fed into the second stream (e.g., Optical Flow modality) to refine the final prediction. We show that such sequential fusion can provide a probabilistic perspective about model correlation and complementarity between two different data modalities for improved depression recognition. We conduct experiments on the benchmark dataset (AVEC2014), and the results show the superiority of our method against several state-of-the-art alternatives. The main contributions of this paper are:

- We proposed a sequential chained-fusion approach for depression recognition. With a probabilistic perspective, the proposed approach models the correlation and complementarity between facial appearance and facial dynamics at several network layers, such that the complementary and correlation information of different visual cues extracted from videos could be well exploited in model learning (Section 3).
- We evaluate our approach on the benchmark dataset and empirically show improvement over several state-of-the-art alternatives (Section 4).

The rest of this paper is organized as follows: Section 2 provides a brief review of related work; Section 3 explains in detail the proposed sequential fusion method; Section 4 presents experimental settings, results and discussions; finally, Section 5 concludes the paper.

2. Related work

This section briefly reviewed two related topics: 1) visual-based depression recognition and 2) multimodal learning with deep architectures.

Automated Depression Recognition. Depression analysis based on various behavioral signals has drawn increasing attention in the affective computing community. Such feasible signals include the vision- and speech-based cues of human communication. While several works in the literature focus on this research topic, we are interested in the visual-based approaches for depression recognition. In the AVEC 2013 competition[19], a facial descriptor named the local phase quantization (LPQ) [20] was used as a baseline for facial depression recognition, where the extracted LPQ features for each video frame are further employed to train a support vector regression (SVR). In [7], Cummins et al. used the pyramid of histogram of gradients (PHOG) [21] and the space-time interest points (STIPs) [22] for extraction of behavioural cues for depression analysis. Meng et al. [9] proposed to use motion history histogram (MHH) [23] feature to model motion in videos, and then use the partial least squares (PLS) [24] for training regression model. In [13], the motion cue is encoded by the LPQ-TOP feature extracted from sub-volumes of the cropped facial regions, by which the behavior pattern dictionary can be obtained based on sparse coding.

In the AVEC 2014 competition [5], the local motion descriptor LGBP-TOP [25] and the SVR were employed as the baseline video description and prediction model, respectively. In [26], various local motion features extracted from sub-volumes of the detected faces were used for training an SVR-based prediction model. Jan et al. [27] proposed a 1D MHH based on some local descriptors to train a PLS regressor for final prediction. In [28], the baseline LGBP-TOP combined with LPQ was used as the video descriptor for depression prediction.

The aforementioned depression recognition methods proposed to use hand-crafted descriptors, which are generally less effective to model and reveal high-level semantic cues. Recently, depression feature learning based on deep CNNs achieves considerable progress. For example, Zhu et al. [15] proposed jointly learning the facial appearance and dynamics based on a two-stream CNN, in which two different features are fused at a fully-connected layer. Improved performance reported in their experiments indicated the efficacy of such a simple fusion manner. Most recently, Uddin et al. [29] introduced a new two-stream network for deep spatiotemporal feature learning, in which spatial information is extracted by a ResNet network, and they used a volume local directional number (VLDN) based feature descriptor to model facial motions. Zhou et al. [30] proposed a deep network named DepressNet to learn facial depression features with visual explanation, such that facial salient regions with different depression levels can be detected by using the generated activation mapping. Later, Zhou et al. [31] proposed to jointly learn the feature embedding and label distribution to address the issue of deep representation learning on a limited amount of labeled depression data, and the improvement by such learning scheme was reported in their experiments in comparison several state-of-the-art alternatives. Besides deep CNNs, there are several methods for image feature learning such as binary descriptor [32].

Multimodal Deep Representation Learning. In the multimodal setting, visual data consists of multiple input modalities [33–37], and each one may have a different representation and structure. Intuitively, useful representations could be learned from such multimodal data by fusing them into a joint representation to characterize the real-world semantic concept that the visual data corresponds to. In practice, however, it is much more difficult to model and discover the nonlinear correlation and diversity across modalities than those among features in the same modality. Recently, a good number of multimodal deep learning methods have been proposed to better exploit useful information from different modalities for more robust visual analysis [15,29,36,38–46]. In [47], Srivastava et al. proposed a multimodal learning method with a deep Boltzmann machine (DBM) to jointly learn multimodal feature representations. They approached this by adding a concatenated layer that connects DBMs from different modalities. In [42,43], multi-metric learning methods were applied to exploited multi features of human faces. Yan et al. [42] proposed a jointly learning method to capture multiple features from multiple distance metrics. In [43], Hu et al. further studied multi-view metric learning in a sharable and individual way, which shown the superiority of sharing features from different view. In [39], a two-stream CNN with an additional multimodal fusion layer was proposed for RGB-D object recognition. Motivated by the observation that the data from different modalities may contain modal-specific patterns as well as common patterns, Wang et al. [36] proposed a shareable and specific multimodal feature learning framework for RGB-D object recognition. Li et al. [44] proposed a global-local framework to extract pose, appearance and motion features for RGB-D gesture recognition. By imposing the representation learning of associations between different modalities, Zolfaghari et al. [41] designed a chained multi-stream network to fully exploit the pose, motion, and appearance cues for action classification and detection. Feichtenhofer et al. [45] proposed a two-stream CNN architecture to fuse a spatial and temporal network at a convolutional layer instead of at the softmax layer, which boosted performance on action recognition problem with a substantial saving in parameters. To further boost performance of action classification and detection, Feichten-

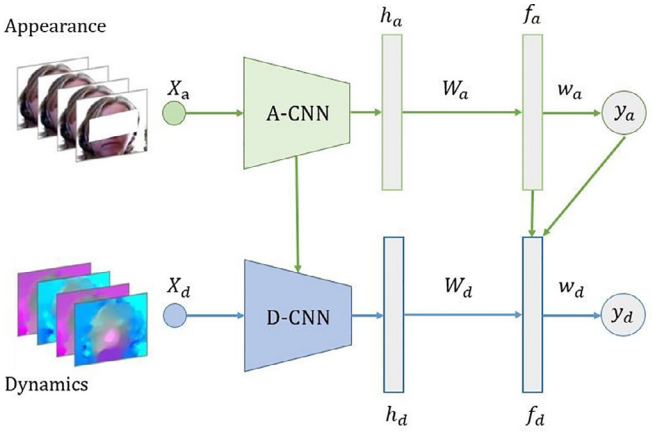

Fig. 1. Overview of our proposed sequential fusion approach for facial depression recognition. The (mid- and high-level) features extracted from the first stream (RGB modality) together with the predicted label are fed into the second stream (dynamics modality) for refinement of the final prediction.

hofer et al. [46] presented SlowFast Networks for video recognition which was inspired by biological studies on the retinal ganglion cells in the primate visual system.

Our proposed method is related to the multimodal deep learning methods for visual recognition, which aimed to fuse multiple responses from the CNNs trained with different modalities of visual data. However, our work differs from them in the fusion manner. Our work aims at deep fusion between features from different modalities in a sequential manner, while only a simple fusion at the fully-connected layer was performed in [29] as well as [15]. Most similar to ours is the work [41] which also performed fusion in a chained manner; however, the fusion of mid-level features in the CNN training was not involved in their approach, and hence our sequential fusion represents a more *deep* fusion mechanism that can be more suitable for multimodal representation learning with limited labeled depression data, as indicated in our experiments.

3. Our approach

Predicting the severity of facial depression is a process of learning spatio-temporal features related to human emotion categorization [48] from face videos. The facial appearance of a subject is one of the important visual patterns for depression recognition. At the same time, facial dynamics characterized by optical flow captures the temporal variations of appearance across frames. As such, we propose a sequential fusion approach to investigate the correlation and complementarity between two different data modalities for depression analysis. As shown in Fig. 1, we use a two-stream network architecture, where the encoders for each stream have the same backbone structure (can be any off-the-shelf CNNs).

Fundamentally, depression estimation can be viewed as a regression task, and hence we employ the mean square error (MSE) as the loss function. Mathematically, the loss L is defined as:

$$L = \frac{1}{2N}\sum_{i=1}^{N} \|y_i - \hat{y}_i\|^2, \tag{1}$$

where N is the number of the samples, $\hat{y}_i$ is the output prediction of the second stream of our network for the ith sample, and y_i is its ground-truth label.

3.1. Appearance-CNN

CNNs have been proved with powerful capability on image classification tasks over large-scale image data. Conversely, CNN is not a proper option to capture features from the dataset with a limited size. For depression estimation, available datasets are usually with limited data and subjects. To handle this issue, we employ the pre-training and fine-tuning strategies to train the facial appearance CNN (A-CNN).

Due to time restrictions or computational restraints, it's not always possible to build a deep model from scratch which is the reason why we use pre-trained model. To achieve facial representations, we train two pre-trained deep networks (e.g., GoogLeNet [49] and ResNet-50 [50]) over CASIA-WebFace database [51], which is a public face recognition database containing 494,414 facial images from 10,575 subjects.

After the pre-training step, we can obtain the general facial features through the pre-trained model while those features are not relevant to facial depression. Hence, depression data are fed into the pre-trained model for fine-tuning, such that the final model is capable of accurate depression estimation.

3.2. Dynamics-CNN

Along with facial appearance, facial dynamics is also an indispensable component in our proposed model. The dynamics-CNN (D-CNN) is built upon the same backbone as the A-CNN with the optical flow data. Unlike the static RGB data, facial dynamics model the motion patterns inherent in faces that can be highly indicative for visual depression analysis.

We compute the optical flow with the duality-based approach [52], which is a decent method with sufficiently fast speed. To feed the optical flow data into the D-CNN, we transform the optical flow signal into a 3-dimensional image data (x, y, m), where x, y and m represent the x-component, y-component and the magnitude of the flow, respectively. For better observation and calculation, we multiply each image by 16 and convert it to the closest integer between 0 and 255 [53].

Similarly, we employ the same architecture, configurations, and loss used in the appearance CNN to train the dynamics CNN. The difference lies in that the input to the dynamics stream is the *flow* image computed from video frames, and there is no pre-training step in D-CNN.

3.3. Sequential fusion

Since Simonyan et al. [54] proposed a simple fusion framework with two separate CNNs on raw images and optical flow, respectively, multi-stream architectures entered the public consciousness and became a popular approach to handle multimedia tasks.

The baseline fusions are illustrated in Fig. 2, where the solution (a) applies the normal concatenation to form the input features of a set of fully-connected layers, which combines features from two streams directly, and the two CNNs are independent. In such a case, the learning manners of the two streams exert a slight influence on each other. Furthermore, as the input to the dynamics, CNN is the optical flow between several sequential frames; the changes between continuous frames are minor and difficult to distinguish. In order to achieve a compact and discriminative multimodal deep representation, a proper fusion method is needed to gather features from facial appearance and dynamics.

To integrate the two individual deep networks mentioned above, we introduce a sequential fusion approach based on a two-stream architecture. By making use of the features from both modalities, a Markov chain is established to integrate the two streams, which may refine the depression prediction sequentially. Considering different modality as the main input, a refined prediction is achieved by combining the hidden features (e.g., high-level features) and the predictions from the previous stream (see Fig. 2(b) and (c)). However, such fusion is made only after the

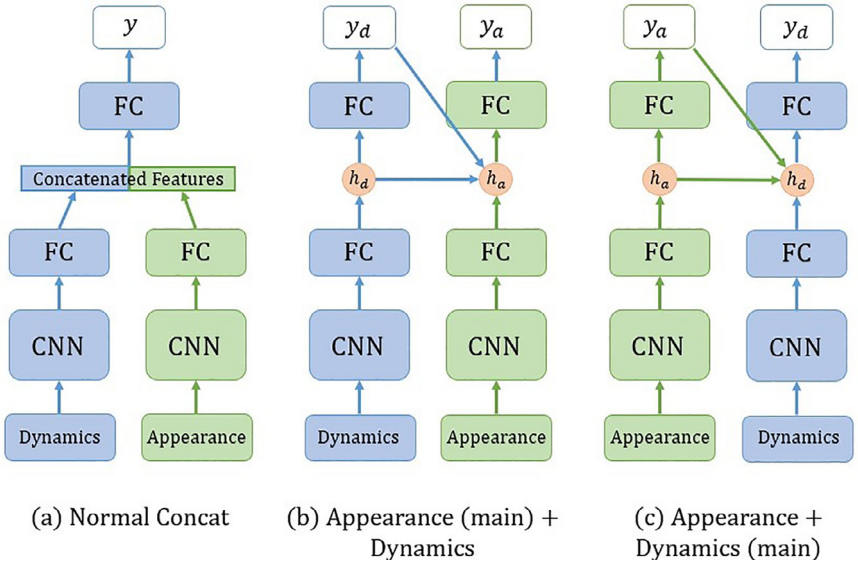

Fig. 2. Different fusion baselines for facial depression recognition. (a) Normally concatenated features from the dynamics and appearance streams; (b) The features extracted from the dynamics stream together with the predicted label are fed to the appearance stream for final prediction; (c) The features extracted from the appearance stream together with the predicted label are fed to the dynamics stream for final prediction.

fully-connected layers, which means the *mid-level features* before the fully-connected layers have not been exploited for better fusion. With the limited promotion of the feature fusion, the refinement of the predictions is improved, though the improvement is not significant.

Unlike the baseline fusion schemes shown in Fig. 2, our proposed sequential fusion (shown in Fig. 1) perform feature fusion not only after the fully-connected layer, but also on the blocks of CNNs, by which both high-level and mid-level features of the CNNs can be well exploited to model the correlation and complementarity between different data modalities. In what follows, we give a probabilistic interpretation for such a fusion mechanism.

For different input modalities, we assume that the depression predictions are conditionally independent. Consequently, we can factorize the joint probability into the conditional probabilities according to the conditional independence property. In a Markov chain, we predict the outputs $Y = \{y_1, y_2, \ldots, y_S\}$ on the given sequence of inputs $I = \{I_1, I_2, \ldots, I_S\}$ with $P(y|I)$ maximized. Due to the Markov property, we have

$$P(y|I) = P(y_1|I)\prod_{s=2}^{S} P(y_s|I, y_1, \ldots, y_{s-1}) \tag{2}$$

To model the likelihood in (2), the hidden feature and the prediction probability are respectively denoted by

$$\begin{aligned} h_s &= f([h_{s-1}, CNN(I_s), (y_1, y_2, \ldots, y_{s-1})]), \\ P(y_s|I, y_{<s}) &= \mathcal{N}(w_s^T h_s; \bar{y}_s, \sigma^2) \end{aligned} \tag{3}$$

where $s \in \{1, 2, \ldots, S\}$, f is an activation function, h_{s-1} denotes the hidden feature from the previous stream, y_s denotes the prediction of the sth stream, w_s denotes the regression coefficient vector, $\bar{y}_s$ is the ground-truth label, and σ is a certain standard deviation of the Gaussian $\mathcal{N}$. Here, $CNN(\cdot)$ denotes the convolutionary part and the first fully-connected layer of the network, which can be any off-the-shelf backbones (e.g., VGG and ResNet).

In the proposed approach, the prediction of the dynamics stream is made conditioned on the appearance stream as well as the input dynamics data, which means that the final prediction is effected not only by the input of the current stream but also by the features and predictions from previous streams (see Fig. 1). When the input modalities is $I = \{I_a, I_d\}$, we have:

$$\begin{aligned} h_a &= CNN(I_a), \\ P(y_a|I) &= \mathcal{N}(w_a^T h_a; \bar{y}_a, \sigma^2) \end{aligned} \tag{4}$$

and

$$\begin{aligned} h_d &= f([h_a, CNN(I_d), y_a]), \\ P(y_d|I, y_a) &= \mathcal{N}(w_d^T h_d; \bar{y}_d, \sigma^2) \end{aligned} \tag{5}$$

As known to all, maximization of $P(y_d|I, y_a)$ is equivalent to the minimization of the MSE loss defined in Eq. (1). Hence, our sequential fusion mechanism has a clear probabilistic interpretation.

It should be noticed that depressed patients are usually slow to initiate actions with stiff facial expressions. For depression analysis, motion features could be more discriminative than appearance cues. It is observed in our experiment that using the dynamics stream as the main stream to fuse the appearance stream can perform better prediction, and it also facilitates the training of the two-stream network. In the inference stage, only the prediction of the mainstream is used as the final depression prediction.

4. Experiments

To validate the effectiveness of our depression recognition approach, we conduct experiments on the AVEC 2014 benchmark dataset and compare its performance with several state-of-the-art algorithms as well as the baselines. In what follows, a description of the dataset, data pre-processing, and experimental setting are first presented. Then, we present the results and analysis.

4.1. Dataset

We conduct the experiments on a database of the Audio/Visual Emotion Challenge (AVEC) 2014 [5], which is the most widely-used depression sub-challenge database for depression recognition. In AVEC dataset, to evaluate the severity, a depression level score is measured by a self-reported 21 multiple-choice inventory–Beck Depression Index (BDI) [55]. The BDI scores range from 0 to 63, where the lower score represents more mild symptoms. The score evaluated in [0, 13], [14, 19], [20, 28] and [29, 63] indicate minimal, mild, moderate and severe depression, respectively. For each video clip, there are 3–5 annotators predicting the BDI score.

In AVEC 2014, there are 84 subjects, and each subject needs to perform two different tasks named "Northwind" and "FreeForm" according to the instructions. All subjects in the two tasks speak German. There are 150 videos for each task, and the recordings were equally split into three partitions: training, development, and test set. Each partition includes 50 videos and has similar distributions in terms of gender, age, and depression levels for the partitions. All videos are recorded by webcam in a human-computer interaction scenario, and each video is approximately 2-min length on average. There are at least three annotators per clip, and most clips are annotated by 5.

4.2. Data pre-processing

As the raw data have a certain degree of noisy and redundant information which is irrelevant to depressive expressions. To extract meaningful information from noise, it is necessary to apply multiple pre-processing steps on the raw data before feeding it to the model. To avoid the waste of computing resources and speed up the training, subsampling is performed on the video frames with an interval of 10 frames which is determined experimentally.

To deal with the raw data, face detection and landmark localization of each subject in the video are implemented by Dlib [56]. After that, the facial region of an image size of 256×256 is cropped and aligned according to the eye locations.

After the above steps, we compute optical flow over a sequence of facial regions extracted from each video clip. As the input to dynamics stream, optical flow is computed between two frames which can capture facial motions known as face "flow images". A "flow image" has three components (x, y, m), where the first two channels are x, and y flow values and the third channel is the magnitude of the optical flow normalized between 0 and 255 with a median of 128.

4.3. Experimental setting

We use two popular network architectures, i.e., VGG-11 [57] and ResNet-50 [50], as the backbone of our two-stream network to train the appearance and dynamic CNNs. In our experiments, the appearance CNN is pre-trained on a large-scale face dataset CASIA WebFace [51], while the dynamics CNN is trained from scratch. The MSE is adopted as the loss function for our depression regression.

The total number of training iterations for the appearance and dynamics CNNs are 400,000 and 600,000, respectively. We set the initial learning rate to 0.001, and decrease the learning rate by polynomial decay with power equals to 0.5, and set the momentum to 0.9 with a weight decay of 0.0002. For the joint training, the total number of iterations is 200,000 with an initial learning rate of 0.0001.

We use the mean absolute error (MAE) and root mean square error (RMSE) to measure the overall recognition performance. They are defined by $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$ and $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$, where N is the number of data samples, y_i and $\hat{y}_i$ are the ground truth and the prediction for the ith sample.

4.4. Experimental results

4.4.1. Performance of individual models

We first investigate the impact of different fusion models (see Fig. 2) in our fusion framework for depression analysis. Six baselines are defined for evaluation of the performance of individual models: 1) Appearance CNN, 2) Dynamics CNN, 3) Normal Concat: fusion by normal concatenation on the fully-connected layer of the two-stream CNN, 4) Appearance (main)+Dynamics: Dynamics stream is fed to Appearance stream for sequential fusion, 5) Dynamics (main)+Appearance: Appearance stream is fed to Dynamics stream for sequential fusion, 6) DeepFusion: Appearance stream is fed to Dynamics stream for a sequential fusion of both mid-level and high-level features.

Table 1
Ablation study of our approach on AVEC 2014.

Model	MAE	RMSE
Appearance CNN	6.60	8.88
Dynamics CNN	8.64	10.71
Normal Concat	6.72	8.68
Appearance(main)+Dynamics	6.71	8.58
Appearance+Dynamics(main)	6.41	8.70
DeepFusion (proposed)	**6.16**	**8.13**

Table 2
Depression prediction results with different backbones on AVEC 2014.

Model	MSE	RMSE
Appearance CNN(VGG)	8.19	10.34
Dynamics CNN(VGG)	9.54	11.49
Normal Concat(VGG)	9.54	11.50
DeepFusion(VGG)	**7.54**	**9.79**
Appearance CNN(ResNet-50)	6.60	8.88
Dynamics CNN(ResNet-50)	8.64	10.71
Normal Concat(ResNet-50)	6.72	8.68
DeepFusion(ResNet-50)	**6.16**	**8.13**

Table 3
Comparison with previous methods on AVEC 2014.

Methods	MAE	RMSE
LGBP-TOP+SVR [5] (2014)	8.86	10.86
MRLBP-TOP+DPFV+SVR [59] (2018)	7.21	9.01
SlowFast Networks [46] (2019)	6.78	8.40
C3D(Global+Local) [60] (2019)	6.59	8.31
VLDN+Bi-LSTM+TMP [29] (2020)	6.86	8.78
DepressNet [30] (2020)	6.60	8.88
DJ-LDML [31] (2020)	6.59	8.30
Spectral-Representation [61] (2020)	**5.95**	**7.15**
DLGA-CNN [62] (2021)	6.51	8.30
DeepFusion (proposed)	6.16	8.13

As shown in Table 1, our proposed DeepFusion achieves the MAE/RMSE of 6.16/8.13, which consistently outperforms the other baselines. Specifically, two sequential fusion models (dynamics CNN as the mainstream) perform better than other individuals, indicating the efficacy of such fusion mechanism. On the other hand, we can see that appearance CNN as the mainstream for sequential fusion may not be suitable for depression analysis, as the motion cues may play a more important role in features fusion for final prediction. Finally, the proposed DeepFusion performs better than Dynamics (main)+Appearance, indicating that integration of mid-level features in the sequential fusion framework can further boost the overall prediction performance.

In addition, we investigate the impact of different backbones in our proposed approach. As shown in Fig. 2, prediction models built upon ResNet-50 [50] consistently perform better than those built upon VGG-11 [57]. Also, our proposed DeepFusion achieves the best performance on both backbones in terms of MAE and RMSE.

4.4.2. Comparison with previous methods

We compared our proposed approach with several state-of-the-art depression recognition methods, and the results are presented in Table 3. For a fair comparison, the compared eight methods are based on the visual modality (e.g., face videos). Specifically, six of

them are also deep learning-based solutions, and the other two are shallow learning models with hand-crafted video descriptors.

In [5], the baseline model for AVEC 2014 employed the epsilon-SVR with intersection kernel [58] trained using LGBP-TOP features. On the AVEC 2014 dataset, our approach beats the baseline approach by dropping the MAE by 2.70 and RMSE by 2.83. In [59], an SVR trained with the dynamic feature descriptors MRLBP-TOP and DPFV achieved the MAE/RMSE of 7.21/9.01. Also, our approach outperforms this shallow learning-based solution by a significant margin.

Our approach achieves the second-best performance on the AVEC 2014 dataset when compared to other seven deep learning-based works [29–31,46,60–62]. In [46], the SlowFast networks transferred from action recognition model achieved the MAE/RMSE of 6.78/8.40. In [60], the combination of the global and local Convolutional 3D networks achieved the MAE/RMSE of 6.59/8.31. In [29], the MAE/RMSE of 6.86/8.78 was achieved by utilizing Bi-LSTM, whose input is the output of a deep CNN and TMP. In a very recent work [62], both local and global attention CNN are introduced for depression recognition and reduced the MAE/RMSE to 6.51/8.30. In [30], deep depression representation with visual explanation achieved the MAE/RMSE of 6.60/8.88, and later in [31], the deep metric learning-based solution achieved the MAE/RMSE of 6.59/8.30. Our approach consistently outperforms the aforementioned deep learning-based methods in terms of MAE and RMSE. The best-performed solution among the compared methods is the spectral representation of behavior primitives [61], which achieved the MAE/RMSE of 5.95/7.15.

5. Conclusion

In this paper, we proposed a deep multimodal learning method for the representation fusion of facial appearance and dynamics. To model the correlated and complementary depression patterns in multimodal learning, a chained-fusion mechanism is introduced to jointly learn facial appearance and dynamics in a unified framework. We showed that such sequential fusion provides a clear probabilistic perspective of the model correlation and complementarity between two different data modalities for improved depression recognition. Experimental results on a benchmark dataset demonstrated the efficacy of our method when compared to several state-of-the-art alternatives. In future work, investigation of the private-share model for multimodal depression representation learning appears to be an interesting topic.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W.H. Organization, et al., Depression and Other Common Mental Disorders: Global Health Estimates, Technical Report, World Health Organization, 2017.
[2] S. Ross, N. Heath, A study of the frequency of self-mutilation in a community sample of adolescents, J. Youth Adolesc. 31 (1) (2002) 67–77.
[3] J.-P. Lépine, M. Briley, The increasing burden of depression, Neuropsychiatr. Dis. Treat. 7 (Suppl 1) (2011) 3.
[4] S. Ji, X. Li, Z. Huang, E. Cambria, Suicidal ideation and mental disorder detection with attentive relation networks, Neural Comput. Appl. (2021).
[5] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, Avec 2014: 3d dimensional affect and depression recognition challenge, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, 2014, pp. 3–10.
[6] J.R. Williamson, T.F. Quatieri, B.S. Helfer, R. Horwitz, B. Yu, D.D. Mehta, Vocal biomarkers of depression based on motor incoordination, in: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013, pp. 41–48.
[7] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, J. Epps, Diagnosis of depression by behavioural signals: a multimodal approach, in: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 2013, pp. 11–20.
[8] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, F. De la Torre, Detecting depression from facial actions and vocal prosody, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009, pp. 1–7.
[9] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, Y. Wang, Depression recognition based on dynamic facial and vocal expression features using partial least square regression, in: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 2013, pp. 21–30.
[10] Y. Yang, C. Fairbairn, J.F. Cohn, Detecting depression severity from vocal prosody, IEEE Trans. Affect. Comput. 4 (2) (2012) 142–150.
[11] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, Depaudionet: an efficient deep model for audio based depression classification, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 35–42.
[12] J.M. Girard, J.F. Cohn, M.H. Mahoor, S.M. Mavadati, Z. Hammal, D.P. Rosenwald, Nonverbal social withdrawal in depression: evidence from manual and automatic analyses, Image Vis. Comput. 32 (10) (2014) 641–647.
[13] L. Wen, X. Li, G. Guo, Y. Zhu, Automated depression diagnosis based on facial dynamic analysis and sparse coding, IEEE Trans. Inf. Forensics Secur. 10 (7) (2015) 1432–1441.
[14] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, M. Breakspear, Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors, IEEE Trans. Affect. Comput. 9 (4) (2016) 478–490.
[15] Y. Zhu, Y. Shang, Z. Shao, G. Guo, Automated depression diagnosis based on deep networks to encode facial appearance and dynamics, IEEE Trans. Affect. Comput. 9 (4) (2017) 578–584.
[16] A.T. Beck, R.A. Steer, R. Ball, W.F. Ranieri, Comparison of beck depression inventories-ia and-ii in psychiatric outpatients, J. Pers. Assess. 67 (3) (1996) 588–597.
[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[18] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1097–1105.
[19] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, Avec 2013: the continuous audio/visual emotion and depression recognition challenge, in: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 2013, pp. 3–10.
[20] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[21] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, 2007, pp. 401–408.
[22] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[23] H. Meng, N. Pears, Descriptive temporal template features for visual motion recognition, Pattern Recognit. Lett. 30 (12) (2009) 1049–1058.
[24] S. De Jong, Simpls: an alternative approach to partial least squares regression, Chemom. Intell. Lab. Syst. 18 (3) (1993) 251–263.
[25] T.R. Almaev, M.F. Valstar, Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, 2013, pp. 356–361.
[26] H. Pérez Espinosa, H.J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, V. Reyez-Meza, Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, 2014, pp. 49–55.
[27] A. Jan, H. Meng, Y.F.A. Gaus, F. Zhang, S. Turabzadeh, Automatic depression scale prediction using facial expression dynamics and regression, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, 2014, pp. 73–80.
[28] H. Kaya, F. Çilli, A.A. Salah, Ensemble CCA for continuous emotion prediction, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, 2014, pp. 19–26.
[29] M.A. Uddin, J.B. Joolee, Y.-K. Lee, Depression level prediction using deep spatiotemporal features and multilayer bi-LTSM, IEEE Trans. Affect. Comput. (2020).
[30] X. Zhou, K. Jin, Y. Shang, G. Guo, Visually interpretable representation learning for depression recognition from facial images, IEEE Trans. Affect. Comput. 11 (3) (2020) 542–552.
[31] X. Zhou, Z. Wei, M. Xu, S. Qu, G. Guo, Facial depression recognition by deep joint label distribution and metric learning, IEEE Trans. Affect. Comput. (2020), doi:10.1109/TAFFC.2020.3022732. 1–1
[32] Y. Duan, J. Lu, Z. Wang, J. Feng, J. Zhou, Learning deep binary descriptor with multi-quantization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1183–1192.
[33] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (264–270) (2019).

[34] P. Chhokra, A. Chowdhury, G. Goswami, M. Vatsa, R. Singh, Unconstrained kinect video face database, Inf. Fusion 44 (2018) 113–125.
[35] L. Stappen, A. Baird, E. Cambria, B. Schuller, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2) (2021) 88–95.
[36] A. Wang, J. Cai, J. Lu, T.-J. Cham, Mmss: multi-modal sharable and specific feature learning for RGB-D object recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1125–1133.
[37] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics, in: IEEE SSCI, Singapore, 2013, pp. 108–117.
[38] A. Wang, J. Lu, J. Cai, T.-J. Cham, G. Wang, Large-margin multi-modal deep learning for RGB-Dobject recognition, IEEE Trans. Multimed. 17 (11) (2015) 1887–1898.
[39] A. Eitel, J.T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard, Multimodal deep learning for robust RGB-D object recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2015, pp. 681–687.
[40] I. Lenz, H. Lee, A. Saxena, Deep learning for detecting robotic grasps, Int. J. Robot. Res. 34 (4–5) (2015) 705–724.
[41] M. Zolfaghari, G.L. Oliveira, N. Sedaghat, T. Brox, Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2904–2913.
[42] H. Yan, J. Lu, W. Deng, X. Zhou, Discriminative multimetric learning for kinship verification, IEEE Trans. Inf. Forensics Secur. 9 (7) (2014) 1169–1178.
[43] J. Hu, J. Lu, Y.-P. Tan, Sharable and individual multi-view metric learning, IEEE Trans. Pattern Anal. Mach. Intell. 40 (9) (2017) 2281–2288.
[44] B. Li, W. Li, Y. Tang, J.-F. Hu, W.-S. Zheng, Gl-pam RGB-D gesture recognition, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3109–3113.
[45] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.
[46] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6202–6211.
[47] N. Srivastava, R. Salakhutdinov, et al., Multimodal learning with deep boltzmann machines., in: NIPS, 1, Citeseer, 2012, p. 2.
[48] Z. Wang, S. Ho, E. Cambria, A review of emotion sensing: categorization models and algorithms, Multimed. Tools Appl. 79 (2020) 35553–35582.
[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[51] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, arXiv:1411.7923(2014).
[52] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime tv-l 1 optical flow, in: Joint Pattern Recognition Symposium, Springer, 2007, pp. 214–223.
[53] P. Weinzaepfel, Z. Harchaoui, C. Schmid, Learning to track for spatio-temporal action localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3164–3172.
[54] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. 27 (2014) 568–576.
[55] J.T. Olin, L.S. Schneider, E.M. Eaton, M.F. Zemansky, V.E. Pollock, The geriatric depression scale and the beck depression inventory as screening instruments in an older adult outpatient population., Psychol. Assess. 4 (2) (1992) 190.
[56] D.E. King, Dlib-ml: a machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[57] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556(2014).
[58] S. Maji, A.C. Berg, J. Malik, Classification using intersection kernel support vector machines is efficient, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[59] L. He, D. Jiang, H. Sahli, Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding, IEEE Trans. Multimed. 21 (6) (2018) 1476–1486.
[60] W.C. de Melo, E. Granger, A. Hadid, Combining global and local convolutional 3d networks for detecting depression from facial expressions, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–8.
[61] S. Song, S. Jaiswal, L. Shen, M. Valstar, Spectral representation of behaviour primitives for depression analysis, IEEE Trans. Affect. Comput. (2020).
[62] L. He, J.C.-W. Chan, Z. Wang, Automatic depression recognition using CNN with attention mechanism from videos, Neurocomputing 422 (2021) 165–175.