Multi Task Sequence Learning for Depression Scale Prediction from Video

Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li
National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences
Beijing, China
{linlin.chao, jhtao, mhyang, yli}@nlpr.ia.ac.cn
Jianhua Tao
Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China

Abstract—Depression is a typical mood disorder, which affects people in mental and even physical problems. People who suffer depression always behave abnormal in visual behavior and the voice. In this paper, an audio visual based multimodal depression scale prediction system is proposed. Firstly, features are extracted from video and audio are fused in feature level to represent the audio visual behavior. Secondly, long short memory recurrent neural network (LSTM-RNN) is utilized to encode the dynamic temporal information of the abnormal audio visual behavior. Thirdly, emotion information is utilized by multi-task learning to boost the performance further. The proposed approach is evaluated on the Audio-Visual Emotion Challenge (AVEC2014) dataset. Experiments results show the dimensional emotion recognition helps to depression scale prediction.

Keywords—depression recognition; affective computing; multi-task learning; deep learning

I. INTRODUCTION

Depression affects the psychological state of a wide range of the human population and can be life-threatening for men, women and even children. According to EU Green Papers from 2005 [1] to 2008 [2], mental health problems affect one in four citizens at some point during their lives and too often lead to suicide. It is therefore very necessary to predict and diagnose the depression for patients.

Depression is a state of low mood and aversion to activity that can affect a person’s thoughts and behavior. In this low mood, people’s voice, facial expression and body movements often behave abnormal, which can be indicators to depression scale. Therefore, automatic voice and visual analysis can be regarded as a good start to predict depression scale. In the work presented here, we extract both audio features and visual features and fuses these features in feature-level to predict the depression scale.

Facial expressions, eye gazed and head motion are important visual features used by psychologists to gauge depression in patients. Advances in the field of computer vision allow us to automatically observe these visual features. However, most research has focused on static images. In [3], the authors underline the importance of spatial-temporal information for affect sensing. A growing body of researches [4, 5] in cognitive sciences also argue that the dynamics of facial and vocal behavior are crucial for their interpretation. Moreover, a number of recent studies [6]-[10] in affective computing show the importance of temporal modeling for affect understanding. Besides visual information, audio signals can also help psychologists to gauge depression. Most of the audio analysis are focused in two aspects, frame level or utterance level. However, the frame level signal only contains part of the information conveyed by people. The utterance level features are always the statistics of frame level features. So the temporal information conveyed among several frames or utterance is not fully explored. In this context, we use the Long Short Memory Recurrent Neural Network to encode the temporal information for both audio and visual signals. LSTM-RNN can model both long and short range context, which is suited to incorporate the whole video information without losing too much temporal information.

Besides, depressed people may feel sad, anxious, hopeless, empty, worried, hurt, irritable or restless, which can be expressed by their emotions. Effectively discovering and exploiting such emotion information would help in predicting the depression scale. In the work presented here, we also try to utilize the dimensional emotion information to predict the depression scale. As dimensional emotion models the emotion from time to time, which provide rich information to understanding the human behavior. We utilize the multi-task learning framework to learn the dimensional emotion and depression together. With the help of dimension emotion information, we hope the performance of depression prediction can improve further.

Section 2 provides a brief description of the related work. The related work mainly focuses on audio visual based depression prediction. Section 3 introduce the dataset we utilize, which has dimensional emotion annotation and depression scale annotation. Section 4 describes our proposed approach. In this part, the extracted features, the LSTM-RNN encoder and multi task learning framework are introduced in details. The feature part include features from audio and visual modalities. For visual features, we also include the face shape feature and appearance feature. The landmarks of the face is utilized to encode the shape information of face. For the appearance feature, features
extracted from convolutional neural network (CNN) are utilized. The experiment and experiment results are presented in section 5.

II.RELATED WORK

A. Audio visual based depression prediction

There are growing number of studies for depression prediction form audio and visual signals [32]-[35]. Depression recognition sub-challenge of AVEC has focused on this problem in the 2013 and 2014 [31][16]. Williamson et al. [34][35], the winner of AVEC2013 and AVEC2014, looks at the change in motor control, which can reflect the mechanisms controlling speech production and facial expression. They derive multi-scale correlation structure and timing feature sets from audio based vocal features in AVEC2013. In AVEC2014, they also add the features about multi-scale correlation structure and timing feature set from video-based facial action units. Their multi-scale correlation features explore the temporal characteristic of depression patients by multi scale, long and short time delay analysis of the sequential signals. Meng et al. [11] presents a novel method, which comprehensively models visual and vocal modalities with dynamic features to automatically predicts the scale of depression. Motion History Histogram (MHH) is used to extract the dynamics from corresponding video and audio data to represent characteristics of subtle changes in facial and vocal expression of depression. Gupta et al. [12] extract features from audio, visual and text modalities, which creating a pool of 42092 candidate features, then investigate multiple feature selection approaches. After feature selection, the final result is obtained by Support Vector Regression (SVR).

From the above methods, we can see feature is one of the vital factor for depression scale prediction. Combing audio and visual modalities has significant advantages compared to single modality. More importantly, the sequential temporal information of the video, can provide important information, which is not fully explored. In this context, we focused on temporal modeling for depression scale prediction. Besides, features extracted from visual modality, face shape features and audio features are explored.

B. Multi task learning

Also related to the present work are the methods which perform synergistic multi-task learning. One of the most recognized example of multi-task learning from the deep leaning community is the convolutional network based for natural language processing [13] from Osadchy et al. This work tries to learn multi related tasks jointly. While in [14] [15], the face attribute inference and head pose estimation are utilized to help detection for facial landmarks by multi-task learning. In this context, the emotion information is utilized to help the depression scale prediction by multi task learning framework.

III.AUDIO-VISUAL DATASET

The dataset utilized is from the AVEC2014 corpus that includes audio and video recordings of subjects performing a human-computer interaction task [16]. There are 84 German subjects’ recordings are collected. The subjects’ age varied between 18 and 63 years, with a mean of 31.5 years.

In each recoding, subjects performed two different tasks in German language. One is reading a phonetically-balanced passage. The passage is an excerpt of the fable Die Sonne und der Wind (The Northw Wind and the Sun). The other one is replying to a free-response question. In free-response question task, subjects are asked responded to one a number of questions, which are designed to activate the subjects’ emotion states. For example, “What was your best gift, and why?” and “Discuss a sad childhood memory”.

During the human-computer interaction, subjects’ behavior are recorded by a webcam at 30 frames per second and a headset microphone. For each recording, there are two label sets. One is the self-reported Beck depression inventory (BDI) [17] score, which reflects the depression severity. The BDI scores range from 0 to 63 and the higher the score the more severity of the depression. The other label is the dimensional emotion annotation for arousal, valence and dominance dimensions. The emotion labels are annotated from time to time and every frames in the recording are labeled. The value of the emotion annotation for each dimension are scaled from -1 to 1. There are total 150 recordings, with 50 for training, validation and testing respectively.

IV.APPROACH

Fig. 1 depicts the framework of the proposed approach. There are mainly three parts, multimodal feature extraction, multimodal sequence learning for the video, multi task learning for dimensional emotion and depression scale prediction. The details of each part are introduced in the follow sub-sections.

<table>
<thead>
<tr>
<th>Table 1 Audio features</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
</tr>
<tr>
<td>LSF</td>
</tr>
<tr>
<td>ZCR</td>
</tr>
<tr>
<td>OBSI</td>
</tr>
<tr>
<td>OBSI</td>
</tr>
<tr>
<td>MFCC</td>
</tr>
<tr>
<td>Energy</td>
</tr>
<tr>
<td>Envelope</td>
</tr>
<tr>
<td>Loudness</td>
</tr>
<tr>
<td>Mel Spectrum</td>
</tr>
<tr>
<td>Auto Correlation</td>
</tr>
<tr>
<td>Perceptual Spread</td>
</tr>
<tr>
<td>Perceptual Sharpness</td>
</tr>
<tr>
<td>Magnitude Spectrum</td>
</tr>
</tbody>
</table>

A. Multimodal feature extract

1) Audio features

We utilize the YAAFE toolbox [18] to extract audio features. All the waves are resampled to 16 kHz and 27 features are extracted. Features are listed in table 1. There are also four feature transforms used on MFCC features. The feature transforms include the first and second derivate,
cepstrum and auto correlation peaks integrator. All the features are extracted in the default parameters. Finally, 939 dimensions features are extracted for each frame and the frame length is 1024.

2) Face shape feature

As the scene in each recording is relatively clean, without any context prediction cues, we mainly focus on the face part. For extracting faces from the videos, we utilize the code provided by the authors in [19]. This code combines the OpenCV Viola Jones’ face detector [20] with their recently proposed tracking method. This tracking method is able to handle large variation of head poses, which helps to extract faces.

We use the landmarks’ location of the face as face shape feature. In the tracking module, 49 face landmarks are returned. After feature normalization for each recording, these features can also reflect the head movement and head pose.

3) Face appearance feature

For face appearance feature, we utilize the features extracted from a pre-trained CNN model. CNN is a type of feed-forward artificial neural network which is inspired from biology. The individual neurons are designed to simulate cells within visual cortex, which are sensitive to small sub-regions of the input space. We employ the Caffe [21] implementation, which is commonly used in several latest works [22]-[23]. Fer2013[24] dataset is utilized to train the CNN model. Fer2013 dataset is an facial expression dataset, including happy, sad, surprise, neutral, angry and surprise emotion types. It contains of 28,709 48 x48 training images, each with seven emotion labels, as well as a 7177 image test set. Images have been taken “from the wild”. The architecture of the CNN model is shown in Fig.2. The 256 nodes’ values of the ReLU hidden layer are used for the final appearance features.

B. Sequence learning

After extracting audio feature, face appearance feature and face shape feature (PCA dimension reduction and whiten for audio feature and face shape feature separately), these features are fed into the LSTM based neural network. The first layer of the network is a multimodal fusion layer. This layer maps the three different features to one vector. The multimodal fusion process can be expressed via the following equation:

\[ h_{t+1} = m_{t} \cdot b \]
\[ m_t = \tanh (W_{am}a_t + W_{ms}m_t + W_{fm}f_t + b_m) \]  
\[ a_t, s_t, f_t, m_t \text{ represent the audio feature, face shape feature, face appearance feature and multimodal feature at time } t \text{ separately. Then the multimodal feature sequence } m_t \text{ are fed into the LSTM layer.} \]

LSTM-RNN has the ability to learn long-term dynamic while avoiding the vanishing and exploding gradients problems. As research on LSTMs has progressed, hidden units with varying connections within memory unit have been proposed. We use the LSTM unit as described in [25] (Fig.3), which is a slight simplification of the one described in [26]. The LSTM updates for timestep \( t \) given inputs \( m_t, h_{t-1}, \) and \( c_{t-1} \) are:

\[ i_t = \sigma_d(W_{si}m_t + W_{hi}h_{t-1} + b_i) \]  
\[ f_t = \sigma_d(W_{sf}x_t + W_{hf}h_{t-1} + b_f) \]  
\[ o_t = \sigma_d(W_{so}x_t + W_{ho}h_{t-1} + b_o) \]  
\[ g_t = \tanh (W_{sc}m_t + W_{hc}h_{t-1} + b_c) \]  
\[ c_t = f_t * c_{t-1} + i_t * g_t \]  
\[ h_t = o_t * \tanh (c_t) \]  

\( \sigma \) is the output function for LSTM layer.

C. Multi-task learning

The traditional multi-task learning seeks to improve the generation of multiple related tasks by learning them jointly. In contrast to conventional multi-task learning, our aim is to optimize the main task, depression scale prediction, with the assistances of related task, dimensional emotion recognition. However, dimensional emotion recognition needs output for every timestep, which is a sequential outputs problem. While depression scale prediction just need one output, which is a sequential inputs and fixed output problem. Thus, for depression prediction we add a mean pooling layer after LSTM encoding layer, then the pooling results are input into the linear regression layer. The linear regression layer outputs the predicted depression scale. For dimensional emotion recognition, the output of the LSTM layer are directly input into the linear regression layer, which outputs the dimension emotion prediction.

Suppose we have the training data denoted as \((x_i, y^d_i, y^e_i)\), where \( i = \{1, ..., N\} \), with \( y^d, y^e \) represent the depression label and emotion label. As the two tasks are both regression problems, the cost function can be formulated as

\[ c = \sum^N_i l(y^d_i, f(x_i, W^d)) + \sum^N_i \frac{1}{T_i} \sum_{t=1}^{T_i} l(y^e_t, f(x^t_i, W^e)) \]

where \( l \) represents the least square loss function for regression parameterized by a weight vector \( W^d \) and \( W^e \) separately. The network is trained to minimize the cost function.

V. Experiments

A. Experiment setup

We follow the challenge criterion of AVEC2014, with training set for training, development set for validation. All the training processes are monitored by depression scale prediction task. As each recordings are split to two recordings based on the two tasks, we use the split recordings, and the final results are the average of the two tasks for each recording. The best test performance is obtained according to the validation performance. We utilize the code from Theano [27][28]. In the experiment, performance without multi-task learning is also compared.

512 memory cells are utilized in the LSTM layer. Adadelta[29] optimization algorithm with batch size 5 is utilized. The learning rate are set by grid search. The maximum training epoch is 50 with early stopping regularization is applied. Dropout in the linear regression layer are also applied to prevent over fitting.

B. Experiment results

The experiment results are shown in Table 2. From the table, we can see the results from multi-task learning significantly better than the one without multi-task learning. This improves that emotion recognition and depression scale prediction are related tasks and dimensional emotion recognition can helps depression scale prediction. Besides, we also find the proposed approach over fits quickly, especially for multi-task learning. This may because the training data is relatively small.

We also compare our approach with the AVEC2014 depression sub-challenge results. Fig 4\(^1\) shows the challenge result. From Fig 4 and Table 2, we can see our approach achieves competitive results. However, compared to several leading methods, there are still room for improvement.

| Table 2 Performance of the proposed depression scale prediction method measured in root mean square error (RMSE) and mean absolute error (MAE) averaged over all sequences in training, development and testing set. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | Training        | Development     | Testing         |
|                 | MAE  | RMSE | MAE  | RMSE | MAE  | RMSE |
| Without Multi-task learning | 7.31 | 9.67 | 8.94 | 11.16 | 8.70 | 10.61 |
| Multi-task learning          | 4.29 | 5.78 | 8.76 | 11.26 | 7.91 | 9.98 |

\(^1\) http://sspnet.eu/wp-content/uploads/2014/12/DSC_2014.png
VI. CONCLUSIONS

This article presents our approach that models depression scale prediction from video using neural networks. Long short memory recurrent neural network is utilized to explore the effectiveness of temporal modeling for depression scale prediction. Multi modal features are extracted from both audio and visual modality. These features are fused in feature level. Besides, the relationship between dimensional emotion recognition and depression scale prediction is investigated by multi task learning.

Experiment result proves that dimensional emotion recognition helps to the depression scale prediction. This motivate us employ the dimensional emotion recognition methods to depression scale prediction, which will be further explored in the future.

ACKNOWLEDGMENT

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61403386, No.61305003, No.61332017, No.61375027, No.61273288, No.61233009, No.61203258), and the Major Program for the National Social Science Fund of China (13&ZD189).

REFERENCES


