PersEmoN: A Deep Network for Joint Analysis of Apparent Personality, Emotion and Their Relationship

Le Zhang, Songyou Peng, and Stefan Winkler, Fellow, IEEE

Abstract—Personality and emotion are both central to affective computing. Existing works solve them individually. In this paper we investigate if such high-level affect traits and their relationship can be jointly learned from face images in the wild. To this end, we introduce PersEmoN, an end-to-end trainable and deep Siamese-like network which we call emotion network and personality network, respectively. It consists of two convolutional network branches, one for emotion and the other for apparent personality. Both networks share their bottom feature extraction module and are optimized within a multi-task learning framework. Emotion and personality networks are dedicated to their own annotated dataset. An adversarial-like loss function is further employed to promote representation coherence among heterogeneous dataset sources. Based on this, the emotion-to-personality relationship is also well explored. Extensive experiments are provided to demonstrate the effectiveness of PersEmoN.

Index Terms—Affective Computing, Emotion, Personality, Adversarial Learning, Multi-Task Learning, Deep Learning.

1 INTRODUCTION

Proliferation of cameras, availability of cheap storage and rapid developments in high-performance computing have spurred the rise in Human-Computer Interaction (HCI), in which affective computing plays an inevitable role. For instance, in video-based interviews, automatically computed personalities of candidates can serve as an important cue to assess their qualifications. However, affective computing remains a challenging problem in both computer vision and psychology despite many years of research.

We focus on the fundamental problem of analyzing (apparent) personality\(^1\), emotion and their relationship. Personality reflects the coherent patterning of behavior, cognition and desires (goals) over time and space. Emotion is an integration of feeling, action, appraisal, and wants at a particular time and location [1]. We can understand the emotion-to-personality relationship as weather to climate, i.e. what one expects is personality while what one observes in a particular moment is emotion. Although they have distinct definitions, the personality-to-emotion relationship has been revealed previously. Eysenck’s personality model [2] showed that extraverts require more external stimulations than introverts. In other words, extraversion is accompanied by low cortical arousal. He also concluded that neurotics could be more sensitive to external stimulation and easily become upset or nervous due to minor stressors.

In this paper we consider the Big Five personality traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness) [3]. Instead of classifying pre-defined emotion categories, we use a finer-grained representation based on the circumplex model [4], in which emotions are distributed in a two-dimensional circular space spanned by the dimensions of arousal and valence. This is advantageous in the sense that emotional states can be represented at any level of valence and arousal.

There is a plethora of research in the literature on analyzing emotion and personality. [5] studied lexical cues from informal texts for recognizing personalities. [6] showed a high correlation of personality with non-verbal behavioral measures such as the amount of speech and physical activity. [7] investigated the physiological correlation of emotion and personality using commercial sensors and found that the emotion-to-personality relationship is better captured by non-linear rather than linear statistics. [8] proposed a three-layer neural network-based architecture for predicting the sixteen personality factors from faces analyzed using facial action coding system.

Deep convolutional neural networks (CNNs) reign undisputed as the new de-facto method for face based applications such as face recognition [9, 10], alignment [11], and so on. This motivates us to study the following fundamental problems:

1) As both face recognition and affective computing can have faces as input, how transferable are deeply learned face representations for emotion and personality analysis?
2) Is it beneficial to explore emotion, personality and their relationship in a single deep CNN?

These tasks are non-trivial. Among the most significant challenges are:

\(^1\) For simplicity, we will use the term “personality” to represent “apparent personality” in this paper.
• The scarceness of large-scale datasets which encompass both emotion and personality annotations for learning such a rich representation for personality, emotion and emotion-to-personality relationship. In particular, existing datasets only contain emotion attributes, while other datasets may only be annotated with the personality labels. Manually annotating data for both emotion and personality may partly alleviate this. However, it is costly, time-consuming, and error-prone due to subjectivity.

• The discrepancy of existing datasets: datasets are usually collected in different environments which may exhibit significant variations in illumination, scale, pose, etc. Each dataset may have vastly different statistical distributions.

• Emotion is typically annotated at frame level, whereas an entire video is needed for personality labeling. How can we encapsulate both frame and video level understanding into a single network?

We address these challenges by proposing \textit{PersEmoN}, an end-to-end trainable and deep Siamese-like network [12]. It consists of two CNN branches which we call emotion network and personality network, respectively. Emotion network and personality network share their bottom feature extraction module and are optimized within a multi-task learning framework. An adversarial-like loss function is further employed to promote representation coherence between heterogeneous dataset sources. We show that \textit{PersEmoN} works well for analysis personality, emotion and their relationship. Moreover, \textit{PersEmoN} also provides a promising solution for automatically annotating the personality based on the emotion. A demo version of this paper has been presented in [13].

2 Related Work

The wealth of research in this area is such that we cannot give an exhaustive review. Instead, we focus on describing the most important threads of research on using deep learning for face recognition, emotion and personality analysis.

2.1 Deep Learning for Face Recognition

Deep learning was applied to face recognition in the pioneer work of DeepFace [14] and series of DeepID [9, 15, 16]. Inherited from them, most of latest face recognition methods consider the task as a multi-class classification problem and train deep face features on large public datasets such as LFW [17], VGG-Face [10] or FaceNet [18]. While it has been shown that the trained representations are, to some extent, transferable between face recognition and affective computing [3, 19], a direct application of shared CNN representations trained for both emotion and personality without large-scale datasets encompassing both emotion and personality annotations is rarely studied. Inspired by the recent advances in face recognition achieved by light-structured networks [20], we introduce \textit{PersEmoN} with a SphereFace [20] based network backbone to show that such a strategy is advantageous.

2.2 Deep Learning for Emotion Analysis

Emotion analysis has been investigated from different perspectives. [21] proposed a deep belief network for unsupervised audio-visual emotion recognition. However, its feasibility of large-scale supervised learning remains unclear. [22] investigated the usage of deep CNNs and Bayesian classifiers for group emotion recognition in the wild. In addition, [23] introduced the convolutional deep belief network networks to learn salient multi-modal features of emotions. Unlike popular classification approaches for discrete emotion categories, many recent works delve into different representations of human expressions and emotions, such as facial action units [24] or arousal-valence space [4, 25]. This paper focuses on the latter.

2.3 Deep Learning for Personality Analysis

[26] identified personality with a Deep Bimodal Regression framework based on both video and audio input. A similar work from [27] introduced a deep audio-visual residual network for multimodal personality trait recognition. Besides, [28] developed a volumetric convolution and Long-Short-Term-Memory (LSTM) based network to learn audio-visual temporal patterns. However, performances from all above-mentioned methods rely heavily on ensemble strategies and here we report better results with a single visual stream with \textit{PersEmoN}. [29] employed a pre-trained CNN was employed to extract facial expressions as well as ambient information for personality analysis. Although they achieved promising results, the system is not end-to-end trainable and needs a stand-alone regressor. For more related work on personality analysis, please refer to recent surveys [30, 31].

3 Methodology

In comparison to the aforementioned studies, our work aims to investigate whether emotion and personality analysis can benefit from the face representations learned from a well-annotated face recognition dataset, \textit{without having a dataset with both emotion and personality annotations}. To this end, we show that state-of-the-art face recognition networks perform well for both emotion and personality analysis. We also explore the feasibility of jointly training emotion and personality analysis. More specifically, we propose \textit{PersEmoN} within a multi-task learning framework to learn better representations for both emotion and personality than those obtained by solving each task individually. On top of such representations, we demonstrate the feasibility of establishing a good emotion-to-personality relationship.

3.1 \textit{PersEmoN} Overview

An overview of \textit{PersEmoN} can be found in Fig. 1. We first detect and align faces for both personality and emotion datasets with well-established MTCNN [11]. For the personality dataset, we employ a sparse sampling strategy. The personality network consists of a feature extraction module (FEM) and personality analysis module (PAM) to predict the Big Five personality factors. A consensus aggregation function is employed to aggregate raw personality scores.
before feeding them into RAM. Similarly, the emotion network shares the FEM module with the personality network and has its own emotion analysis module (EAM) targeted at predicting the arousal and valence dimensions [19] of emotion. An emotion-to-personality relationship analysis module (RAM) is also employed. In the following section, we elaborate on the different modules mentioned above.

### 3.2 Personality and Emotion Networks

A shared FEM, embodied with a truncated SphereFace network [20] with its last two layers removed, is employed for both branches. Those two branches are dedicated to emotion- and personality-annotated datasets, respectively, and jointly optimized with the FEM.

As personality is defined over a period of time, existing personality datasets only provide video-level annotations. To utilize rich information from each video frames for more effective network training, personality network operates on a pool of sparsely sampled faces from the entire video. Each face in this pool can produce its own preliminary prediction of the personality score. We take inspiration from recent advances in video based human action recognition [32] to employ a consensus strategy among all the faces from each video to give a video-level prediction on the personality. The loss values of video-level predictions, other than those of face-level ones, are optimized by iteratively updating the model parameters. We use \( V \) and \( Y \) to represent a generic video input and its ground truth label. Given the \( i \)th video \( \{ V_{i}^{P}, Y_{i}^{P} \} (i \in \mathbb{N}^{P}) \), where \( \mathbb{N}^{P} \) stands for the index set of personality videos, and \( P \) denotes the data source, i.e., personality dataset here. We divide them into \( K \) segments \( \{ S_{i1}^{P}, S_{i2}^{P}, \ldots, S_{iK}^{P} \} \) of equal duration. Now our personality network models a sequence of faces as follows:

\[
P(V_{i}^{P}, W^{P}) = P(I_{i1}^{P}, I_{i2}^{P}, \ldots, I_{iK}^{P}, W^{P}) = G(F(I_{i1}^{P}, W^{P}), F(I_{i2}^{P}, W^{P}), \ldots, F(I_{iK}^{P}, W^{P}))
\]

(1)

Here \( (I_{i1}^{P}, I_{i2}^{P}, \ldots, I_{iK}^{P}) \) is a pool of faces where each face \( I_{ik}^{P} \) is randomly sampled from its corresponding segment \( S_{ik}^{P} \). The function \( F(I_{ik}^{P}, W^{P}) \) represents the personality network with parameters \( W^{P} \) which operates on face \( I_{ik}^{P} \) and provide preliminary results on the personality scores. The segmental consensus function \( G \) aggregates the raw outputs from multiple faces to obtain a final personality score for each video. Although the proposed method is generic and applicable for a wide range of functions such as max, average, recurrent aggregation, we use the average function similar to [32]. Based on this consensus, we optimize the personality network with the smooth \( \ell_{1} \) loss function [33] defined as:

\[
L_{per}(W^{P}) = \sum_{i \in \mathbb{N}^{P}} \text{smooth}_{\ell_{1}}(Y_{i}^{P} - P(V_{i}^{P}, W^{P}))
\]

(2)

The smooth \( \ell_{1} \) function is given below; \( m \) represents a margin parameter.

\[
\text{smooth}_{\ell_{1}}(x) = \begin{cases} \frac{1}{2}x^2 & |x| < m \\ |x| - 0.5 & \text{otherwise} \end{cases}
\]

(3)

The emotion network works in a simpler manner by directly processing input faces, since frame level annotations are already available. More specifically, given a face image \( \{ I_{i}^{E}, Y_{i}^{E} \} (i \in \mathbb{N}^{E}) \), the emotion network produces emotion scores as:

\[
E(I_{i}^{E}, W^{E}) = F(I_{i}^{E}, W^{E})
\]

(4)

Similarly, the loss function for the emotion network is:

\[
L_{emo}(W^{E}) = \sum_{i \in \mathbb{N}^{E}} \text{smooth}_{\ell_{1}}(Y_{i}^{E} - E(I_{i}^{E}, W^{E}))
\]

(5)

### 3.3 Representation Coherence

Datasets for personality and emotion are usually collected separately. People may appear in various scales and poses under different illumination conditions. Besides, each dataset may exhibit different statistical distributions. Representations learned from each domain individually without pursuing coherence between them may present significant discrepancy. A representation with good transferability should be domain-invariant in the sense that the learned representations are coherent for different data samples from different domains [34]. This is also beneficial to exploring the emotion-to-personality relationship in our case. To this end, a classifier trained using the coherent representation cannot distinguish examples from those two domains.

We take inspiration from [34] by training a domain classifier, denoted as \( D \) with parameters \( W^{D} \), to perform binary classification to distinguish which domain a particular datum comes from. For each feature representation from the FEM, we learn the domain classifier with the following softmax loss:

\[
L_{D}(W^{D}) = - \sum_{i} \left\{ \sum_{k=1}^{K} \frac{\log q(I_{ik}^{P}, W^{P}, W^{D}), \ i \in \mathbb{N}^{P}}{\log q(I_{i}^{E}, W^{E}, W^{D}), \ i \in \mathbb{N}^{E}} \right\}
\]

(6)

where \( q(I, W, W^{D}) = \text{softmax}(W^{D} \cdot F(I, W)) \).

As in [34], an adversarial-like learning objective is introduced in the FEM which aims at “maximally confusing” the two domains by computing the cross entropy between the output predicted domain labels and a uniform distribution over domain labels:

\[
L_{adv}(W^{P}, W^{E}) = \sum_{i \in \mathbb{N}^{P}} \sum_{k=1}^{K} log q(I_{ik}^{P}, W^{P}, W^{D}) \\
+ \sum_{i \in \mathbb{N}^{E}} \sum_{k=1}^{K} log q(I_{i}^{E}, W^{E}, W^{D}) \]

(7)

Similar to the adversarial-learning, we perform iterative updates for both \( L_{D}(W^{D}) \) and \( L_{adv}(W^{P}, W^{E}) \) given the fixed parameters from the previous iteration.

### 3.4 Emotion-to-Personality Relationship Analysis

Here we investigate whether personality can be inferred directly from emotion attributes. This is challenging due to the paucity of datasets which encompass both emotion and personality annotations for us to learn such a relationship. We insert a relationship analysis module (RAM), which receives the emotion scores from the emotion analysis
network and predicts personality scores. More specifically, the input of RAM can be obtained by:

$$E(V_i^P, W^E) = E_i = E(I_{i1}^P, I_{i2}^P, \ldots, I_{iK}^P, W^E) = (F(I_{i1}^P, W^E), F(I_{i2}^P, W^E), \ldots, F(I_{iK}^P, W^E))$$

As we already defined, $\{I_{i1}^P, I_{i2}^P, \ldots, I_{iK}^P\}$ is a pool of faces from the personality dataset where each face $I_{iK}^P$ is randomly sampled from its corresponding segment $S_{iK}^E$. $F(I_{iK}^P, W^E)$ represents the emotion network with parameters $W^E$ which operates on face $I_{iK}^P$ to give preliminary results on the emotion scores. RAM employs the same consensus strategy among all the faces from the video to output the aggregated personality score $R$ of video $V_i^P$:

$$R(E_i, W^R) = R(G(E_i), W^R),$$

where $W^R$ represents the weights of RAM. RAM is trained by optimizing the following objective function:

$$\mathcal{L}_{RAM}(W^R) = \sum_{i \in \mathbb{N}^P} \text{smooth}_t(Y_i^P - R(E_i, W^R))$$

3.5 Overall Loss Functions

Every module of PersEmoN is differentiable, allowing end-to-end optimization of the whole system. The learning process of PersEmoN aims to minimize the following loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{per}(W^P) + \lambda_2 \mathcal{L}_{emo}(W^E) + \lambda_3 \mathcal{L}_D(W^D) + \lambda_4 \mathcal{L}_{adv}(W^P, W^E) + \lambda_5 \mathcal{L}_{RAM}(W^R)$$

4 EXPERIMENTS

4.1 Dataset and Evaluation Protocol

We choose two large-scale challenging datasets to investigate PersEmoN. The Aff-Wild emotion dataset [19] consists of 298 YouTube videos (252 for training and 46 for testing) with a total length of about 30 hours (over 1M frames). The videos show the reaction of individuals to various clips from movies, TV series, trailers, etc. Each video is labeled by 6–8 annotators with frame-wise valence and arousal values, with a total of 200 annotators. Both valence and arousal values range from −1 to 1. An example of the relationship between emotions arousal/valence values is illustrated in Fig. 2. For personality, we use the ChaLearn personality dataset [3], which consists of 10k short video clips with 41.6 hours (4.5M frames) in total. In this dataset, people face and speak to the camera. Each video is annotated with personality attributes as the Big Five personality traits in [0, 1]. The annotation was done via Amazon Mechanical Turk.

Since this dataset aims at helping job interviews, there is another labeled value which reflects the willingness to interview this individual, but we do not consider it in our paper.

To assess the quality of emotion predictions from our PersEmoN, we calculate the mean square errors (MSEs) between the predicted values of personality traits and ground truth. For the evaluation of the personality recognition, we apply two metrics used in ECCV 2016 ChaLearn First Impression Challenge [3], namely mean accuracy $A$ and coefficient of determination $R^2$, which are defined as
Arousal 0.161 early
late 0.095 

Video 0.140
0.125 late early

Modality 0.130
0.088 early
0.134
0.123
0.108

Valence
were evaluated by the official organizer.

employed in FATAUVA-Net and multi-scale inputs adopted in MM-Net, multiple datasets used for cascade learning approaches, such as ensemble of memory networks used in PersEmoN.

Videos on the test data.

shows competitive accuracy to these state-of-the-art methods.

The margin parameter in all the smooth ℓ1 loss (Eq. (3)) is set to m = 0.05.

4.2 Implementation

We initialize FEM with a truncated 20 layer version of the SphereFace model [20]. PAM is embodied with a fc layer with 5 outputs, while EAM has only 2 output neurons in the fc layer. We use sigmoid and tanh to squash the outputs for PAM and EAM respectively. We use a single-hidden-layer feed-forward network to analyze the emotion-to-personality relationship. More specifically, the RAM module is implemented with two fc layers where the first one receives 2 emotion scores as input and output 100 features with ReLU nonlinearity. The same consensus function and sigmoid nonlinearity are used to obtain the personality traits for RAM.

PersEmoN is implemented in Caffe [35]. We train the whole network with an initial learning rate of 0.01. For each mini-batch, we randomly select 100 images from the Aff-Wild dataset and 10 videos from Chalearn. For each video, 10 frames are further sparsely sampled in a randomized manner, i.e. K = 10. Hence, the overall batch size is equal to 200. We train the network for 50k iterations and decrease the learning rate by a factor of 10 in the λth iteration. λ1 = 1, λ2 = 1, λ3 = 0.1, λ4 = 0.1 and λ5 = 0.1. The margin parameter in all the smooth ℓ1 loss (Eq. (3)) is set to m = 0.05.

4.3 Evaluation of Emotion

We first report the results of emotion predictions on the Aff-Wild dataset. PersEmoN is compared with a strong baseline method CNN-M and 3 benchmark methods from the Aff-Wild challenge [19]. As demonstrated in Table 1, our method shows competitive accuracy to these state-of-the-art methods on the test data.

Simplicity is central to our design; the strategies adopted in PersEmoN are complementary to those more complicated approaches, such as ensemble of memory networks used in MM-Net, multiple datasets used for cascade learning employed in FATAUVA-Net and multi-scale inputs adopted in DRC-Net. Furthermore, all these other methods are much more difficult to train than ours. Multiple LSTM layers are used in MM-Net and DRC-Net, while FATAUVA-Net cannot perform end-to-end but cascade training.

4.4 Evaluation of Personality

Recognition of Big Five personality traits appears more interesting to us because personality is a higher-level feature compared to emotion. Table 2 lists the comparison of the details of several latest personality recognition methods. In contrast to other approaches, ours can be trained end-to-end using only one pre-trained model. Moreover, unlike most methods which fuse both acoustic and visual cues, our PersEmoN uses only video as input.

The quantitative comparison between PersEmoN and state-of-the-art works on personality recognition is shown in Table 3. The teams from NJU-LAMDA to BU-NKU-v1 are the top five participants in the 1st ChaLearn Challenge on First Impressions [3]. Note that BU-NKU was the only team not using audio in the challenge, and their predictions were rather poor comparatively. After adding the acoustic cues, the same team won the 2nd ChaLearn Challenge on First Impressions [39]. Importantly, PersEmoN only considers visual streams. Yet as is evident in Table 3, even when only taking into account PAM, PersEmoN already achieves superior performance over others, not only on the average A and R2 scores, but both scores for all traits.

Since RAM can also predict the personality attributes from the output of EAM, as shown in Fig. 1, it can provide our personality network with complementary information. To demonstrate this, we fuse the predicted attributes of both RAM and PAM; we use late fusion by a weighted average which gives the weight of 6 for the personality network and 1 for the RAM. The results are presented in Table 3 as “PAM+RAM”. In this case, we observe another performance boost and the highest overall accuracy.

4.5 Emotion-to-Personality Relationship

Big Five personality traits are usually analyzed from lifelog data or questionnaires [40]. Here we show the possibility of determining personality traits from 2-dimensional affective components. As can be noticed in Table 3 under “Ours (RAM)”, we achieve satisfactory personality predictions with only 2-dimensional arousal-valence inputs.
An illustration of the emotion-to-personality relationship is shown in Fig. 3, where each “disk” represents a certain personality trait with respect to the corresponding values of arousal and valence. The discoveries are consistent with [41]: Agreeableness and Conscientiousness are fairly near each other (the two traits share similar emotions), while Neuroticism is located far away from Openness. The “disk” for Extraversion (not shown in the Figure) is close to Agreeableness. This demonstrates that our RAM network indeed has the ability of learning the emotion-to-personality relationship. Based on this, we believe that PersEmoN can serve as a strong practical baseline for automatically annotating personality based on arousal and valence.

![Fig. 3. Illustration of the relationship between various personality traits and the arousal-valence emotion space, acquired from the input and output of RAM. Best viewed in color.](image)

Table 4 illustrates the effectiveness of this strategy. As the annotations for the test set of Aff-Wild are not released, we divide the original training set into training and validation set with a ratio of 10 : 1 and evaluate all models on the validation set for the emotion task using MSE. We believe our improvement originates from the back-propagation training of CNN, during which the shared parameters within the FEM will directly impact the generalization ability of the whole system.

### 4.6.2 Consensus Function \( \mathcal{G} \)

Average temporal pooling has been reported to work well in modeling long-term temporal dependencies for deeply learned representations by [32]. This is also in line with our empirical results on personality recognition. To demonstrate this, we compare average pooling with two other alternatives. One is max pooling, which helps to select the most salient information in its receptive field and has been heavily encoded in popular network structure such as ResNet, VGG and so on. The other is recurrent aggregation, for which we choose the popular LSTM [42]. LSTM has been shown to work better than conventional recurrent networks due to its learnable memory gate to avoid gradient vanishing or explosion. In our implementation, both feature representations from FEM as well as LSTM are jointly optimized. We achieve an accuracy of 91.4\%, 90.6\% and 90.1\% for average pooling, max pooling and LSTM, respectively. Max pooling performs worse than average pooling and better than LSTM. This indicates that selecting the most

| Emotion | Personality | Relationship | Aff-Wild | ChaLearn |
|---------|-------------|--------------|----------|----------|
| ✔️      | ✔️          | ✔️           | 0.096    | -        |
| ✔️      | ✔️          | ✔️           | 0.057    | -        |
| ✔️      | ✔️          | ✔️           | 0.033    | -        |
| ✔️      | ✔️          | ✔️           | 0.071    | 0.027    |

**Table 4.** Effectiveness of jointly training of PersEmoN. Values are MSEs of prediction.

#### 4.6 Ablation Study

**4.6.1 Effectiveness of Joint Training**

Our novel multi-task learning aims at learning a generalizable representation, which is applicable not only to the task in question, but also to other tasks with significant commonalities. In PersEmoN, since a shared FEM is employed by all tasks, additional tasks act as regularization, which requires the system to perform well on a related task. The backpropagation training from different tasks will directly impact the representation learning of shared parameters. It prevents overfitting by solving all tasks jointly and allowing for the exploitation of additional training data.

Table 4 illustrates the effectiveness of this strategy. As the annotations for the test set of Aff-Wild are not released, we divide the original training set into training and validation set with a ratio of 10 : 1 and evaluate all models on the validation set for the emotion task using MSE. We believe our improvement originates from the back-propagation training of CNN, during which the shared parameters within the FEM will directly impact the generalization ability of the whole system.

**Table 3.** Personality prediction benchmarking using mean accuracy and coefficient of determination \( R^2 \) scores. Note that there are no \( R^2 \) scores reported for BU-NKU-v2.

|          | Average | Extraversion | Agreeableness | Conscientiousness | Neuroticism | Openness |
|----------|---------|--------------|---------------|-------------------|-------------|----------|
|          | \( A \) | \( R^2 \)    | \( A \) | \( R^2 \) | \( A \) | \( R^2 \) | \( A \) | \( R^2 \) | \( A \) | \( R^2 \) |
| PAM+RAM  | 0.917   | 0.485       | 0.920 | 0.552 | 0.914 | 0.349 | 0.921 | 0.570 | 0.914 | 0.500 | 0.915 | 0.457 |
| Ours (PAM) | 0.916   | 0.478       | 0.920 | 0.544 | 0.913 | 0.338 | 0.921 | 0.571 | 0.915 | 0.489 | 0.914 | 0.448 |
| Ours (RAM) | 0.903   | 0.373       | 0.911 | 0.449 | 0.908 | 0.264 | 0.902 | 0.349 | 0.908 | 0.442 | 0.907 | 0.364 |
| NJULAMDA | 0.916   | 0.485       | 0.913 | 0.481 | 0.913 | 0.338 | 0.917 | 0.344 | 0.910 | 0.475 | 0.912 | 0.437 |
| evolgen  | 0.902   | 0.440       | 0.915 | 0.315 | 0.912 | 0.239 | 0.912 | 0.488 | 0.910 | 0.458 | 0.912 | 0.414 |
| DCC      | 0.911   | 0.441       | 0.911 | 0.431 | 0.910 | 0.296 | 0.914 | 0.478 | 0.910 | 0.448 | 0.911 | 0.402 |
| ucas     | 0.910   | 0.439       | 0.913 | 0.489 | 0.909 | 0.292 | 0.911 | 0.520 | 0.906 | 0.457 | 0.910 | 0.439 |
| BU-NKU-v1 | 0.909   | 0.394       | 0.916 | 0.514 | 0.907 | 0.234 | 0.913 | 0.487 | 0.902 | 0.363 | 0.908 | 0.372 |
| BU-NKU-v2 | 0.913   | -           | 0.918 | -    | 0.907 | -    | 0.915 | -    | 0.911 | -    | 0.914 | -    |
salient information from a video frame does not necessarily capture its overall statistics better. The reason for the failure of LSTM could be that personality is an orderless concept where temporal dependencies may not be so relevant.

4.6.3 Number of Segments $K$

In our implementation, $K = 10$. We empirically find that the personality results are not sensitive when $K$ is within $[5, 20]$. However, when both emotion and personality network are jointly optimized, we observe that a balanced input can always be beneficial in both domains. We use a batch size of 100 for both emotion and personality datasets. In this way, 10 input videos for personality are used in each batch. Setting $K$ to a larger value, for example 100, will lead to a lower number of either input videos for personality or emotion frames. This further reduces the final performance in both domains.

4.6.4 Coherence Strategy

As reported by [34], a representation with good transferability should be domain invariant. We observe that this strategy leads to around 1% improvement in terms of MSE for Aff-Wild and 0.5% on mean accuracy for the Chalearn dataset, respectively. We visualize the distribution of the deeply learned features from FEM (the fc5 layer of SphereFace) in Fig. 4. More specifically, we project the 512-dimensional features on both emotion and personality datasets into 2 dimensional space and visualize their distributions using t-SNE [43]. Without a coherence strategy, distributions of those deep features on different domains can be well classified, i.e. except for the center part, features from emotion dataset are mainly distributed in the outer ring of the $x−y$ plane. Using the coherence strategy, a large number of features from the emotion dataset are pulled inside the ring, making the two distributions more similar.

5 Conclusions

For the first time, we investigate the feasibility of jointly analyzing apparent personality, emotion, and their relationship within a single deep neural network. This is challenging due to the scarceness of datasets which encompass both emotion and personality annotations. To tackle this issue we propose PersEmoN, an end-to-end trainable deep network with two CNN branches called emotion and personality network. With shared bottom feature extraction layers, these two networks regularize each other within a multi-task learning framework, where each one is dedicated to their own annotated dataset. We further employ an adversarial-like loss function to promote representation coherence between heterogeneous dataset sources, which leads to further performance boosts. We demonstrate the feasibility of PersEmoN on two personality and emotion datasets. We find that the proposed joint training of both emotion and personality networks can lead to a more generalizable representation for both tasks.

Acknowledgement

This research is supported by the SERC Strategic Fund from the Science and Engineering Research Council (SERC), A*STAR (project no. a1718g0046).

References

[1] W. Revelle and K. R. Scherer, “Personality and emotion,” in Oxford Companion to Emotion and the Affective Sciences. Oxford University Press, 2009, pp. 304–305.
[2] H. J. Eysenck, Dimensions of Personality. Transaction Publishers, 1950, vol. 5.
[3] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, “Chalearn LAP 2016: First round challenge on first impressions – dataset and results,” in Proc. ECCV, 2016.
[4] J. A. Russell, “A circumplex model of affect,” J. Personality and Social Psychology, vol. 39, no. 6, p. 1161, 1980.
[5] S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker, “Lexical predictors of personality type,” in Proc. Interface and Classification Society of North America, 2005.
[6] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrina, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, “Salsa: A novel dataset for multimodal group behavior analysis,” IEEE TPAMI, vol. 36, no. 8, pp. 1707–1720, 2016.
[7] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe, “ASCERTAIN: Emotion and personality recognition using commercial sensors,” IEEE Trans. Affective Computing, 2016.
[8] M. Gavrilescu and N. Vizireanu, “Predicting the sixteen personality factors (16PF) of an individual by analyzing facial features,” EURASIP J. Image and Video Processing, vol. 2017, no. 1, p. 59, 2017.
[9] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in Proc. NIPS, 2014.
[10] O. M. Parkhi, A. Vedaldi, A. Zisserman et al., “Deep face recognition,” in Proc. BMVC, 2015.
[11] 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, no. 10, pp. 1499–1503, 2016.
[12] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a “Siamese” time delay neural network,” in Proc. NIPS, 1994.
[13] S. Peng, L. Zhang, S. Winkler, and M. Winslett, “Give me one portrait image, i will tell you your emotion and personality,” in ACM MM, 2018.
[14] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in Proc. CVPR, 2014.
[15] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in Proc. CVPR, 2014.
[16] Y. Sun, D. Liang, X. Wang, and X. Tang, “DeepID3: Face recognition with very deep neural networks,” arXiv preprint arXiv:1502.00873, 2015.
[17] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, 2007.
[18] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in Proc. CVPR, 2015.
[19] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, “Aff-wild: Valence and arousal in-the-wild challenge,” in Proc. CVPRW, 2017.

[20] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in Proc. CVPR, 2017.

[21] Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotion recognition,” in Proc. ICASSP, 2013.

[22] L. Surace, M. Patacchiola, E. Battini Sömmez, W. Spataro, and A. Cangelosi, “Emotion recognition in the wild using deep neural networks and bayesian classifiers,” in Proc. ICMI, 2017.

[23] H. Ranganathan, S. Chakraborty, and S. Panchanathan, “Multimodal emotion recognition using deep learning architectures,” in Proc. WACV, 2016.

[24] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild,” in Proc. CVPR, 2016.

[25] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” IEEE Trans. Affective Computing, 2017.

[26] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, “Deep bimodal regression for apparent personality analysis,” in Proc. ECCV, 2016.

[27] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. van Lier, “Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition,” in Proc. ECCV, 2016.

[28] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramaniam, and A. Mittal, “Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features,” in Proc. ECCV, 2016.

[29] F. Gürpınar, H. Kaya, and A. A. Salah, “Combining deep facial and ambient features for first impression estimation,” in Proc. ECCVW, 2016.

[30] J. Junior et al., “First impressions: A survey on computer vision-based apparent personality trait analysis,” arXiv preprint arXiv:1804.08046, 2018.

[31] H. J. Escalante et al., “Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos,” arXiv preprint arXiv:1802.00745, 2018.

[32] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in Proc. ECCV, 2016.

[33] R. Girshick, “Fast R-CNN,” in Proc. ICCV, 2015.

[34] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, “Simultaneous deep transfer across domains and tasks,” in Proc. ICCV, 2015.

[35] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Proc. ICMI, 2014.

[36] J. Li, Y. Chen, S. Xiao, J. Zhao, S. Roy, J. Feng, S. Yan, and T. Sim, “Estimation of affective level in the wild with multiple memory networks,” in Proc. CVPRW, 2017.

[37] W.-Y. Chang, S.-H. Hsu, and J.-H. Chien, “FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation,” in Proc. CVPR, 2017.

[38] B. Hasani and M. H. Mahoor, “Facial affect estimation in the wild using deep residual and convolutional networks,” in Proc. CVPRW, 2017.

[39] H. J. Escalante, V. Ponce-López et al., “Chalearn joint contest on multimedia challenges beyond visual analysis: An overview,” in Proc. ICPR, 2016.

[40] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, “AMIGOS: A dataset for mood, personality and affect research on individuals and groups,” arXiv preprint arXiv:1702.02510, 2017.

[41] M. S. Yik and J. A. Russell, “Predicting the big two of affect from the big five of personality,” J. Research in Personality, vol. 35, no. 3, pp. 247–277, 2001.

[42] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[43] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” JMLR, vol. 9, pp. 2579–2605, Nov 2008.

Le Zhang received the B.Eng degree from University of Electronic Science and Technology Of China (UESTC) in 2011. He received his M.Sc and Ph.D degree form Nanyang Technological University (NTU) in 2012 and 2016, respectively. Currently, he is a scientist at Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore. Prior to that, he is a postdoctoral researcher at the University of Illinois Advanced Digital Sciences Center (ADSC), Singapore. His current research interests include deep learning and computer vision.

Songyou Peng received the Erasmus Mundus M.Sc. in Computer Vision and Robotics in 2015. Between 2016 and 2017, he spent some time doing research at INRIA Grenoble and Technical University of Munich. He is currently a research engineer at Institute for Infocomm Research, A*STAR, Singapore. He is also an adjunct research engineer at Advanced Digital Sciences Center, University of Illinois. His research interests are computer vision and machine learning.

Stefan Winkler is the deputy director of AI Singapore. He is also an adjunct distinguished scientist and a director of the Video & Analytics Program with the University of Illinois's Advanced Digital Sciences Center (ADSC), Singapore. Prior to that, he co-founded a startup, worked for a Silicon Valley company, and held faculty positions with the National University of Singapore and the University of Lausanne, Switzerland. He has published more than 100 papers and the book Digital Video Quality (Wiley). He is an associate editor of the IEEE Transactions on Image Processing. He is a fellow of the IEEE.