Deep Bimodal Regression of Apparent Personality Traits from Short Video Sequences

Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu, *Member, IEEE*

Abstract—Apparent personality analysis (APA) is an important problem of personality computing, and furthermore, automatic APA becomes a hot and challenging topic in computer vision and multimedia. In this paper, we propose a deep learning solution to APA from short video sequences. In order to capture rich information from both the visual and audio modality of videos, we tackle these tasks with our Deep Bimodal Regression (DBR) framework. In DBR, for the visual modality, we modify the traditional convolutional neural networks for exploiting important visual cues. In addition, taking into account the model efficiency, we extract audio representations and build a linear regressor for the audio modality. For combining the complementary information from the two modalities, we ensemble these predicted regression scores by both early fusion and late fusion. Finally, based on the proposed framework, we come up with a solution for the Apparent Personality Analysis competition track in the ChaLearn Looking at People challenge in association with ECCV 2016. Our DBR is the winner (first place) of this challenge with 86 registered participants. Beyond the competition, we further investigate the performance of different loss functions in our visual models, and prove non-convex loss functions for regression are optimal on the human-labeled video data.

Index Terms—Apparent personality analysis, deep learning, bimodal data, convolutional neural networks, regression.

1 Introduction

Automated analysis of human affective behavior has attracted increasing attention from researchers in psychology, computer science, linguistics, neuroscience, and other related disciplines. In the field of computer science, personality computing studies how machines could automatically recognize or synthesize human personality [1].

In particular, apparent personality analysis (APA) is an important problem of personality computing. Personality is relevant to any computing area involving understanding, prediction or synthesis of human behavior, which is also a strong predictor of important life outcomes like happiness, satisfaction, longevity, quality of relationships with peers, family, occupational choice, community involvement, criminal activity and political ideology [2], [3]. The Big Five traits (or the Five Factor Model) is currently the dominant paradigm in personality analysis, and is widely accepted in the computing community as well [1]. In the Big Five model, personality traits are decomposed into five components, including *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*. A learning algorithm for personality traits prediction is to automatically predict the Big Five traits values based on cues like still facial images, speech signals or human-centered videos.

In the ChaLearn Looking At People (LAP) challenge [4] in association with ECCV 2016, the organizers proposed a novel and large benchmark dataset for apparent personality analysis. The dataset contains 10,000 human-centered short video sequences of about 15 seconds each. These videos are collected from YouTube, and annotated with the Big Five traits by Amazon Mechanical Turk (AMT) workers. For dealing with the personality traits prediction from short videos task, in this paper, we propose the Deep Bimodal Regression (DBR) framework for APA. As shown in Fig. 1, DBR treats human-centered videos as having two modalities, i.e., the visual and the audio modality. Then, in these two modalities, deep visual regression networks and audio regression models are built for capturing both visual and audio information for the final personality traits predictions.

Specifically, in the visual modality, we firstly extract frames from each original video. Then, deep convolutional neural networks are adopted to learn both visual representations and deep regressors for predicting the Big Five traits values. Inspired by our previous work [5], [6], in these visual deep regression networks, we modify the traditional CNN architecture by discarding the fully connected layers. And then, the deep descriptors of the last convolution layer are both averaged and max pooled into 512-d feature vectors. After that, the standard ℓ_2-normalization is followed. Finally, the feature vectors are concatenated into the final 1024-d image representations, and a regression (fc+sigmoid) layer is added for end-to-end training. The modified CNN model is called Descriptor Aggregation Network (DAN). Furthermore, the ensemble of multiple layers is used for boosting the regression performance of the visual modality, which is regarded as DAN$^+$. Beyond DAN and DAN$^+$, Residual Networks [7] is also utilized in our visual modality of DBR. As discussed in Sec. 5.3.1, the epoch fusion is used as the early fusion to boost the visual regression performance.

Additionally, because the ground truth of the ChaLearn LAP competition dataset is obtained by the workers of AMT, there should be label noises in these human-labeled data. Thus, we introduce different loss functions into DANs,

- *The first two authors contributed equally to this work. All authors are with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. J. Wu is the corresponding author. E-mail: {weixs, zhangcl, zhangh, wujx}@lamda.nju.edu.cn.*
- *This research was supported by National Natural Science Foundation of China under Grant 61772256 and Grant 61422203, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.*

including convex loss functions (e.g., the ℓ_1 and ℓ_2 loss) and non-convex loss functions (e.g., the Tukey's biweight loss function [8]). The experimental results in Sec. 5.4.1 could validate the effectiveness of non-convex loss functions for personality traits predictions, which also proves non-convex loss functions can reduce label noises and further improve the regression performance.

On the other hand, in the audio modality, not only we use the handcrafted audio features to build a personality traits regressor, but also reimplement a deep learning based audio network [9] for directly learning the audio representations from the raw audio signals. For the handcrafted one, the log filter bank (logfbank) [10] features are extracted from the original audio of each video. Based on the logfbank features, we train a linear regressor to obtain the Big Five traits values. In consideration of its computational efficiency, in the ChaLearn LAP challenge, we choose the logfbank features as the audio representation, and then build a linear regressor based on logfbank. Finally, the visual modality and audio modality are lately fused by averaging the scores of these deep visual models and the audio model. In addition, the deep learning based audio networks could achieve better personality traits prediction results than the handcrafted audio features, cf. Table 1.

In consequence, based on the proposed DBR framework, we come up with a solution for the Apparent Personality Analysis track in the ChaLearn Looking at People (LAP) challenge. In the Final Evaluation phase, our DBR framework achieved the best regression accuracy (0.9130 mean accuracy), which ranked *the first place* in this competition. The source codes and models of the APA challenge are available on our project page, i.e., http://lamda.nju.edu.cn/weixs/project/APA/APA.html.

The paper is an extension based on our previous work [11] published in proceedings of the ChaLearn Looking at People Workshop associated with ECCV 2016. The rest of this paper is organized as follows. Sec. 2 introduces the related work about apparent personality analysis, visual-based deep learning and audio representations. Sec. 3 gives detailed descriptions about the apparent personality analysis from short videos. The details of the proposed Deep Bimodal Regression framework are presented in Sec. 4. In Sec. 5, we present our implementation details of DBR, the experimental results of this challenge, and the analysis of the proposed DAN models. Sec. 6 concludes the paper.

2 Related Work

In this section, we will briefly review the related work for apparent personality analysis, visual-based deep learning and audio representations.

2.1 Apparent Personality Analysis

Originally, personality analysis is a task that is specific to the psychology domain. Previous researches in personality analysis usually need psychology scientists to figure out the results, or need participants to do specific tests containing large number of questions which can reflect their personalities. However, such process will cost a lot of time and funds. With personality computing becoming more and more popular [1], automatically Apparent Personality Analysis (APA) also becomes a hot topic in computer science, psychology, neuroscience and so on.

In the field of computer science, APA methods were proposed for recognizing personality from nonverbal aspects of verbal communication [12], multimodal combinations of speaking style (e.g., prosody, intonation) and body movements [13], [14], combining acoustic with visual cues or physiological with visual cues [15], [16], and so on.

In particular, a similar task to automatic personality analysis in computer vision is the emotion analysis task, e.g., [17], [18]. Emotion analysis can be regarded as a multiple class classification problem, where usually six basic emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*) are recognized by the algorithms. However, in apparent personality analysis, it needs to predict the Big Five traits (*openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*) which are independent of each other and whose scores are continuous values in the range of $[0, 1]$. Thus, it is obvious to see the apparent personality analysis tasks is more realistic but difficult than emotion analysis.

In this paper, we pay our attention on one kind of specific automatic apparent personality analysis task, i.e., predicting personality traits from short video sequences.

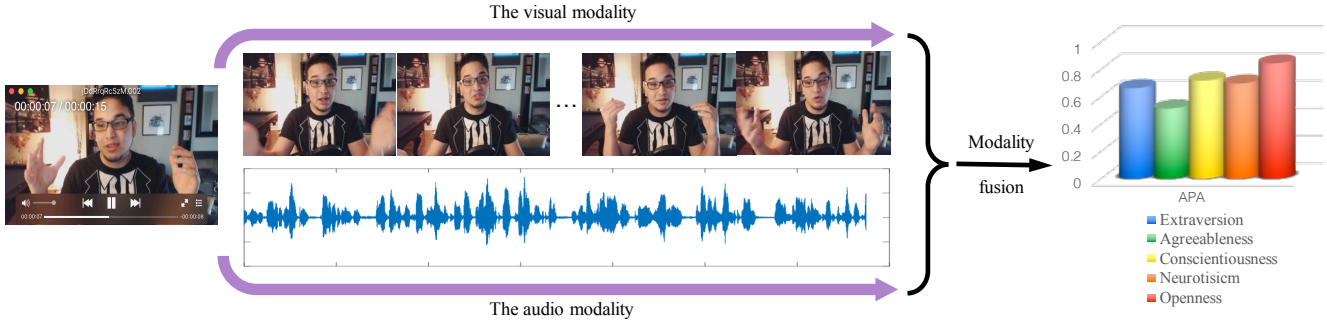
2.2 Visual-Based Deep Learning

Deep learning refers to a class of machine learning techniques, in which many information processing layers organized in a sequential structure are exploited for pattern classification and for feature or representation learning.

Recently, for image-related tasks, Convolutional Neural Networks (CNNs) [19] allow computational models that are composed of multiple processing layers to learn representations of images with multiple levels of abstraction, which have been demonstrated as an effective class of models for understanding image content, giving state-of-the-art results on image recognition, segmentation, detection and retrieval. Specifically, the CNN model consists of several convolution layers and pooling layers, which are stacked up with one on top of another. The convolution layer shares many weights, and the pooling layer sub-samples the output of the convolution layer and reduces the data rate from the layer below. The weight sharing in the convolution layer, together with appropriately chosen pooling schemes, endow the CNN with some invariance properties (e.g., translation invariance).

Specifically, in [20], emotion analysis for user generated videos was performed through the extraction of deep convolution network features and through zero-shot emotion learning, a method that predicts emotions not observed in the training set. To implement this task, image transfer encoding (ITE) was proposed to encode the extracted features and generate video representations. Recently, Poria et al. [21] developed a convolutional recurrent neural network (RNN) to extract visual features. In their study, a CNN and RNN have been stacked and trained together for dealing with the emotion analysis problem.

In our DBR framework, we employ and modify multiple CNNs to learn the image representations for the visual modality, and then obtain the Big Five traits predictions.



(a) Short video sequences (b) Proposed Deep Bimodal Regression framework (c) Predicted Big Five Traits

Figure 1. Framework of the proposed Deep Bimodal Regression (DBR) method for video-based apparent personality analysis. In DBR, the original videos are treated as having two natural modalities, i.e., the visual modality for images and the audio modality for speeches. After learning the (deep) regressors on these two modalites, the final predicted personality traits are obtained by late fusion.

2.3 Audio Representations

In the past decades, many handcrafted representations for audio have been proposed: some of them are time domain features, and others are frequency domain features. Among them, there are several famous and effective audio features, to name a few, Mel Frequency Cepstral Coefficients (MFCC) [10], Linear Prediction Cepstral Coefficient (LPCC) [22] and Bark Frequency Cepstral Coefficient (BFCC) [22]. These handcrafted audio features are computationally efficient and widely used in diverse applications, e.g., speech emotion recognition [23], [24], music information retrieval [25], [26], facial/lip synchronization [27], [28], etc.

Recently, deep learning based methods attract more and more attentions in the audio processing community. Bhargava and Rose [29] used stacked bottleneck deep neural networks trained on windowed speech waveforms and obtained results only slightly worse than corresponding MFCC on the same architecture. Sainath et al. match the performance of a large-vocabulary speech recognition system based on log-Mel filter bank energies by using a Convolutional, LSTM-DNN [30]. They observed that a time convolution layer helps in reducing temporal variation, another frequency convolution layer aids in preserving locality and reducing frequency variation, while the LSTM layers serve for contextual modeling of the speech signal. Palaz et al. [31] used CNNs directly trained on the speech signal to estimate phoneme class conditional probabilities and observed that the features learnt between the first two convolution layers tend to model the phone-specific spectral envelope of subsegmental speech signal, which leads to a more robust performance in noisy conditions. In 2016, Trigeorgis et al. [9] proposed a convolutional recurrent model that operates on the raw signal, to perform an end-to-end spontaneous emotion prediction task from speech data.

3 Apparent Personality Analysis from Short Video Sequences

Apparent personality analysis (APA) is an important problem of human affective behavior analysis, which is inherently a multidisciplinary enterprise involving different research fields, including computer vision, psychology, linguistics, multimedia, speech analysis and machine learning [32]. In this paper, we focus on the APA task from short video sequences.

Video analysis is one of the key tasks in computer vision and multimedia research, especially human-centered video analysis. In recent years, human-centered videos have become ubiquitous on the internet, which has encouraged the development of algorithms that can analyze their semantic contents for various applications, including first-person video analyses [33], [34], activity recognition [35], [36], gesture and pose recognition [37], [38] and many more [39], [40], [41], [42]. Here, the goal of APA from videos is to develop algorithms for recognizing personality traits of users in short video sequences.

In the following, we will briefly describe the personality traits predictions in APA and the corresponding ChaLearn Looking At People (LAP) 2016 Challenge [4].

3.1 Personality Traits Predictions

Personality plays an important role in the way people manage the images they convey in self-presentations and employment interviews, trying to affect the audience first impressions and increase effectiveness [43], [44], [45]. The Big Five traits model is currently the dominant paradigm in personality research. Personality traits are usually decomposed into five components, including *openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*. Generally, automatic personality traits prediction requires the learning algorithms/methods to return five continuous values in the range of $[0, 1]$ which correspond to the Big Five traits values. Effective personality traits prediction is challenging due to several factors: cultural and individual differences in tempos and styles of articulation, variable observation conditions, the small size of faces in images taken in typical scenarios, noise in camera channels, infinitely many kinds of out-of-vocabulary motion, and real-time performance constraints.

3.2 ChaLearn Looking At People (LAP) Challenge

The ECCV ChaLearn LAP 2016 challenge [4] consisted of a single track competition to quantitatively evaluate the

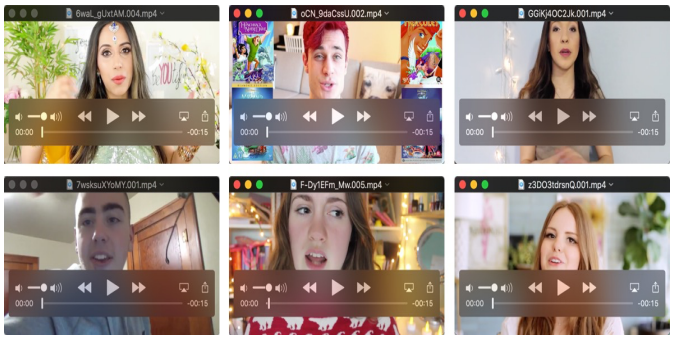


Figure 2. Sampled video demos of the ChaLearn Looking At People competition dataset [4].

recognition of the apparent Big Five personality traits on bimodal audio+RGB data from YouTube videos. In this challenge, the organizers proposed a novel data set consisting of 10,000 human-centered short video sequences. The ground truth consists of five fractional scores (corresponding to the Big Five traits values) in the range between 0 to 1 for each video, which was obtained from workers of Amazon Mechanical Turk (AMT). The video demos are shown in Fig. 2.

The ChaLearn LAP challenge had two phases:

- A development phase during which the participants had access to 6,000 manually labeled continuous video sequences of 15 seconds each. Thus, 60% of the videos are used for training. The participants could get immediate feedback on their prediction performance by submitting results on an unlabeled validation set of 2,000 videos. These 2,000 videos used in validation represent 20% over the total set of videos.
- A final phase during which the participants could submit their predictions on 2,000 new test videos (the remainder 20% over the total set of videos). The prediction scores on test data were not revealed until the end of the challenge.

The challenge attracted totally 86 registered participants, and lasted over two months. At last, ten teams entered the final phase. Our *NJU-LAMDA* team achieved the best personality traits prediction results, and ranked the first place in this challenge.

4 The Proposed DBR Framework

In this section, we will introduce the proposed Deep Bimodal Regression (DBR) framework for the apparent personality analysis task in the ChaLearn LAP challenge. As shown in Fig. 1, DBR has three main parts: the first part is the visual modality regression, the second part is the audio one, and the last part is the ensemble process for fusing information of the two modalities.

4.1 Deep Regression for the Visual Modality

The deep regression of the visual modality contains three subparts: image extraction, deep regression network training and regression scores prediction. Furthermore, as aforementioned, because the ground truth for personality traits was obtained from workers of Amazon Mechanical Turk, there should exist label noises. We here introduce several different loss functions into DANs for the networks training, and show the non-convex loss functions could resist the label noises and achieve better performance.

4.1.1 Image Extraction

The inputs of traditional convolutional neural networks are single images. But for the APA task, the original inputs are the human-centered videos where each video has its corresponding Big Five traits values. In order to utilize powerful CNNs to capture the visual information, it is necessary to extract images from these videos. For example, for a fifteen seconds length video whose frame rate is 30fps, there are 450 images/frames from each original video. However, if all the images/frames are extracted, the computational cost and memory cost will be quite large. Besides, in fact, nearby frames look extremely similar. Therefore, we down-sample these images/frames to roughly 100 images per video. That is to say, we extract about 6 images in one second from a video. After that, the extracted images/frames from one video are labeled with the same Big Five traits values as the values of their corresponding video. In consequence, based on the images, we can train deep regressors by CNNs for apparent personality analysis.

4.1.2 Deep Regression Network Training

In the visual modality of DBR, the main deep CNN models are modified based on our previous work [5], [6], which are called Descriptor Aggregation Networks (DANs). What distinguishes DAN from the traditional CNN is: an additional layer equipped with both average- and max-pooling is added between the last convolutional layer (Pool_5) and the final fully connected layer (fc+sigmoid). The similar designed principle can be also found in recent CNN models such as Inception [46] and ResNet [7]. Meanwhile, in our DANs, each pooling operation is followed by the standard ℓ_2-normalization. After that, the obtained two 512-d feature vectors are concatenated as the final image representation. Thus, in DAN, the deep descriptors of the last convolution layers are aggregated as a single visual feature. Finally, because APA is a regression problem, a regression (fc+sigmoid) layer is added for end-to-end training. Our DAN uses a CNN with the VGG-16 [47] architecture, which is illustrated in Fig. 3. Our choice is motivated (1) by the deep but manageable architecture, (2) by the impressive accuracy achieved using VGG-16 on the ImageNet challenge [48], (3) and that pre-trained models for classification, especially VGG-Face for face recognition [49], are publicly available allowing warm starts for training.

Because DAN has no traditional fully connected layers, it will bring several benefits, such as reducing the model size, reducing the dimensionality of the final feature, and accelerating the model training. Moreover, the model performance of DAN is better than traditional CNNs with the fully connected layers, cf. Table 1 and also the experimental results in [6].

For further improving the regression performance of DAN, the ensemble of multiple layers is employed. Specifically, the deep convolutional descriptors of $\mathrm{ReLU}5\_2$ are also incorporated in the aforementioned aggregation approach,

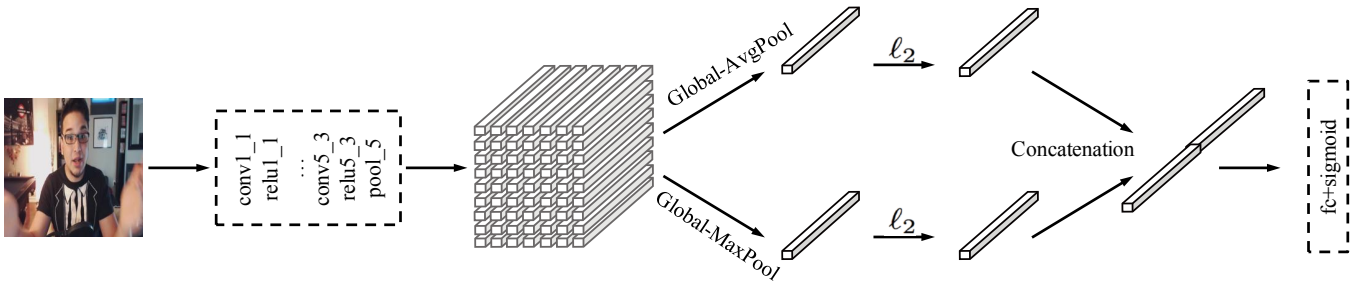


Figure 3. Architecture of the proposed Descriptor Aggregation Network (DAN) model. Note that we removed traditional fully connected layers. The deep descriptors of the last convolution layer (Pool_5) are firstly aggregated by both average- and max-pooling, and then concatenated into the final image representation for regression.

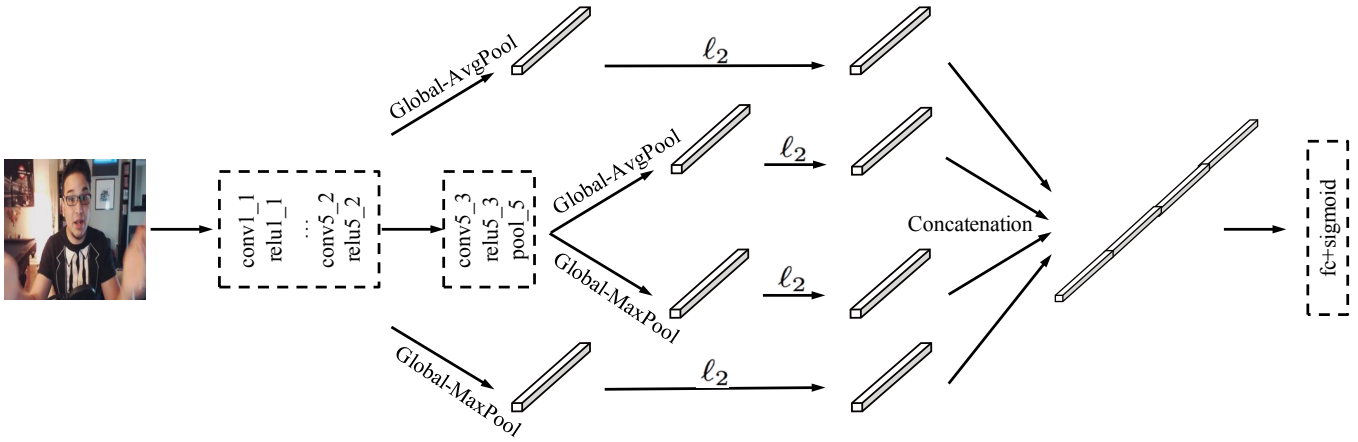


Figure 4. Architecture of the proposed DAN$^+$ model. In DAN$^+$, not only the deep descriptors of the last convolution layer (Pool_5) are used, but the ones of ReLU_{5_2} are also aggregated. Finally, the feature vectors of multiple layers are concatenated as the final image representation for regression.

which is shown in Fig. 4. Thus, the final image feature is a 2048-d vector. We call this end-to-end deep regression network as "DAN$^+$".

4.1.3 Personality Traits Prediction

In the phase of predicting regression values, images are also extracted from each testing video. Then, the predicted regression scores of images are returned based on the trained visual models. After that, we average the scores of images from a video as the predicted scores of that video.

4.1.4 Robust Deep Regression Learning

For deep regression of DAN and DAN$^+$, two traditional convex loss functions, i.e., ℓ_1 and ℓ_2, are used as the loss to be minimized during the network training. In addition, considering the ground truth label noises, we further deploy Tukey's biweight function [8] as one kind of non-convex loss function into the visual modality for robust deep regression learning.

Assume the input image is $\boldsymbol{x}$ and the output is a real-valued vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ with N elements, $y_i \in \mathbb{R}$. We can define the loss of the i-th value of vector $\boldsymbol{y}$ by:

$$l_i = y_i - \hat{y_i}\,, \tag{1}$$

where $\hat{y_i}$ represents the predicted value for the i-th value of $\boldsymbol{y}$. Thus, the ℓ_1 loss function is as follows:

$$\mathcal{L}_{\ell_1} = \sum_{i=1}^{N} |l_i|\,, \tag{2}$$

and similarly, the ℓ_2 loss function is shown as:

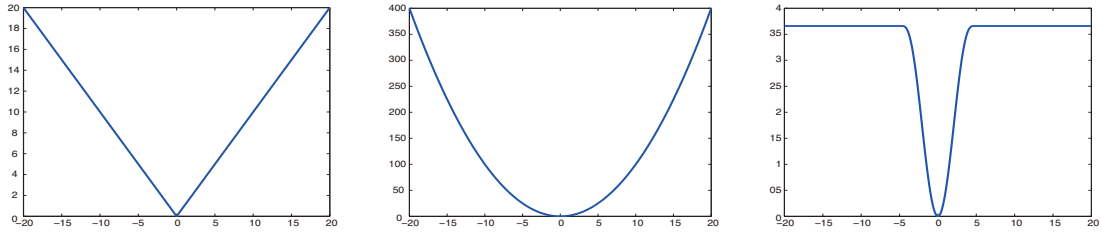
$$\mathcal{L}_{\ell_2} = \sum_{i=1}^{N} (l_i)^2\,. \tag{3}$$

Tukey's biweight loss function can be defined as:

$$\mathcal{L}_{\text{Tukey's}} = \begin{cases} \dfrac{c^2}{6}\left[1 - \left(1 - \left(\dfrac{l_i}{c}\right)^2\right)^3\right] & \text{if } |l_i| \le c \\ \dfrac{c^2}{6} & \text{otherwise} \end{cases}, \tag{4}$$

where c is a tuning constant, which is set to 4.6851, giving approximately 95% asymptotic efficiency as ℓ_2 minimization on the standard normal distribution of residuals. The functionality of ℓ_1, ℓ_2 and Tukey's biweight loss functions are illustrated in Fig. 5.

As seen from these figures, the ℓ_1 and ℓ_2 loss functions are convex, but the Tukey's biweight one is non-convex. Thus, Tukey's biweight function suppresses the influence of label noises during back-propagation by reducing the



(a) ℓ_1 loss function. (b) ℓ_2 loss function. (c) Tukey's biweight loss function.

Figure 5. Different loss functions. Among them, the ℓ_1 and ℓ_2 loss functions are convex, but Tukey's biweight one is a non-convex function.

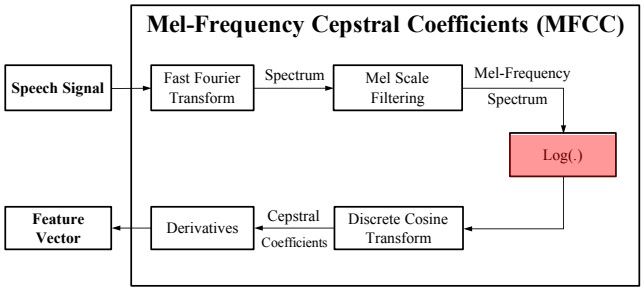


Figure 6. Pipeline of extracting the MFCC and logfbank features.

magnitude of their gradient close to zero. The experimental results of Sec. 5.4.1 could validate the effectiveness of the Tukey's biweight loss function on this kind of human-labeled video data.

4.2 Regression for the Audio Modality

In the audio modality, we use traditional handcrafted spectral audio features as audio representations, and also employ deep learning based audio models for both audio representation learning and personality traits regression.

4.2.1 Handcrafted Audio Features

In the past decades, the Mel Frequency Cepstral Coefficients (MFCC) [10] features have been widely used in the speech recognition community. As shown in Fig. 6, MFCC refers to a kind of short-term spectral-based features of a sound, which is derived from spectrum-of-a-spectrum of an audio clip. MFCC can be derived in five steps. During the five steps, the log filter bank (logfbank) features can be also obtained, which is shown in the red bounding box of Fig. 6.

In our experiments of the competition, we extract the MFCC and logfbank features from the audios of each original human-centered video for APA. MFCC of one frame is a 13-d feature vector, and logfbank is 26-d. Then, we directly concatenate these frames' feature vectors into a single feature vector of 39,767-d for MFCC and 79,534-d for logfbank. The results of logfbank are slightly better than the ones of MFCC, cf. Fig. 10. Empirical comparisons and details can be found in Sec. 5.3.1. Thus, in our solution of the ChaLearn LAP challenge, the logfbank features are used as the audio representations. After extracting the logfbank features from the original audios of videos, we use a model composed of a fully-connected layer followed by a sigmoid function layer to train a linear regressor for the audio modality regression. The whole pipeline of the audio modality can be seen in Fig. 7.

4.2.2 Deep Learning based Audio Models

Deep learning is one kind of powerful representation learning schemes, which is also effective for audio processing. Beyond the handcrafted audio features, we also try to use deep learning based audio networks to learn an audio representation for APA. After finishing the ChaLearn LAP challenge, we reimplemented an end-to-end speech emotion recognition system of [9], and further modified it to fit the personality traits prediction task.

The end-to-end audio network in [9] combines both convolutional neural networks with LSTM networks [50] for automatically learning an optimal representation of the speech signal directly from the raw audio data. The convolutional recurrent model is presented in Fig. 8. Specifically, the input raw waveform is preprocessed to have zero-mean and unit variance to account for variations in different levels of loudness between the speakers. There are two convolution layers. In the first one, it uses $F = 40$ space time finite impulse filters with a 5ms window in order to extract fine-scale spectral information from the high sampling rate signal. In the second one, $M = 40$ space time finite impulse filters of 500ms window are employed to extract more long-term characteristic of the speech and the roughness of the audio signal. For the recurrent LSTM layer, it is a traditional bidirectional LSTM architecture [50] with 128 cells each. Please refer to [9] for more details.

4.3 Modality Ensemble

After the training of both the visual and audio modalities, modality ensemble is used as the late fusion approach for getting the final regression scores. The ensemble method we used in DBR is the simple yet effective simple averaging method. In APA, the predicted result of a trained regressor is a five-dimensional vector which represents the Big Five traits values, i.e., $\boldsymbol{s}_i = (s_{i1}, s_{i2}, s_{i3}, s_{i4}, s_{i5})^T$. We treat each predicted result of these two modalities equally. For example, the predicted results of the visual modality are $\boldsymbol{s}_1$, $\boldsymbol{s}_2$ and $\boldsymbol{s}_3$, and the results of the audio one are $\boldsymbol{s}_4$ and $\boldsymbol{s}_5$. The final ensemble results are calculated as follows:

$$\text{Final score} = \frac{\sum_{i=1}^{5} \boldsymbol{s}_i}{5}. \tag{5}$$

5 Experiments

In this section, we first describe the dataset of apparent personality analysis in the ECCV ChaLearn LAP 2016 challenge . Then, we give a detailed description about the implementation details of the proposed Deep Bimodal Regression

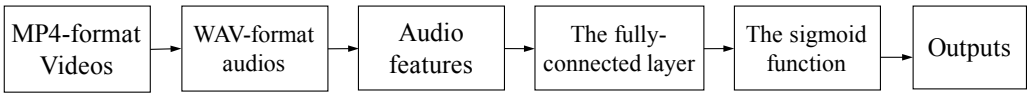


Figure 7. Pipeline of the handcrafted audio feature based regression for the audio modality. In consideration of its computational efficiency, the log filter bank (logfbank) features are used as the audio representations/features in the ChaLearn LAP challenge. Then, a linear regressor is trained on logfbank for predictions.

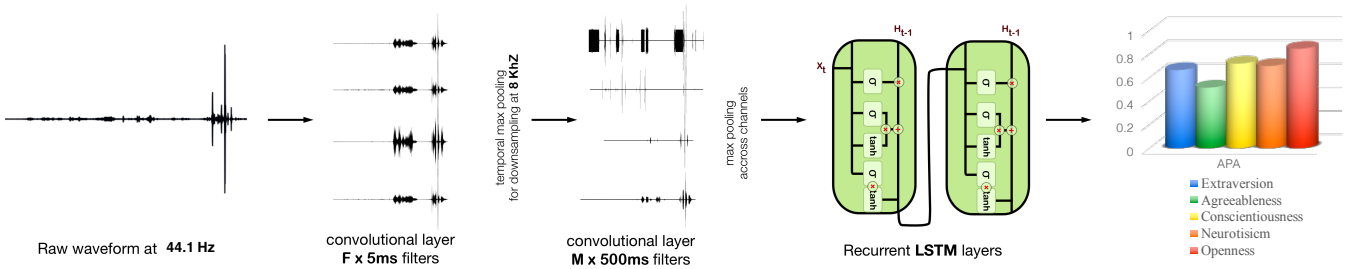


Figure 8. Illustration of the convolutional recurrent network [9] for personality traits predictions. The input are the raw waveform signals from the APA short video sequences. In the following, there are two stacked convolution layers: one uses $F = 40$ space time finite impulse filters with a 5ms window, and the other is a $M = 40$ one of a 500ms window. Then, the outputs of the last convolution layer are fed into the recurrent LSTM layers. The final layer is the ℓ_2 loss for minimizing the personality traits regression loss and training the whole network.

framework. Finally, we present and analyze the experimental results of the proposed framework on the competition dataset.

5.1 Datasets and Evaluation Criteria

The apparent personality analysis at the ECCV ChaLearn LAP 2016 competition is the first version for this track. In total, 10,000 videos are labeled to perform automatic apparent personality analysis. For each video sample, it has about fifteen seconds length. In addition, the RGB and audio information are provided, as well as continuous ground-truth values for each of the 5 Big Five traits annotated by Amazon Mechanical Turk workers.

The dataset is divided into three parts: the training set (6,000 videos), the validation set (2,000 videos) and the evaluation set (2,000 videos). During the Development phase, we train the visual and audio models of DBR on the training set, and verify its performance on the validation set. In the Final Evaluation phase, we use the optimal models in the Development phase to return the predicted regression scores on the final evaluation set.

For evaluation, given a video and the corresponding traits values, the accuracy is computed simply as one minus the absolute distance among the predicted values and the ground truth values. The mean accuracy among all the Big Five traits values is calculated as the principal quantitative measure:

$$\text{Mean accuracy} = \frac{1}{5N}\sum_{j=1}^{5}\sum_{i=1}^{N} 1 - \left|t_{i,j} - \hat{t}_{i,j}\right|, \qquad (6)$$

where $t_{i,\cdot}$ is the ground truth for the i-th video, $\hat{t}_{i,\cdot}$ represents the predicted personality traits value, and N is the number of predicted videos.

5.2 Implementation Details

In this section, we describe the proposed DBR framework's implementation details in the ChaLearn LAP challenge.

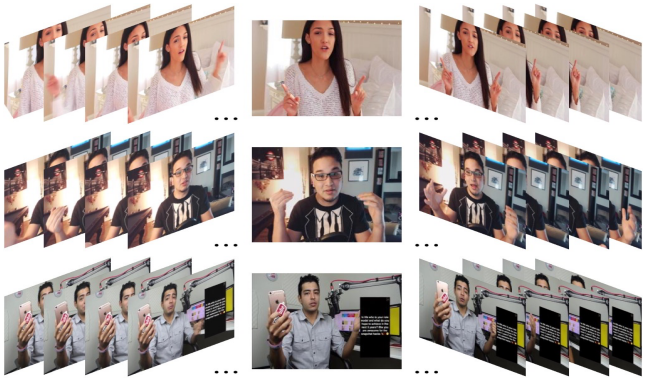


Figure 9. Examples of extracting images from videos. For each video, we extract about 100 images.

5.2.1 *Details of the Visual Modality*

As aforementioned, in the visual modality, we firstly extract about 100 images from each video by the rates of 6fps. After that, we resize these images into the 224×224 image resolution. In consequence, there are 560,393 images extracted from the training videos, 188,561 images from the validation ones, and 188,575 images from testing. Fig. 9 illustrates three examples of extracted image from videos.

In our experiments, the visual DAN models in the proposed DBR framework are implemented using the open-source library MatConvNet [51]. We adopt the pre-trained VGG-Face model [49] as the initialization of the convolution layers in our DANs. Beyond the DAN models, we also employ a popular deep convolutional network, i.e., Residual Network [7] pre-trained on ImageNet, as another regression network for boosting the visual regression performance. In the training stage, the learning rate is 10^{-3} for all the layers of all the used networks. The weight decay is 5×10^{-4}, and the momentum is 0.9. All the visual models (i.e., VGG-Face, ResNet, DAN, DAN$^+$ in Table 1) were fine-tuned in two epochs with batch size 32 on the data of the competition.

5.2.2 Details of the Audio Modality

In the audio modality, we firstly extract the audio features from the original videos, and then learn a linear regressor based on these audio features. In the APA competition, the open-source library FFmpeg[1] is employed for extracting audios from the original videos. Regarding the parameters of FFmpeg, we choose two channels for the WAV format audio outputs, 44,100Hz for the sampling frequency, and $320k$bps for the audio quality. The average memory cost of each audio file is about 2.7MB in disk. Based on the extracted audios, we use the Python open source library to extract the MFCC and logfbank features.[2]

For the regression model training, we use the Torch platform.[3] Thus, GPUs can be used for accelerating the model training. The linear regressor is composed of a fully-connected layer and a sigmoid function layer to regress the values of the Big Five traits' ground truth (in the range of $[0, 1]$). For optimization, the traditional stochastic gradient descent (SGD) method is used, and the momentum is set as 0.9. The batch-size of the audio features is 128. The learning rate of SGD is 8.3×10^{-4}. The weight decay is 6.5, and the learning rate decay is 1.01×10^{-6}. For the hyper-parameters here, we selected the optimal values by the random search method proposed in [52].

All the experiments above were conducted on a Ubuntu 14.04 Server with 512GB memory and K80 Nvidia GPUs support. For the computational time, we took about 2 days on training networks in both the visual and audio modalities. During inference, the personality traits prediction time is about 0.39s per video, which makes real time prediction possible.

5.3 Results of the ChaLearn LAP Challenge

This ChaLearn LAP apparent personality traits challenge lasted over two months and attracted 86 participants who are grouped in several teams. In this section, we first present the experimental results of the Development phase and analyze our proposed DBR framework. Then, we show the Final Evaluation results of this apparent personality analysis competition.

5.3.1 Development

In Table 1, we present the main results of both the visual and audio modality in the Development phase. Note that, during the competition, we employed the ℓ_1 loss function for training all the visual deep convolutional neural networks, including VGG-Face, ResNet, DAN and DAN$^+$.

For the visual modality, we also fine-tune the available VGG-Face [49] model on the competition data for comparison. As shown in Table 1, the regression accuracy of DAN (0.9100) is better than VGG-Face (0.9072) with the traditional VGG-16 [47] architecture, and even better than Residual Networks (0.9080). Meanwhile, because DAN has no traditional fully connected (FC) layers, the number of the DAN parameters is only 14.71M, which is much less than 134.28M of VGG-16 and 58.31M of ResNet. It will bring storage efficiency. Meanwhile, without the parameter redundant FC layers, the inference speed of DAN and DAN$^+$ is faster than both VGG-based models and ResNet.

In addition, from the results of the first and second epoch, we can find the regression accuracy becomes lower when the training epochs increase, which might be overfitting. Thus, we stop training after the second epoch. Then, we average the predicted scores of these two epochs as the epoch fusion.[4] The performance of the epoch fusion is better than that of single epoch. Therefore, the averaged regression scores of the epoch fusion are the predictions of the visual modality, which is the early fusion in DBR.

For the audio modality, because the handcrafted audio features of this competition (i.e., MFCC and logfbank) are in large scale, considering the computational efficiency, we simply train a linear regressor by Torch on GPUs in the competition. In order to choose the optimal audio representation of the audio modality, we randomly split the training set (6,000) into two parts: one has 5,000 samples, and the other has 1,000 samples. On the MFCC and logfbank features, we separately learn two linear regressors on the 5,000 samples. Then the rest 1,000 samples are used to validate the performance of different audio features. Fig. 10(a) and Fig. 10(b) shows the learning curves of MFCC and logfbank, respectively. The vertical axis is the regression error. It can be seen from these figures that, logfbank could outperform MFCC by 0.75%. Therefore, the logfbank features are chosen as the optimal audio representation in the competition. Additionally, beyond the competition, we also train a deep audio network mentioned in Sec. 4.2.2 on the ChaLearn LAP competition dataset, which obtains 0.8950 mean accuracy on the Big Five traits. The result is also reported in Table 1. Its regression accuracy is 0.5% higher than the accuracy of logfdbank.

After the model training of both two modalities, we obtain three deep visual regression networks (i.e., ResNet, DAN and DAN$^+$) and one audio regression model (a linear regressor). As described in Sec. 4.3, we average all the four predicted Big Five traits scores, and get the final APA predictions. Finally, we can get 0.9141 mean accuracy in the Development phase, which is reported in Table 1.

Furthermore, in Table 1, we report the regression accuracy of different settings of modality fusion. The first setting (i.e., "ResNet(ℓ_1)+DAN(ℓ_1)+DAN$^+$(ℓ_1)+linear regressor") was used in the competition. Beyond that, we also present results where fusion takes place with networks trained using the Tukey's biweight loss function as well as results where [9] is used for fusion. As the results shown, the optimal modality fusion setting could be the three visual networks trained using the Tukey's biweight loss plus two audio models (i.e., "ResNet(T)+DAN(T)+DAN$^+$(T)+linear regressor+ [9]" in Table 1).

5.3.2 Final Evaluation

In the Final Evaluation phase, we directly employ the optimal models in the Development phase to predict the Big Five traits values on the testing set. The final challenge

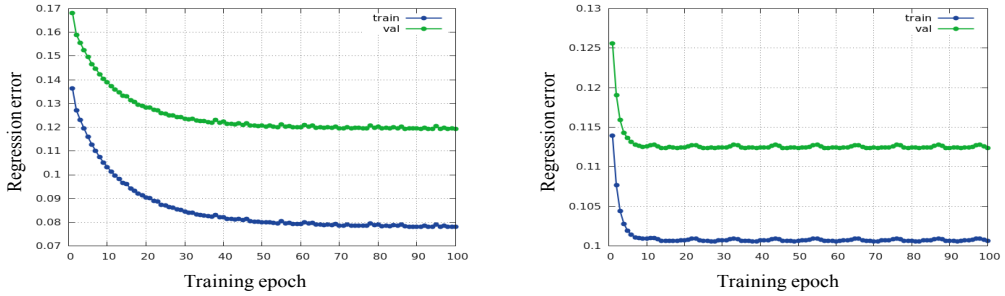
1. http://ffmpeg.org/
2. https://github.com/jameslyons/python_speech_features
3. http://torch.ch/

4. The epoch fusion method is a traditional ensemble method [53] in neural networks, which is also known as doing ensemble with different checkpoints of a single model. Concretely, this method usually takes different checkpoints of a single network over time (e.g., after every epoch) and uses those to form an ensemble.

Table 1
Regression mean accuracy comparisons in the Development phase. Moreover, the number of parameters, feature dimensionality, and inference time per video of different models are also listed. DAN and DAN$^+$ are our proposed models in the ChaLearn LAP challenge. For modality fusion, we also report the regression accuracy when employing the robust Tukey's biweight loss function proposed in Sec. 4.1.4. In this table, for example, "DAN (ℓ_1)" presents DAN training with the ℓ_1 loss, and "DAN (T)" presents it training with the Tukey's biweight loss.

Modality	Model	♯ Para.	Dim.	Inference Speed	*Epoch 1*	*Epoch 2*	*Epoch Fusion*
Visual	VGG-Face (ℓ_1)	134.28M	4,096	400 ms	0.9065	0.9060	0.9072
	ResNet (ℓ_1)	58.31M	512	450 ms	0.9072	0.9063	0.9080
	DAN (ℓ_1)	14.71M	1,024	389 ms	0.9082	0.9080	0.9100
	DAN$^+$ (ℓ_1)	14.72M	2,048	390 ms	0.9100	0.9103	0.9111
Audio	Linear regressor	0.40M	79,534	–	0.8900	–	0.8900
	[9]	1.50M	8,350	–	0.8950	–	0.8950
Modality fusion	ResNet(ℓ_1)+DAN(ℓ_1)+DAN$^+$(ℓ_1)+linear regressor[1]	–	–	–	–	–	0.9141
	ResNet(ℓ_1)+DAN(ℓ_1)+DAN$^+$(ℓ_1)+ [9]	–	–	–	–	–	0.9186
	ResNet(T)+DAN(T)+DAN$^+$(T)+ [9]	–	–	–	–	–	0.9201
	ResNet(T)+DAN(T)+DAN$^+$(T)+linear regressor+ [9]	–	–	–	–	–	0.9212

[1]This fusion strategy was used in the competition of the ECCV ChaLearn LAP 2016 challenge.



(a) Learning curves of MFCC. (b) Learning curves of logfbank.

Figure 10. Learning curves of two different audio features, i.e., MFCC and logfbank.

results are obtained by ResNet, DAN, DAN$^+$ built on the visual modality and the linear regressor built on the audio modality (cf. Table 1). As shown in Table 2, our final result (0.9130) ranked the first place, which outperformed the other participants. Moreover, for the regression accuracy of each Big Five trait value, our proposed DBR framework achieved the best result in four traits, i.e., *Agreeableness*, *Conscientiouness*, *Neurotisicm*, and *Openness*.

Since we just use the simple average method to do the late fusion, for further improving regression performance of the proposed method, advanced ensemble methods, e.g., stacking, can be used to learn the appropriate weights for the late fusion. Additionally, the ensemble of multiple audio models, e.g., handcrafted audio features with regressors and deep learning based audio networks, can be also applied into our DBR framework to achieve better personality traits prediction performance.

5.4 Insight Experiments

In the following, we present various insight experiments. These experiments are both quantitative and qualitative, and give a deeper understanding of the method.

5.4.1 Comparisons of Regression Loss Functions

Table 3 reports the regression mean accuracy of DAN with different loss functions on the ChaLearn LAP competition validation set. Furthermore, we conduct the pairwise t-test on Tukey's biweight with ℓ_1 (marked by "$\circ$" in Table 3) and Tukey's biweight with ℓ_2 (marked by "$\bullet$"), respectively. As shown in Table 3, the regression results of DAN with the ℓ_1 loss function are the worst among them. While, ℓ_1 is also the loss function used in the competition for DAN training. For the other loss functions, DAN with ℓ_2 loss outperforms DAN with ℓ_1. Furthermore, DAN with the Tukey's biweight loss function achieves the best regression performance for predicting personality traits, which validates the robustness and generalization ability of such a non-convex loss function on the human-labeled data. After equipping with robust regression loss, e.g., the Tukey's biweight loss function, our DBR model will further improve the apparent personality traits prediction accuracy (cf. the results of modality fusion shown in Table 1).

5.4.2 Visualization of DAN Models

In order to further justify the effectiveness of the deep visual regression networks of DBR, we visualize the feature maps of these networks (i.e., ResNet, traditional VGG models, DAN with ℓ_1 loss, DAN with ℓ_2 loss, DAN with Tukey's biweight loss and DAN$^+$) in Fig. 11. In that figure, we randomly sample 12 extracted images from different APA videos, and show the Pool_5 feature maps. As shown in those figures, the strongest responses in the corresponding feature maps of these deep networks are quite different from each other, especially the ones of ResNet vs. the ones of DAN/DAN$^+$. That is to say, different models are focusing on different regions in the input images to make their own predictions.

Concretely, for almost all the cases, ResNet could pay its attention on the human beings. While DAN/DAN$^+$ will fo-

Table 2
Comparison of performances of the proposed DBR framework with that of the top ten teams in the Final Evaluation phase. **"NJU-LAMDA" is ours.**

Rank	Team name	*Mean Accuracy*	*Extraversion*	*Agreeableness*	*Conscientiouness*	*Neurotisicm*	*Openness*
1	**NJU-LAMDA** [11]	**0.9130**	0.9133	**0.9126**	**0.9166**	**0.9100**	**0.9123**
2	evolgen[1] [54]	0.9121	0.9150	0.9119	0.9119	0.9099	0.9117
3	DCC[2] [55]	0.9109	0.9107	0.9102	0.9138	0.9089	0.9111
4	ucas	0.9098	0.9129	0.9091	0.9107	0.9064	0.9099
5	BU-NKU	0.9094	**0.9161**	0.9070	0.9133	0.9021	0.9084
6	pandora	0.9063	0.9064	0.9075	0.9049	0.9044	0.9081
7	Pilab	0.8936	0.8921	0.9000	0.8867	0.8914	0.8977
8	Kaizoku	0.8826	0.8751	0.8916	0.8803	0.8767	0.8890
9	ITU-SiMiT	0.8814	0.8780	0.8937	0.8746	0.8775	0.8834
10	sp	0.8758	0.8769	0.8842	0.8727	0.8752	0.8702

[1]*Department of Computer Science and Engineering, Indian Institute of Technology Madras, India.*
[2]*Donders Institute for Brain, Radboud University, The Netherlands.*

Table 3
Comparison of DAN with different regression loss functions, i.e., the ℓ_1, ℓ_2 and Tukey's biweight loss functions. Note that, "$\circ$" presents Tukey's biweight is significantly different from ℓ_1, and "$\bullet$" presents Tukey's biweight is siginificantly different from ℓ_2. The significance level is 85%.

Loss Function	*Epoch 1*	*Epoch 2*	*Epoch Fusion*
ℓ_1	0.9082	0.9080	0.9100
ℓ_2	0.9098	0.9083	0.9108
Tukey's biweight	**0.9106** $\circ\bullet$	**0.9087** $\circ$	**0.9115** $\circ\bullet$

cus on not only the human, but also the environments where human beings are on these videos. These environments in videos might be one kind of side information for personality traits prediction in some sense, because the environments where human beings stay might also reflect their personal characteristics. Particularly, for the last example, there is just a person without any background in that video. Our DAN models could accurately pay attentions on her. However, ResNet failed this time. In addition, the traditional VGG models with fully connected layers cannot focus on the main human beings in these short video sequences.

For our DAN models with different loss functions, it is obvious to find DAN with ℓ_1 loss and DAN$^+$ could extract complementary information for images (especially for the 1^{st}, 8^{th}, 9^{th}, 10^{th} and 11^{th} sampled examples in Fig. 11) in apparent personality analysis, which can also give a qualitative explanation about the effectiveness of the late fusion in DBR.

5.4.3 DAN Representation Embedding

Fig. 12 shows a t-Distributed Stochastic Neighbor Embedding (t-SNE) [56] of the concatenation layer of the visual models trained on the ChaLearn LAP competition dataset. The feature vectors of different deep networks are preprocessed using PCA to a dimensionality of 50. The perplexity parameters of t-SNE is set to 30. These dots in Fig. 12 correspond to their feature embeddings on the validation data. Here we use one of the 5 Big Five traits (i.e., *Extraversion*) as an example for illustration. The rest 4 Big Five traits have the identical trends. The color of these dots represents their corresponding real values of *Extraversion*. The warm color stands for a high value, and the cold color does for a low value. These values are continuous in the range of $[0, 1]$.

Since t-SNE is a powerful tool for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map. We here employ t-SNE to visualize these embedding results in a 2-d map for better recognizing the discriminative abilities of the corresponding feature vectors. From these figures, it is obvious to find there are two apparent clusters (i.e., one red cluster and one blue cluster) in the visualization of DANs (i.e., DAN with ℓ_1, DAN with ℓ_2, DAN with Tukey's biweight and DAN$^+$). However, for ResNet, it just has an apparent red cluster, and the blue dots are scatted. There is even no obvious pattern in the visualization of traditional VGG model.

These observations could prove the feature vectors' discriminative abilities of the proposed DANs are significantly better than the abilities of ResNet and VGG. It is consistent with the reported regression accuracy in Table 1, which validates the effectiveness of the proposed DANs from the qualitative perspective.

6 Conclusions

Automatic apparent personality analysis from videos is an important and challenging problem in computer vision and multimedia research. In order to exploit and capture important cues from both the visual and audio modality, this paper has proposed the Deep Bimodal Regression (DBR) framework. Also, in DBR, we modified the traditional CNNs as Descriptor Aggregation Networks (DANs) for improving the visual regression performance. Finally, we utilized the proposed DBR framework and DANs for the track of apparent personality analysis at the ChaLearn LAP challenge in association with ECCV 2016, and achieved the 1^{st} place in the Final Evaluation phase. Additionally, we also investigated the effectiveness and robustness of non-convex loss functions for personality traits regression.

In the future, we will introduce advanced ensemble methods into our framework and incorporating other more discriminative deep audio representations for apparent personality analysis.

References

[1] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.
[2] B. Engler, *Personality theories*. Nelson Education, 2013.
[3] D. Briley and E. Tucker-Drob, "Genetic and environmental continuity in personality development: A meta-analysis," *Psychological Bulletin*, vol. 140, no. 5, pp. 1303–1324, 2014.

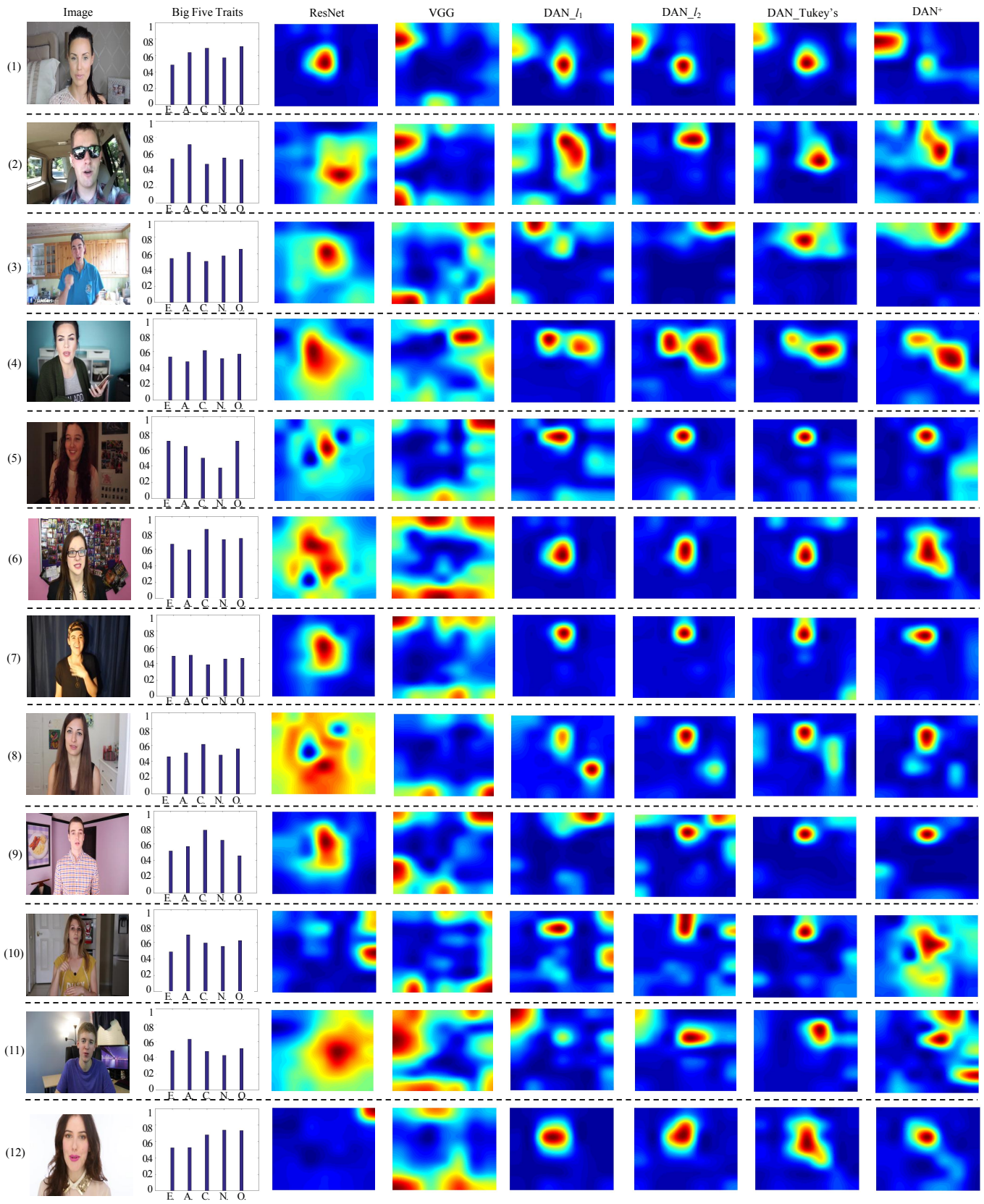
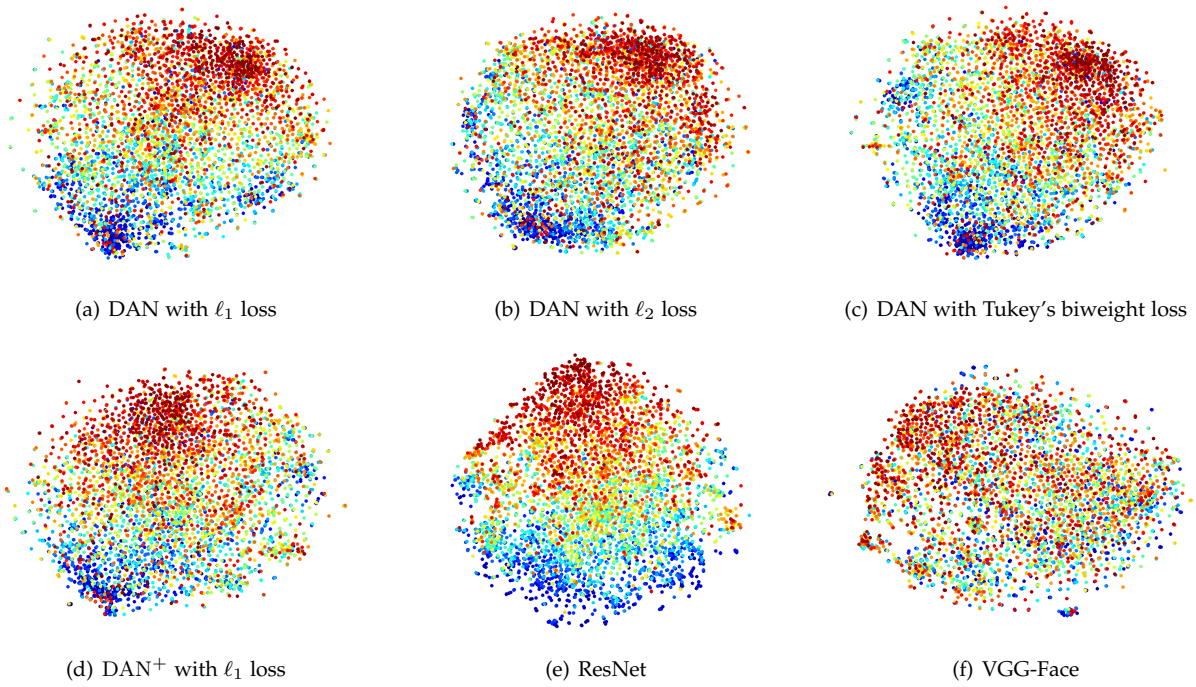


Figure 11. Feature maps of 12 sampled images in the visual modality of DBR. The first column shows the images, and the second column presents their corresponding Big Five traits values (In each subfigure, the horizontal axis shows the Big Five traits, and the vertical axis reports their corresponding predicted values.). The rest columns show the feature maps of ResNet, the VGG model, "DAN with ℓ_1 loss", "DAN with ℓ_2 loss", "DAN with Tukey's biweight loss" and "DAN$^+$", respectively. The ℓ_1 loss function is used as the loss for training DAN$^+$, ResNet and VGG. For each feature map, we sum the responses values of all the channels in the final pooling layer for each deep network. Best viewed in color.



(a) DAN with ℓ_1 loss (b) DAN with ℓ_2 loss (c) DAN with Tukey's biweight loss

(d) DAN$^+$ with ℓ_1 loss (e) ResNet (f) VGG-Face

Figure 12. Representation embedding visualization by t-SNE of different deep models on the ChaLearn LAP competition data [4]. Best in color.

[4] V. Ponce-López, B. Chen, M. Oliu, C. Cornearu, 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 *Proceedings of the European Conference on Computer Vision 2016, in press*, 2016.

[5] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2868–2881, 2017.

[6] X.-S. Wei, C.-W. Xie, and J. Wu, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition," *arXiv preprint arXiv:1605.06878*, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *International Journal of Computer Vision*, vol. 19, no. 1, pp. 57–91, 1996.

[9] G. Trigeorgis, F. Ringeval, R. Bruekner, E. March, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 1–5.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep bimodal regression for apparent personality analysis," in *Proceedings of the European Conference on Computer Vision 2016 ChaLearn Looking at People Workshop, Part III, LNCS 9915*, G. Hua and H. Jégou, Eds. Springer, Switzerland, 2016, pp. 311–324.

[12] F. Mairesse, M. Walker, M. Mehl, and R. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligent Research*, vol. 30, pp. 457–500, 2007.

[13] L. Batrinca, B. Lepri, N. Mana, and F. Pianesi, "Multimodal recognition of personality traits in human-computer collaborative tasks," in *Proceedings of International Conference on Multimodal Interaction*, 2012, pp. 39–46.

[14] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion - a systematic study," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 443–455, 2012.

[15] J. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of International Conference on Multimodal Interaction*, 2012, pp. 53–56.

[16] V. Ponce-López, S. Escalera, M. Pérez, O. Janés, and X. Baró, "Non-verbal communication analysis in victim-offender mediations," *Pattern Recognition Letters*, vol. 67, no. P1, pp. 19–27, 2015.

[17] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the International Conference on Multimodal Interfaces*, 2004, pp. 205–211.

[18] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[20] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Transactions on Affective Computing*, DOI 10.1109/TAFFC.2016.2622690.

[21] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1913–1921.

[22] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

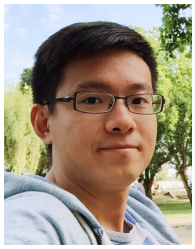
[23] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[24] B. Schuller, "Recognizing affect from linguistic information in 3D continuous space," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 192–205, 2011.

[25] Y. Vaizman, B. McFee, and G. Lanckriet, "Codebook-based audio feature representation for music information retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1483–1493, 2014.

[26] L. Su, C.-C. M. Yeh, J.-Y. Liu, J.-C. Wang, and Y.-H. Yang, "A systematic evaluation of the bag-of-frames representation for music

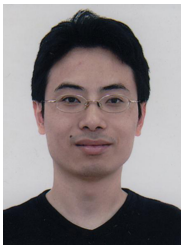
information retrieval," *IEEE Transactions Multimedia*, vol. 16, no. 5, pp. 1188–1200, 2014.
[27] M. Slaney and M. Covell, "Facesync: A linear operator for measuring synchronization of video facial images and audio tracks," in *Advances in Neural Information Processing Systems*, 2001, pp. 1–7.
[28] S. Kshirsagar and N. Magnenat-Thalmann, "Lip synchronization using linear predictive analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2000, pp. 1–4.
[29] M. Bhargava and R. Rose, "Architectures for deep neural network based acoustic models defined over windowed speech waveforms," in *Proceedings of INTERSPEECH*, 2015, pp. 6–10.
[30] T. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4580–4584.
[31] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proceedings of INTERSPEECH*, 2015, pp. 11–15.
[32] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 31, no. 1, pp. 39–58, 2009.
[33] B. Xiong and K. Grauman, "Detecting snap points in egocentric video with a web photo prior," in *Proceedings of the European Conference on Computer Vision 2014, Part V, LNCS 8693*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, Switzerland, 2014, pp. 282–298.
[34] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2629–2638.
[35] M. R. Amer, P. Lei, and S. Todorovic, "Hirf: Hierarchical random field for collective activity recognition in videos," in *Proceedings of the European Conference on Computer Vision 2014, Part VI, LNCS 8694*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, Switzerland, 2014, pp. 572–585.
[36] M. Hasan and A. K. Roy-Chowdhury, "Continuous learning of human activity models using deep nets," in *Proceedings of the European Conference on Computer Vision 2014, Part III, LNCS 8691*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, Switzerland, 2014, pp. 705–720.
[37] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1913–1921.
[38] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in LSTMs for activity detection and early detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1942–1950.
[39] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.
[40] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu, "Inferring forces and learning human utilities from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3823–3833.
[41] X. Yan, H. Chang, S. Shan, and X. Chen, "Modeling video dynamics with deep dynencoder," in *Proceedings of the European Conference on Computer Vision 2014, Part IV, LNCS 8692*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, Switzerland, 2014, pp. 215–230.
[42] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *Proceedings of the European Conference on Computer Vision 2014, Part VI, LNCS 8694*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer, Switzerland, 2014, pp. 800–813.
[43] W. Michel, Y. Shoda, and R. E. Smith, "Introduction to personality: Toward an integration," 2004.
[44] D. Thomas, T. Daniel, and C. John, "Confirming first impressions in the employment interview: A field study of interviewer behavior," *Journal of Applied Psychology*, vol. 79, no. 5, pp. 659–665, 1994.
[45] M. Barrick and M. Mount, "The big five personality dimensions and job performance: A meta-analysis," *Personnel Psychology*, vol. 44, pp. 1–26, 1991.
[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015, pp. 1–14.
[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[49] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of British Machine Vision Conference*, 2015, pp. 1–12.
[50] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[51] A. Vedaldi and K. Lenc, "MatConvNet – Convolutional Neural Networks for MATLAB," in *Proceeding of ACM International Conference on Multimedia*, 2015, pp. 689–692, http://www.vlfeat.org/matconvnet/.
[52] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
[53] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: Chapman & HallCRC (ISBN 978-1-439-830031), 2012.
[54] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *Proceedings of the European Conference on Computer Vision 2016 ChaLearn Looking at People Workshop, Part III, LNCS 9915*, G. Hua and H. Jégou, Eds. Springer, Switzerland, 2016, pp. 337–348.
[55] Y. Güclütürk, M. A. J. v. G. U. Güclü, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *Proceedings of the European Conference on Computer Vision 2016 ChaLearn Looking at People Workshop, Part III, LNCS 9915*, G. Hua and H. Jégou, Eds. Springer, Switzerland, 2016, pp. 349–358.
[56] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[57] C. G. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.
[58] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2830–2838.



Xiu-Shen Wei received the BS degree in Computer Science and Technology in 2012. He is currently a PhD candidate in the Department of Computer Science and Technology at Nanjing University, China. He achieved the first place in the Apparent Personality Analysis competition (in association with ECCV 2016) and the first runner-up in the Cultural Event Recognition competition (in association with ICCV 2015) as the team director. He also received the Presidential Special Scholarship (the highest honor for Ph.D. students) in Nanjing University. His research interests are computer vision and machine learning.

Chen-Lin Zhang received his BS degree in the Department of Computer Science and Technology from Nanjing University, China, in 2016. He is currently working toward the PhD degree in the Department of Computer Science and Technology, Nanjing University, China. His research interests are computer vision and machine learning.

Hao Zhang received his BS degree from Nanjing University, China, in 2016. He is currently working toward the MS degree in the Department of Computer Science and Technology, Nanjing University, China. His research interests are computer vision and machine learning.



Jianxin Wu (M'09) received his BS and MS degrees in computer science from Nanjing University, and his PhD degree in computer science from the Georgia Institute of Technology. He is currently a professor in the Department of Computer Science and Technology at Nanjing University, China, and is associated with the National Key Laboratory for Novel Software Technology, China. He was an assistant professor in the Nanyang Technological University, Singapore, and has served as an area chair for CVPR, ICCV, senior PC member for AAAI and IJCAI, and an associate editor for the Pattern Recognition Journal. His research interests are computer vision and machine learning. He is a member of the IEEE.