# Deep Bimodal Regression for Apparent Personality Analysis

Chen-Lin Zhang, Hao Zhang, Xiu-Shen Wei, and Jianxin Wu(✉)

National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, China
{zhangcl,zhangh,weixs,wujx}@lamda.nju.edu.cn

**Abstract.** Apparent personality analysis from short video sequences is a challenging problem in computer vision and multimedia research. In order to capture rich information from both the visual and audio modality of videos, we propose the 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 the 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 teams.

**Keywords:** Apparent personality analysis · Deep regression learning · Bimodal learning · Convolutional neural networks

## 1 Introduction

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 [14,19,21], activity recognition [1,4], gesture and pose recognition [8,11,22] and many more [13,15,20,23].

Moreover, apparent personality analysis (APA) is an important problem of human-centered video analysis. The goal of APA is to develop algorithms for recognizing personality traits of users in short video sequences. Personality traits are usually decomposed into components called the *Big Five Traits*, including

This work was supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization. X.-S. Wei is the team director of the APA competition, and J. Wu is the corresponding author.


G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part III, LNCS 9915, pp. 311–324, 2016.
DOI: 10.1007/978-3-319-49409-8_25

*openness to experience*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*. Effective apparent personality analysis 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.

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 with 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 analysis.

In the visual modality, we firstly extract frames from each original video. Then, deep convolutional neural networks are adopted to learn deep regressors for predicting the Big Five Traits values. Inspired by our previous work [17,18], 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 convolutional layer are both averaged and max pooled into 512-d feature vectors. After that, the standard $\ell_2$-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 DAN$^+$. Beyond DAN and DAN$^+$, Residual Networks [5] is also utilized in our visual modality of DBR. As discussed in Sect. 4.3, the epoch fusion is used as the early fusion to boost the visual regression performance.

For the audio modality, the log filter bank (logfbank) [3] features are extracted from the original audio of each video. Based on the logfbank features, we train the linear regressor to obtain the Big Five Traits values. Finally, the two modalities are lately fused by averaging the scores of these deep visual models and the audio model. Thus, the final predicted Big Five Traits values are returned.

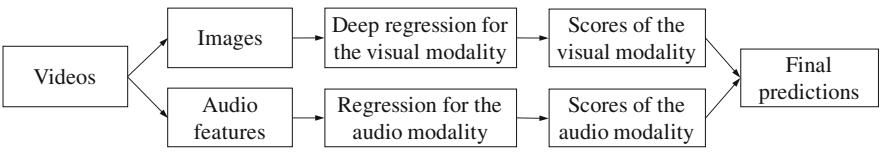



**Fig. 1.** Framework of the proposed Deep Bimodal Regression method. In DBR, the original videos are treated as having two natural modalities, i.e., the visual modality for images and the audio modality for speeches. After learning the (deep) regressors on these two modalites, the final predicted personality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

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 association with ECCV 2016. In the challenge, we are given a large newly collected video data set, which contains 10,000 videos of about 15 s each collected from YouTube, and annotated with the Big Five Traits by Amazon Mechanical Turk workers. In the Final Evaluation phase, our DBR framework achieved the best regression accuracy (0.9130 mean accuracy), which ranked *the first place* in this challenge.

## 2 Related Work

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

### 2.1 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) [7] 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 convolutional 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 convolutional layer, together with appropriately chosen pooling schemes, endows the CNN with some invariance properties (e.g., translation invariance).

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 by end-to-end training.

### 2.2 Audio Representations

In the past few years, many 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) [3], Linear Prediction Cepstral Coefficient (LPCC) [9] and Bark Frequency Cepstral Coefficient (BFCC) [9].

Particularly, the Mel Frequency Cepstral Coefficients (MFCC) [3] features have been widely used in the speech recognition community. 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 four steps. During the four steps, the log filter bank (logfbank) features can be also obtained.

In the proposed DBR framework, we extract the MFCC and logfbank features from the audios of each original human-centered video for APA. In our experiments, the results of logfbank are slightly better than the ones of MFCC. Thus, the logfbank features are used as the audio representations in DBR.

### 2.3 Apparent Personality Analysis

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.

A similar task to personality analysis in computer vision is the emotion analysis tasks, e.g., [2,6]. Emotion analysis can be regarded as a multiple class classification problem, where usually four emotions (*sadness*, *happiness*, *anger* and *neutral state*) 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 with 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.

## 3 The Proposed DBR Framework

In this section, we will introduce the proposed Deep Bimodal Regression (DBR) framework for the apparent personality analysis task. 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.

### 3.1 Deep Regression for the Visual Modality

The deep regression part contains three subparts: image extraction, deep regression network training and regression score prediction.

**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. 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 downsample these images/frames to 100 images per video. That is to say, in each second, we extract 6 images from a video. After that, the extracted images/frames are labeled with the same personality traits values as the ones of their corresponding video. In consequence, based on the images, we can train the deep regressors by CNNs for APA.

**Deep Regression Network Training.** In the visual modality of DBR, the main deep CNN models are modified based on our previous work [17,18], which are called Descriptor Aggregation Networks (DANs). What distinguishes DAN from the traditional CNN is: the fully connected layers are discarded, and replaced by both average- and max-pooling following the last convolutional layers ($\text{Pool}_5$). Meanwhile, each pooling operation is followed by the standard $\ell_2$-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 convolutional 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. The architecture of DAN is illustrated in Fig. 2.

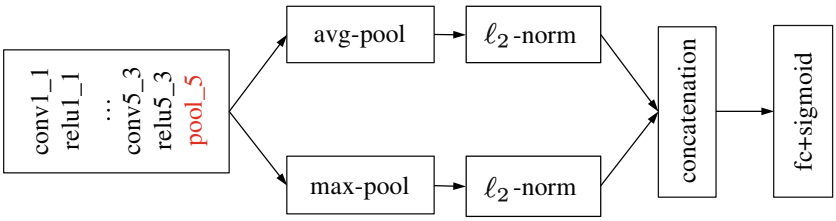



**Fig. 2.** Architecture of the Descriptor Aggregation Network (DAN) model. Note that we removed the fully connected layers. The deep descriptors of the last convolutional layer ($\text{Pool}_5$) are firstly aggregated by both average- and max-pooling, and then concatenated into the final image representation for regression.

Because DAN has no 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 [18]. In the experiments of the proposed DBR framework, we adopt the pre-trained VGG-Face model [10] as the initialization of the convolutional layers in our DANs.

For further improving the regression performance of DAN, the ensemble of multiple layers is employed. Specifically, the deep convolutional descriptors of $\text{ReLU}_{5\_2}$ are also incorporated in the similar aforementioned aggregation approach, which is shown in Fig. 3. Thus, the final image feature is a 2048-d vector. We call this end-to-end deep regression network as "DAN$^+$".

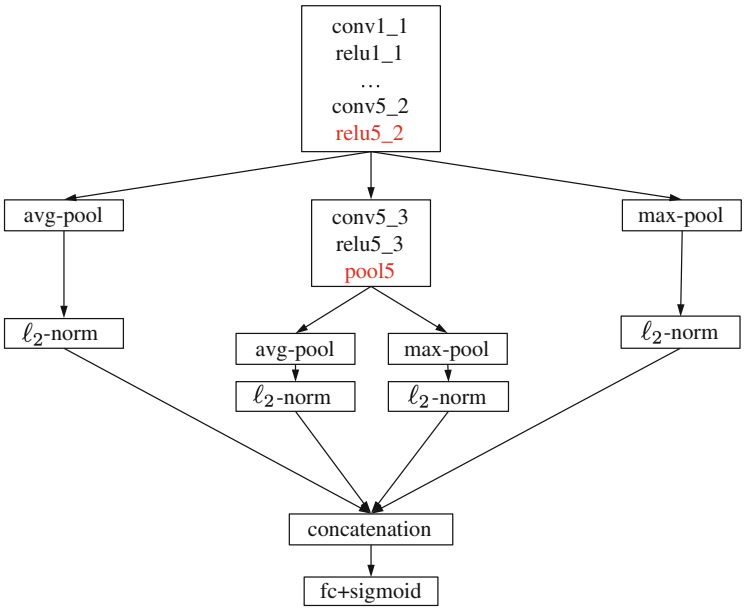



**Fig. 3.** Architecture of the DAN$^+$ model. In DAN$^+$, not only the deep descriptors of the last convolutional layer ($\text{Pool}_5$) are used, but the ones of $\text{ReLU}_{5\_2}$ are also aggregated. Finally, the feature vectors of multiple layers are concatenated as the final image representation for regression.

**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.

### 3.2 Regression for the Audio Modality

As aforementioned, in the audio modality, we choose the log filter bank (logfbank) features as the audio representations. The logfbank features can be extracted directly from the original audios from videos. After that, we use a model composed of a fully-connected layer followed by a sigmoid function layer to train the audio regressor. The $\ell_2$ distance is used as the loss function to calculate the regression loss. The whole pipeline of the audio modality can be seen in Fig. 4.

### 3.3 Modality Ensemble

After the training of both the visual and the audio modalities, modality ensemble is used as the late fusion approach for getting the final regression scores.

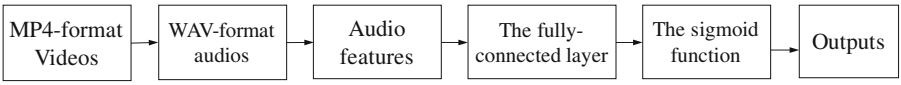



**Fig. 4.** Pipeline of the regresssion for the audio modality. The log filter bank features are used as the audio representations/features. Based on the audio features, a linear regressor is trained for predictions.

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:

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

## 4 Experiments

In this section, we first describe the dataset of apparent personality analysis at the ECCV ChaLearn LAP 2016 competition. Then, we give a detailed description about the implementation details of the proposed DBR framework. Finally, we present and analyze the experimental results of the proposed framework on the competition dataset.

### 4.1 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:

$$\text{Mean accuracy} = \frac{1}{5N}\sum_{j=1}^{5}\sum_{i=1}^{N} 1 - \left|\text{ground\_truth}_{i,j} - \text{predicted\_value}_{i,j}\right|, \quad (2)$$

where $N$ is the number of predicted videos.

### 4.2 Implementation Details

In this section, we describe the implementation details of the proposed DBR framework on the APA competition dataset.

**Details of the Visual Modality.** As aforementioned, in the visual modality, we firstly extract about 100 images from each video. Specifically, for most videos, 92 images are extracted (about 6.1fps). After that, we resize these images into the $224 \times 224$ 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. Figure 5 illustrates three examples of extracting image from videos.

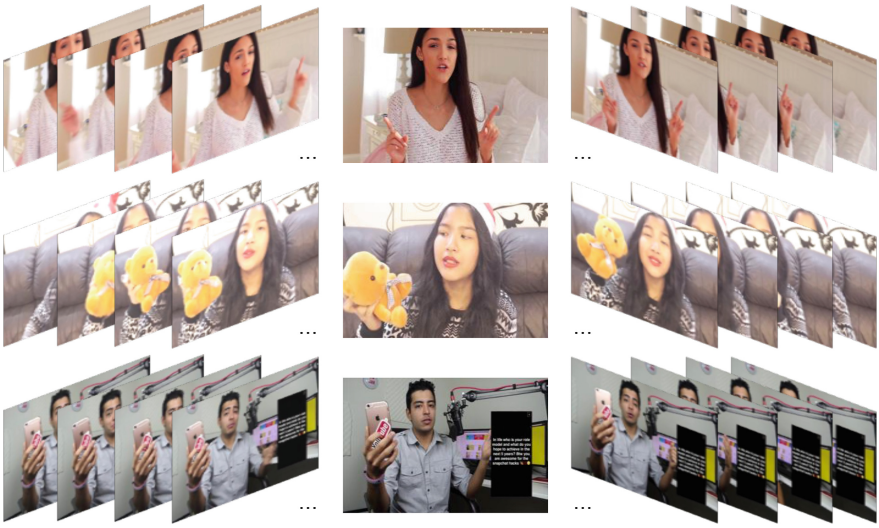


**Fig. 5.** Examples of extracting images from videos. For each video, we extract about 100 images.

In our experiments, the visual DAN models in the proposed DBR framework are implemented using the open-source library MatConvNet [16]. Beyond the DAN models, we also employ a popular deep convolutional network, i.e., Residual Network [5], as another regression network for boosting the visual regression performance. In the training stage, the learning rate is $10^{-3}$. The weight decay is $5 \times 10^{-4}$, and the momentum is 0.9 for all the visual models.

**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, 44,100 Hz for the sampling frequency, and 320 kbps for the audio quality. The average memory cost of each audio file is about 2.7 MB in Disk. Figure 6 presents the wave forms of one sampled audio from its corresponding video. Based on the extracted audios, we use the Python open source library to extract the MFCC and logfbank features.[2]

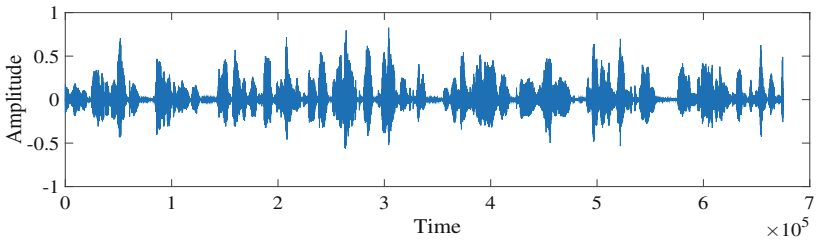



**Fig. 6.** The wave forms of a sampled audio. The horizontal axis stands for the time. Because we set the sampling frequency as 44,100 Hz, the unit of the horizontal axis is 1/44100 s. The vertical axis is the amplitude.

For the regression model training, we use the Torch platform.[3] Thus, the 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 $8.3 \times 10^{-4}$. The weight decay is 6.5, and the learning rate decay is $1.01 \times 10^{-6}$.

All the experiments above were conducted on a Ubuntu 14.04 Server with 512 GB memory and K80 Nvidia GPUs support.

### 4.3 Experimental Results

In this section, we first present the experimental results of the Development phase and analyze our proposed DBR framework. Then, we present the qualitative evaluation of the deep visual regression networks in DBR. Finally, we show the Final Evaluation results of this apparent personality analysis competition.

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

**Development.** In Table 1, we present the main results of both the visual and audio modality in the Development phase.

For the visual modality, we also fine-tune the available VGG-Face [10] 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 [12] architecture, and even better than Residual Networks (0.9080). Meanwhile, because DAN has no traditional fully connected 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.

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. The performance of the epoch fusion is better than the one of each epoch. Therefore, the averaged regression scores of the epoch fusion are the predictions of the visual modality, which is the early fusion in DBR.

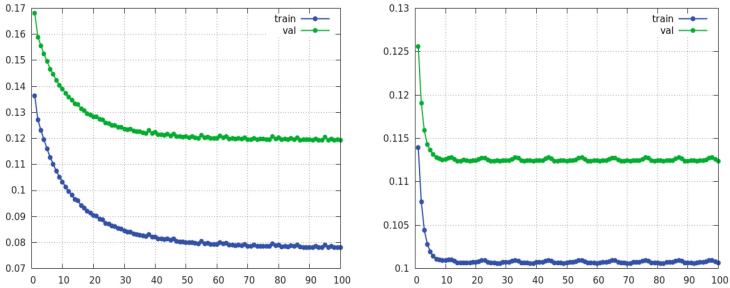
**Table 1.** Regression mean accuracy comparisons in the Development phase. Moreover, the number of parameters, feature dimensionality, and inference time per video of different models are also listed.

| Modality | Model | ♯ Para. | Dim. | *Epoch 1* | *Epoch 2* | *Epoch Fusion* |
|---|---|---|---|---|---|---|
| Visual | VGG-Face | 134.28M | 4,096 | 0.9065 | 0.9060 | 0.9072 |
| | ResNet | 58.31M | 512 | 0.9072 | 0.9063 | 0.9080 |
| | DAN | 14.71M | 1,024 | 0.9082 | 0.9080 | 0.9100 |
| | DAN$^+$ | 14.72M | 2,048 | 0.9100 | 0.9103 | 0.9111 |
| Audio | Linear regressor | 0.40M | 79,534 | 0.8900 | – | 0.8900 |

For the audio modality, MFCC and logfbank are extracted. In the experiments of the competition, we extract 3,059 frames per audio. 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.

Because the audio features of this competition are in large scale, we simply train a linear regressor by Torch on GPUs. 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. Figure 7(a) and (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.

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




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

**Fig. 7.** Learning curves of two different audio features, i.e., MFCC and logfbank. The horizontal axis is the training epoch, and the vertical axis is the regression error.

**Table 2.** Comparison of performances of the proposed DBR framework with that of the top five teams in the Final Evaluation phase. The results in the "extra.", "agree.", "consc.", "neuro." and "open." columns stand for the separate Big Five Traits regression accuracy, respectively. ("NJU-LAMDA" is our team.)

| Rank | Team name | *Mean Acc.* | *Extra.* | *Agree.* | *Consc.* | *Neuro.* | *Open.* |
|---|---|---|---|---|---|---|---|
| 1 | **NJU-LAMDA** | **0.9130** | 0.9133 | **0.9126** | **0.9166** | **0.9100** | **0.9123** |
| 2 | evolgen | 0.9121 | 0.9150 | 0.9119 | 0.9119 | 0.9099 | 0.9117 |
| 3 | DCC | 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 |

model (a linear regressor). As described in Sect. 3.3, 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.

**Qualitative Evaluation of the Deep Visual Regression Networks.** In order to further justify the effectiveness of the deep visual regression networks of DBR, we visualize the feature maps of these three networks (i.e., ResNet, DAN and DAN$^+$) in Fig. 8. In that figure, we randomly sample five extracted images from different APA videos, and show the $\text{Pool}_5$ 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$^+$. It seems that ResNet could pay its attention on the human beings, while DAN/DAN$^+$ will focus on not only the human, but also the environments/backgrounds of these videos. Apparently, different deep visual regression networks could extract complementary information for images in apparent personality analysis.

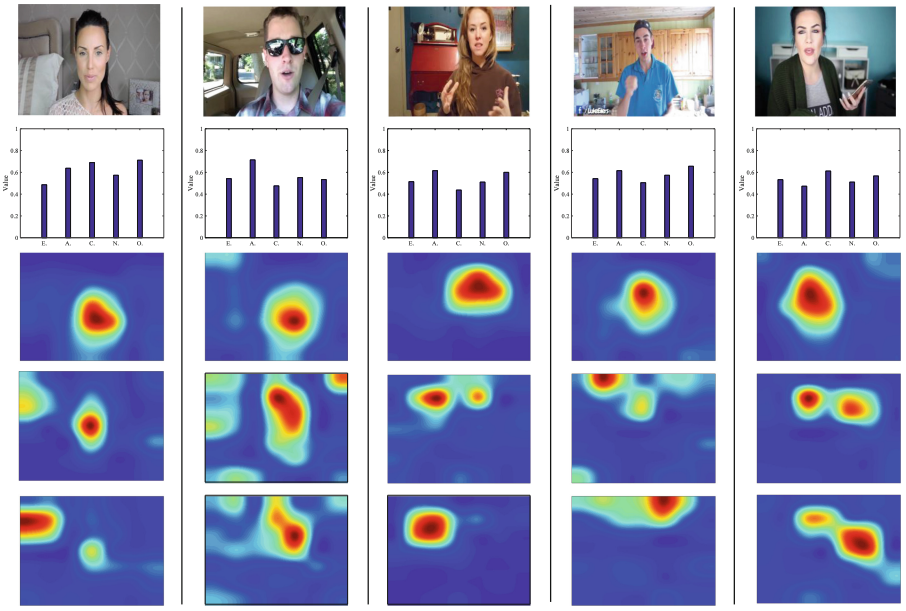


**Fig. 8.** Feature maps of five sampled images in the visual modality of DBR. The first row shows the images, and the second row presents their corresponding Big Five Traits values. The third, fourth and fifth rows show the featmaps of ResNet, DAN and DAN$^+$, respectively. For each feature map, we sum the responses values of all the channels in the final pooling layer for each deep network. These figures are best viewed in color. (Color figure online)

**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 results are shown in Table 2. Our final result (0.9130) ranked the first place, which significantly 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.

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 deep audio networks should be tried to learn the more discriminative audio representations. The ensemble of multiple audio models can be also applied into our DBR framework to achieve better apparent personality analysis performance.

## 5 Conclusions

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.

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

## References


1. Amer, M.R., Lei, P., Todorovic, S.: HiRF: hierarchical random field for collective activity recognition in videos. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 572–585. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_37
2. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the International Conference on Multimodal Interfaces, pp. 205–211 (2004)
3. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
4. Hasan, M., Roy-Chowdhury, A.K.: Continuous learning of human activity models using deep nets. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 705–720. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_46
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: DEAP: a database for emotion analysis using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
8. Ma, S., Sigal, L., Sclaroff, S.: Learning activity progression in LSTMs for activity detection and early detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1942–1950 (2016)
9. Makhoul, J.: Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–580 (1975)
10. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference, pp. 1–12 (2015)
11. Pfister, T., Charles, J., Zisserman, A.: Flowing convNets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1913–1921 (2015)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations, pp. 1–14 (2015)

13. Song, Y., Bao, L., Yang, Q., Yang, M.-H.: Real-time exemplar-based face sketch synthesis. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 800–813. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_51
14. Soo Park, H., Hwang, J.J., Shi, J.: Force from motion: decoding physical sensation in a first person video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3834–3842 (2016)
15. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Niessner, M.: Face2Face: real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
16. Vedaldi, A., Lenc, K.: MatConvNet - convolutional neural networks for MATLAB. In: Proceeding of ACM International Conference on Multimedia, pp. 689–692 (2015). http://www.vlfeat.org/matconvnet/
17. Wei, X.S., Luo, J.H., Wu, J.: Selective convolutional descriptor aggregation for fine-grained image retrieval. arXiv preprint arXiv:1604.04994 (2016)
18. Wei, X.S., Xie, C.W., Wu, J.: Mask-CNN: localizing parts and selecting descriptors for fine-grained image recognition. arXiv preprint arXiv:1605.06878 (2016)
19. Xiong, B., Grauman, K.: Detecting snap points in egocentric video with a web photo prior. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 282–298. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_19
20. Yan, X., Chang, H., Shan, S., Chen, X.: Modeling video dynamics with deep dynencoder. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 215–230. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_15
21. Yonetani, R., Kitani, K.M., Sato, Y.: Recognizing micro-actions and reactions from paired egocentric videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2629–2638 (2016)
22. Zhang, D., Shah, M.: Human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2012–2020 (2015)
23. Zhu, Y., Jiang, C., Zhao, Y., Terzopoulos, D., Zhu, S.C.: Inferring forces and learning human utilities from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3823–3833 (2016)