ImpressionNet: A Multi-view Approach to Predict Socio-facial Impressions

Rohan Kumar Gupta and Dakshina Ranjan Kisku
Department of Computer Science and Engineering, National Institute of Technology Durgapur, West Bengal, India

Keywords: Perception Analysis, Multi-view Learning, Biometrics.

Abstract: The visual facial features do reveal a lot about an individual and can be used to analyse several important social attributes. Existing works have shown that it is possible to learn these attributes through computational models and classify or score subject-faces accordingly. However, we find that there exists local variance in perception. There could be different perspectives of the face which the conventional methods fail to efficiently capture. We also note that Deeper neural networks usually require enough training data and add little to no improvement upon existing ones. In this work, we take social attribute prediction a notch higher and propose a novel multi-view regression approach to incorporate multiple views of face inspired by multi-modal learning. Experimental results show that the proposed approach can achieve superior feature generalisation and diversification on existing datasets using multiple views to improve the coefficient of determination scores and outperforms the state-of-the-art social attribute prediction method. We further propose a method that enables real-time video analysis of multiple subject faces which can have several applications.

1 INTRODUCTION

Social attributes like trustworthiness and dominance have always played a significant role in order to understand how humans perceive one another. Human facial expressions and certain cues do have a lot to do with our assessment and perception of these traits. Independent psychological studies have suggested that we give large importance to traits like trustworthiness when making judgement in social interactions (Frith, 2009). Several attempts have been made in the past in order to computationally analyse and describe these traits.

Keating et al. (Keating et al., 1981) were among the first to analyse how the positions of eyebrows and mouth affect our perceived dominance and happiness in a cultural context. They also noted that the effect of certain gestures such as lowered-brow was restricted to Western subjects indicating that there exists local variance in perception.

Todorov et al. (A et al., 2013) were able to develop computational models based on certain facial features to describe such social attributes through virtual avatars varying on different levels of standard deviation. Mel et al. (McCurrie et al., 2017) proposed a CNN-based approach and showed that such traits could be learned by a deep learning framework to assign a score to the real faces based on the complex features. They have also noted that deeper architectures add no improvement on the existing data.

However, we find that faces are local and have significant subject-dependent variations. There could be different perspectives of the face which can be learned and used for attribute score prediction which the conventional methods fail to efficiently capture. One of the methods to effectively capture multiple perspectives is to use multi-view learning (Xu et al., 2013). The idea is to maximize the mutual agreement on two distinct views of data and combine the results to improve the learning performance and generalization.

We also observe that continuous distributions are a better representation of these attributes rather than discrete values and can also be used for comparative analysis.

In this work, we carry forward this idea and propose a new multi-view based regression approach in order to better generalize and combine the results of different perspectives inspired by multi-modal learning. The idea is to generate conditionally independent features by training two different deep neural network streams which represent the different perspectives. To ensure that the two streams generate conditionally independent facial representations, we utilise a multi-view loss function. The two features are then

Gupta, R. and Kisku, D.
ImpressionNet: A Multi-view Approach to Predict Socio-facial Impressions.
DOI: 10.5220/0010762400003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 33-38
ISBN: 978-989-758-547-0; ISSN: 2184-433X

Figure 1: Diversity in faces across the different regions (AFLW Dataset).

finally used to jointly predict the specific social attribute score on a scale ranging from zero to one.

Most of the multi-view based methods have been used in the past for co-training related tasks in attempts to enhance the feature discrimination in a semi-supervised manner (Niu et al., 2019). Unlike the traditional co-training methods, we propose to use multi-view learning in a supervised manner to accommodate different perspectives and enhance the generalisation ability of our approach.

We also analyse our proposed approach on videos and predict scores on subject faces frame-wise in real-time to assess how the model performs in diverse situations by extending our architecture.

The contributions of this paper are as follows: i) we propose a new multi-view regression method for prediction of social attribute scores in order to analyse and compare faces leveraging diverse annotated faces from public domains. ii) we achieve superior performance and significant improvements in results without using more data indicating better generalisation of facial features. iii) we utilize and modify the proposed approach to enable detection of multiple subject faces in videos and score them simultaneously for a given attribute on a real-time basis.

2 RELATED WORK

Prediction using CNN-based Regression Framework: Mel et al. (McCurrie et al., 2017) have shown that social attributes such as trustworthiness and dominance can be predicted as a score using a convolutional neural network (CNN) based approach. They have hypothesised that the continuous distributions are a more realistic representation of social judgements humans make. They used a MOON framework trained on human faces from annotated AFLW dataset and have achieved the coefficient of determination (R^2) score of 0.38 (trustworthiness) and 0.46 (dominance) on a range of test faces. They have further concluded that low-level features have a significant role in immediate judgements and the data can be learned by a deep learning framework.

Attribute Classification by Comparison from AU Avatars: There have been techniques that predict the attribute metrics by comparing the facial features to the lab-generated facial action unit (AU) avatars. In (Safra et al., 2020), the authors have exploited a set of caucasian face avatars developed by Todorov et al. (A et al., 2013) to compare the facial features extracted from paintings using OpenCV's algorithm. They classified the faces into 7 different levels of trustworthiness and dominance based on standard deviation in Oosterhof and Todorov's model ranging between -3 to +3 SD using the random forest classifier. Notably, the work has only used caucasian faces for a region-specific analysis and may lack diversity in the real world.

In other related work, Liu et al. (Liu et al., 2015), have used cascaded CNNs and trained support vector machines to separate the processes of face localization and attribute prediction for a similar task.

3 PROPOSED METHODOLOGY

As described in the introduction, we use multi-view features obtained from the face images, to train our proposed network for attribute score prediction. We describe our model architecture below as well as the proposed training routine details.

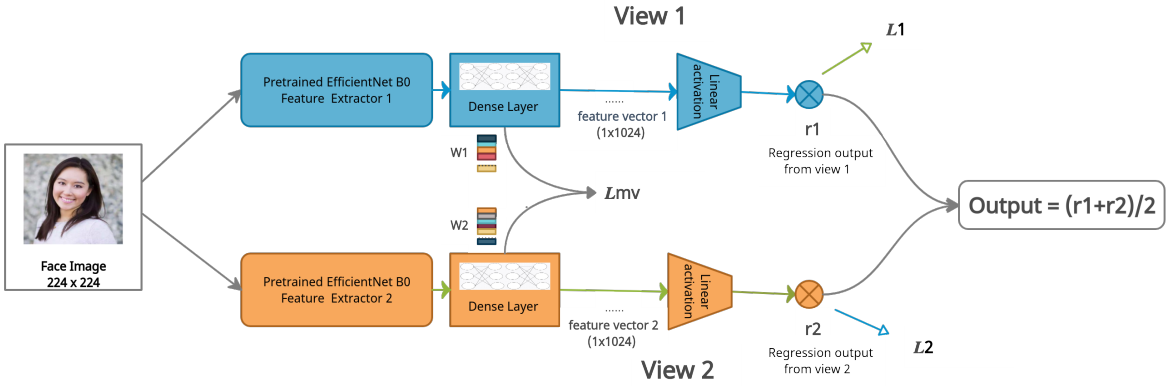

Figure 2: Proposed network architecture. The feature extractors are followed by fully connected layers consisting of dense layer whose weights are penalized by L_{mv}. The final layers have linear activation for regression task and their scores are meaned.

3.1 Pre-processing

The first step in the proposed methodology is to pre-process the images before training. For all face images used in the experiments, RetinaFace (Deng et al., 2020) is utilized for detection and aligning the faces. Face images are cropped to 224x224 pixels and converted to grayscale in order to remove any possibility of bias towards skin tone. The data augmentation performed includes horizontal flipping, rotation, width shift, height shift, shear and zoom.

3.2 Network Architecture

As discussed above, the proposed approach consists of two EfficientNet-B0 (Tan and Le, 2019) network streams pre-trained on ImageNet as feature extractors. The resultant feature vectors are then fed into the respective fully-connected layers that finally output the i-th view regression scores. Both the network streams have similar architecture, called ImpressionNet henceforth, as shown in Figure 2. Note that the weights of the two streams differ although the architecture is identical.

The ImpressionNet itself consists of two modules, the pre-trained backbone feature extractor and the fully connected layers. There is a dense layer consisting of 1024 neurons among the FC layers. We denote the weight parameter of this layer as W_i for the i-th view.

We denote f_i as the feature generated by the i-th stream. Since the features are supposed to be conditionally independent, a multi-view loss function is utilised to orthogonalize the weights of the respective dense layers to avoid collapsing of features into one another. The multi-view loss function L_{mv} is defined as:

$$L_{\text{mv}} = \frac{1}{2}\frac{W_1^T W_2}{|W_1||W_2|}$$

This ensures that multi-view features are different from each other while complimenting each other at the same time. The two streams can be trained simultaneously to predict the respective regression scores represented by r_i. After training the multi-view networks, we obtain the final score by taking the mean of the the i-th view scores as follows:

$$r_{\text{combined}} = \frac{(r_1 + r_2)}{2}$$

With this approach, the two networks tend to get more representative features and the combined generalisation ability of networks is enhanced significantly.

3.3 Loss Function

To train our model for the scoring task, we have used a linear activation function upon fully connected dense layers and have trained the network streams using two different loss functions. We have used the mean-squared error (MSE) function on individual view networks and a combined multi-view loss function to obtain conditionally independent features.

Mean Squared Error (MSE): We use mean squared error as our main loss function which is used to calculate the average square difference between the estimated values and the actual value on the predicted scores of individual network streams given by:

$$L_{\text{mse}} = \frac{1}{n}\sum_{1}^{n}(r_i - p_i)^2$$

Here r_i represent the i-th view output and p_i represent the actual attribute score.

Multi-view Loss (L_{mv}): As discussed above, in addition to the MSE, we also incorporate a secondary loss

Table 1: Coefficient of Determination (R^2) scores for prediction of social attributes by the proposed and baseline method on the databases.

Method	**Trustworthiness (A)**	**Dominance (B)**
Annotated AFLW Database (D1)		
MOON architecture [4]	0.3800	0.4600
proposed multi-view approach	0.4903	0.5523
Chicago Database (D2)		
MOON architecture [4]	0.3012	0.3441
proposed multi-view approach	0.3758	0.4464
Oslo Face Database (D3)		
MOON architecture [4]	0.3479	0.4021
proposed multi-view approach	0.3985	0.4795

Table 2: Coefficient of determination (R^2) scores of different approaches in the ablation study on the different databases (D1 - Annotated AFLW Database, D2 - Chicago Database, D3 - Oslo Face Database).

Method	D1-A	D1-B	D2-A	D2-B	D3-A	D3-B
Baseline	0.4227	0.4849	0.3307	0.3613	0.3641	0.4201
Ensemble-baseline	0.4442	0.4997	0.3395	0.3663	0.3802	0.4226
Ensemble-Baseline + L_{mv}	**0.4903**	**0.5523**	**0.3758**	**0.4464**	**0.3985**	**0.4795**

function to penalize the weight parameters of individual dense layers in order to obtain multi-view features and efficiently train the network given by:

$$L_{\text{mv}} = \frac{1}{2}\frac{W_1^T W_2}{|W_1||W_2|}$$

Here W_1 and W_2 represent the weight parameters of penultimate dense layers of two network streams.

4 EXPERIMENTS AND RESULTS

In order to validate our hypothesis and justify our approach, we have performed a methodical experimental analysis, details of which are described below.

4.1 Dataset Specifications

In this work, we have used three publicly available databases, i.e., Chicago Face Database (Ma et al., 2015), Oslo Face Database (O et al., 2014) and Annotated AFLW Database (McCurrie et al., 2017) which are often used for behavioural and perceptual analysis. By doing so, we have significantly diversified our dataset in order to fairly assess our approach and reduce the human bias.

Chicago Face Database consists of images of about 597 individuals of diverse ethnicity including asians, black, latino and white. All the subjects are represented with neutral facial expressions. The face Images were provided with manually annotated subjective ratings (e.g., trustworthiness).

Oslo Face Database consists of about 630 male and female faces of neutral expression with three gaze directions: left, right, center. The subjects are mainly from the University of Oslo. The face images contain annotations for several social attributes like attractiveness, perceived dominance and trustworthiness.

Annotated AFLW Dataset consists of about 6,300 grayscale images of face sampled from AFLW dataset and annotated for subjective ratings by a crowdsourced psychophysics testing website.

We perform normalization on the social attribute scores on each of these databases so as to bring them under the range of 0-1.

4.2 Training and Testing Protocol

For training our multi-view CNN-based model, we use two EfficientNet-B0 networks pre-trained on Imagenet as our backbone feature generators followed by fully connected layers. In order to train the model on the three databases, we split the annotated images randomly by maintaining an equal ratio of male and female subjects for the training and testing set. The initial number of training image samples before augmentation for experiment is 5840, 160, 530 for the annotated AFLW, Oslo and Chicago Databases respectively. The Adam optimizer with an initial learning rate of 0.001 is applied to jointly train the feature generators and FC layers of two streams. We set the maximum iteration to 75 epochs. The batch size for all the experiments is set to 32. We also use an early stopping function while training in order to select the best model. All experiments are conducted

Figure 3: Video Frames from the processed video feed indicating detected faces and their attribute scores (trustworthiness) in real-time.

with Keras (Chollet et al., 2015) on a NVIDIA Tesla K80 GPU.

4.3 Performance and Evaluation Metrics

In order to validate the proposed approach, we have used the coefficient of determination (R^2) as our metric which is described below:

Coefficient of determination is a well-known statistical measure in regression models that determines the proportion of variance in the prediction that can be explained by the model.

4.4 Performance Analysis

We summarize the results of our proposed network's performance on the test set in Table 1 and choose the MOON architecture proposed by Mel et al. (McCurrie et al., 2017) for a fair comparison. The MOON Architecture utilizes a single network to learn the features, performance of which is also provided in Table 1.

From the results, we can see that when comparing with the MOON architecture which doesn't require multi-view features, our method outperforms the previous work by a fair margin on the R^2 scores. This indicates that optimizing feature learning along with averaging the views is helpful for improving the performance for score prediction and also improves the overall generalization ability of the model.

4.5 Ablation Study

We also provide the ablation study to investigate the effectiveness of the multi-view loss function (L_{mv}) on the proposed multi-view regression approach. We also summarize the results obtained using different strategies in Table 2 for two social attributes: Trustworthiness (A) and Dominance (B).

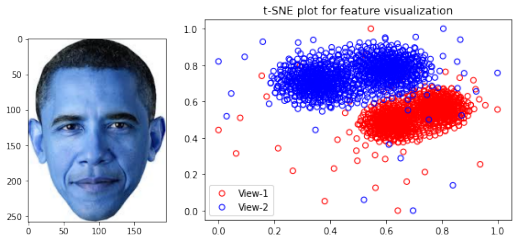

Figure 4: (a) Sample subject face (b) t-SNE plot for visualization of features extracted from corresponding views.

Effectiveness of Multi-view Loss (L_{mv}): From the results in Table 2, we can see that the improvement in the proposed model is mainly gained from the L_{mv}, with an average R^2 score increased by 0.02, which indicates the effectiveness of our multi-view loss function in regression-based task as well. To further analyse the effect of multi-view approach on the two features generated by the different views, we use t-SNE (van der Maaten and Hinton, 2008) to visualize the features of a random face image provided in Fig 4. We can clearly see from the Fig. that features generated from two views are diverse from each other. Thus it further strengthens our claim that L_{mv} can help the network in learning more diverse representations and subsequently improve the scores.

5 MODIFICATIONS FOR REAL-TIME VIDEO PREDICTION

We also propose a modification in the architecture in order to detect multiple subject faces in a video and score them simultaneously. The first step in the proposed methodology is to pre-process the videos. We use OpenCV's VideoCapture (G, 2000) in order to preprocess the video and split the video into individual frames.

In order to identify and process the multiple subject faces which may be present in the frame, we uti-

lize an open-source face detection framework Caffe (Jia et al., 2014). Individual frames are then fed into the face detector which detects and crops the subject faces. We then utilise our trained ImpressionNet model in order to predict the attribute scores on these face images and annotate the scores on the original frame. The individual annotated frames are merged back to form the video in real-time. Fig. 3 shows several frames from a processed video using the above approach. We have also provided few annotated videos from diverse social settings as supplementary material.

6 CONCLUSION

In this paper, we have proposed a fresh multi-view approach for analysis and prediction of social attributes like trustworthiness and dominance based on facial features. The proposed approach achieves superior feature generalisation and diversification resulting in improved coefficient of determination (R^2) scores. Our experiments validate that one can extract more diverse features using multiple views and subsequently improve the performance by combining their results in regressive tasks as well. To justify the diversification and generalisation ability of our approach, we have also performed the ablation study. The obtained results clearly establish the effectiveness of this approach and also indicates that similar methods can also be used for analogous tasks. At last, we also proposed a method which enables the real-time video analysis of multiple subject faces and can have several applications in marketing, surveillance and more.

7 SUPPLEMENTAL MATERIAL

More results and evaluations on video sequences of various test subjects can be seen in the supplemental annotated videos from the link below. We show examples of various social interactions and how it affects the perception of the social attributes like trustworthiness based on the facial expressions analysed by our multi-view regression approach (see, for example, the fluctuations in trustworthiness scores associated in tense situations such as in political interviews or broadcasting studios when the subject is judged highly in Videos). https://anonymous.4open.science/r/3a48498b-871f-4484-93dc-c9982e11fd65/README.md

REFERENCES

A, T., R, D., JM, P., NN, O., and VB., F. (2013). *Validation of data-driven computational models of social perception of faces.* Emotion.

Chollet, Francois, et al. (2015). *Keras*. GitHub.

Deng, Jiankang, Guo, Jia, Ververas, Evangelos, Kotsia, Irene, Zafeiriou, and Stefanos (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society.

Frith, C. (2009). *Role of facial expressions in social interactions*. The Royal Society.

G, B. (2000). *The OpenCV Library*. Dr. Dobb's Journal of Software Tools.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, and Trevor (2014). *Caffe: Convolutional Architecture for Fast Feature Embedding*. ACM.

Keating, C. F., Mazur, A., Segall, M. H., Cysneiros, P. G., Kilbride, J. E., Leahy, P., Divale, W. T., Komin, S., Thurman, B., and Wirsing, R. (1981). *Culture and the perception of social dominance from facial expression*. Journal of Personality and Social Psychology.

Liu, Z., Luo, P., Wang, X., , and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE Computer Society.

Ma, Correll, and Wittenbrink (2015). *The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data*. Behavior Research Methods.

McCurrie, M., Beletti, F., Parzianello, L., Westendorp, A., Anthony, S., and Scheirer, W. (2017). Predicting first impressions with deep learning. In *12th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE Computer Society.

Niu, Xuesong, Han, Hu, Shan, Shiguang, Chen, and Xilin (2019). Multi-label co-regularization for semi-supervised facial action unit recognition. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

O, C., B, L., M, E., J, R., G, L., H, M., F, W., and S, L. (2014). *Rewards of Beauty: The Opioid System Mediates Social Motivation in Humans*. Mol Psychiatry.

Safra, L., Chevallier, C., Grèzes, J., and Baumard, N. (2020). *Tracking historical changes in trustworthiness using machine learning analyses of facial cues in painting*. Nature Communications.

Tan, M. and Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

van der Maaten, L. and Hinton, G. (2008). *Visualizing data using t-SNE*. Journal of machine learning research.

Xu, C., Tao, D., and Xu, C. (2013). *A Survey on Multi-view Learning*. CoRR.