Joint Multi-Feature Fusion and Attribute Relationships for Facial Attribute Prediction

Pingyu Wang #1, Fei Su #2, Zhicheng Zhao #3

# *School of Communication and Information Engineering, Beijing University of Posts and Telecommunications, Beijing, China*

# *Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China*

[1] applewangpingyu@gmail.com [2] sufei@bupt.edu.cn [3] zhaozc@bupt.edu.cn

***Abstract*—Predicting facial attributes from wild images is very challenging due to complex face variations. The key to this problem is to construct rich facial representations and take advantage of attribute relationships. In this paper, we propose a novel multi-task convolutional neural network (MTCNN) and a supervision signal called Online Batch Relation Loss (OBRL) for face attribute prediction in the wild. In particular, MTCNN builds informative facial features by embedding identity, age and race features from IdentityNet, AgeNet and RaceNet respectively. In addition, OBRL can diminish distribution shift of attribute relationships by mining attribute correlation within each mini-batch, while it penalizes the probability divergence between a pair of attributes. In order to learn discriminative attribute features, we feed AttributeNet with fused facial features and partition attributes into nine groups to share intra-group features and reduce redundant computation. Finally, AttributeNet is optimized with the joint supervision of Cross Entropy Loss and OBRL. Experiments on CelebA and LFWA show that the proposed method outperforms the state-of-the-art methods with a significant margin.**

***Index Terms*—Feature Fusion, Multi-Task, Attribute Prediction, Attribute Relationships, Online Batch Relation Loss**

I. Introduction

Understanding facial attributes to describe semantic face representation is a promising technique in the field of computer vision. Various visual tasks take advantage of facial attributes including transfer learning [1], attention learning [2] and face retrieval [3]. However, learning facial attribute representations from massive wild images is very challenging due to many unfavorable variations, such as illumination, pose and occlusion.

Driven by the great improvements brought by the deep convolutional neural network (DCNN) in large scale image classification [4], many traditional methods have been replaced by DCNN for feature extraction in some areas including face recognition [5] and attribute prediction [6]. Previous researches [7] also show that fusing various facial features can improve the performance of face detection and face recognition.

Our work revisits off-the-shelf methods by proposing a novel multi-task convolutional neural network (MTCNN) to form rich facial features and take advantage of attribute relationships. The proposed framework integrates four DCNNs: IdentityNet, AgeNet, RaceNet and AttributeNet. Specifically, IdentityNet, AgeNet and RaceNet are fed with face images to learn face-level representations such as identity, age and race features, then AttributeNet takes the fused face-level representations as input and learns attribute features for attribute prediction. IdentityNet learns to discriminate different identities which helps AttributeNet to capture global identity-related attribute features such as “Attractive”, “Male” and “Chubby”. The age-aware features from AgeNet are beneficial to obtaining robust attribute features such as “Young” and “Black_Hair”. RaceNet learns ethnic features to improve the prediction performance of race-related attributes such as “Blond_Hair” and “Pale_Skin”. In this way, embedding identity, age and race features can help discover facial attributes.

Owing that some facial attributes are spatially-related, we partition 40 facial attributes into nine groups. The proposed framework treats each group as an independent task to share common representations among intra-group attributes and reduce computing resources.

We also notice that facial attributes are highly correlative. For example, if people are wearing heavy makeup or lipstick, the probability that they are women increases, and vice versa. Some former literatures [8] mine global attribute relationships through the entire dataset or the experience of common sense off the line. However, global attribute relationships may suffer from distribution shift when extra data are used to finetune models. Drawing the inspiration from Batch Normalization [9], we put forward an iterative method called Online Batch Relation (OBR) to update the attribute relation matrix within each mini-batch. In this paper, Online Batch Relation Loss (OBRL) is also proposed to encourage AttributeNet to learn relationships among facial attributes. Therefore, we combine Cross Entropy Loss and OBRL to optimize MTCNN. The final results infer that our model can surpass other advanced methods.

Our main contributions are as follows:

1. To be the best of our knowledge, it is the first time that leveraging discriminative identity, age and race features to construct deep attribute representations. Experimental results show that fusing related and informative features can benefit to facial attribute prediction.

978-1-5386-0462-5/17/$31.00 © 2017 IEEE

2. We propose a new loss function OBRL to learn relationships among facial attributes. With the joint supervision of Cross Entropy Loss and OBRL, the highly discriminative features can be obtained for robust face attribute prediction, as supported by our experimental results.

3. The proposed method can reach state-of-the-art performance on CelebA [6] and LFWA [10] datasets.

II. Proposed Method

Figure 1 shows the MTCNN architecture. It contains four models: IdentityNet, AgeNet, RaceNet and AttributeNet. A facial image (left) is resized to 224 × 224, then fed into the model. Identity, age and race features are extracted from IdentityNet, AgeNet and RaceNet respectively, and are fused along channel dimension to form face-level representations. Then the fused features are input into AttributeNet where each attribute group is taken as an independent task. Additionally, AttributeNet is formulated under a multi-task learning framework and optimized with the joint supervision of Cross Entropy Loss and OBRL.

A. Face-level Representation

In this section, we discuss the details of training IdentityNet, AgeNet and RaceNet, which are the foundations of face-level representations. The architecture of those three models, which bears a resemblance to AlexNet [4], is selected to extract face-level features. It is characterized by the decreased stride and smaller receptive field in convolutional layers, which is beneficial to extracting face-level features.

Following the common practice of training DCNN [6], we pre-train IdentityNet, AgeNet and RaceNet by classifying 1,000 generic categories. However, DCNN trained on generic objects is not capable of providing us with clear and precise facial response maps. To address this problem, we implement the first finetuning stage on those three models with face and non-face images from CASIA-WebFace [11] and SUN397 [12] respectively. In the following, we will show the second finetuning process of IdentityNet, AgeNet and RaceNet respectively.

For IdentityNet, we select 100k face identities from the MS-Celeb-1M [13], where each identity has around 100 images. To preserve intra-class invariance of different identities, we employ center loss [14] which simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers.

For AgeNet, we also prepare 500k samples from IMDB-WIKI [15]. Since age prediction is a regression task, we apply the Euclidean loss function and stochastic gradient descent (SGD) by standard back-propagation algorithm to optimize AgeNet in the second finetuning stage.

For RaceNet, since there are no available face datasets containing ethnic targets, we collect a new RaceFace dataset including 12k images from Google Image Search and YouTube. We have labeled each image manually. RaceFace covers black, white and yellow celebrities with complex variations. In the second finetuning stage, RaceNet is optimized with Softmax loss.

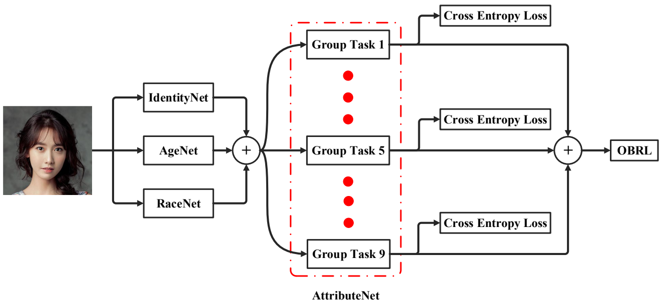

Fig. 1. The proposed MTCNN framework

TABLE I
Attribute Groups

Group	Attributes
Hair Group	Bald, Bangs, Black_Hair, Blond_Hair, Brown_Hair, Gray_Hair, Receding_Hair, Straight_Hair, Wavy_Hair, Wearing_Hat
Eye Group	Arched_Eyebrows, Bags_Under_Eyes, Bush_Eyebrows, Eyeglasses, Narrow_Eyes
Ear Group	Wearing_Earrings
Nose Group	Big_Nose, Pointy_Nose
Mouth Group	Big_Lips, Mouth_Slightly_Open, Smiling, Wearing_Lipstick
Beard Group	5_O_Clock_Shadow, Double_Chin, Goatee, Mustache, No_Beard, Sideburns
Cheek Group	High_Cheekbones, Rosy_Cheeks
Neck Group	Wearing_Necklace, Wearing_Necktie
Global Group	Attractive, Blurry, Chubby, Heavy_Makeup, Male, Over_Face, Pale_Skin, Young

B. Attribute-level Representation

After the second finetuning stage, IdentityNet, AgeNet, and RaceNet encode rich face-level representations in the convolutional layer. We retain all the convolutional layers to extract face-level features in view of this fact that convolutional layers contain more spatial information than fully-connected layers. After that, a concatenation operator is employed to fuse those three features along channel dimension.

With the highly spatial correlation among attributes and reducing redundant computation in feature extraction stage, all facial attributes are split into nine groups which are shown in Table I, and each group shares common attribute-level features. Moreover, we employ a convolutional layer and two fully-connected layers for each task. In summary, AttributeNet is finetuned with the joint supervision of Cross Entropy Loss and OBRL, which will be described in the following.

C. Online Batch Relation Loss

Traditional methods build attribute relationships through the entire dataset, and global relation matrix $R_g \in \mathbb{R}^{M \times M}$ (M is the number of attributes) remains unchanged during the whole training phase. However, those methods have two shortcomings. Firstly, computing global relation matrix through the entire dataset is costly if the dataset has a large number of samples. Secondly, if extra samples are added into the current dataset, we have to recompute R_g since the distribution of attribute relationships may be shifted.

Unlike classic ideas, we propose an iterative and online algorithm called Online Batch Relation (OBR) to approximate real global relation matrix in this paper. We adopt exponential moving average (EMA) method to update global relation matrix on the line. For explanation convenience, we define $R_{sb} \in \mathbb{R}^{M \times M}$ as the element-wise sum of batch relation matrixes, $R_b \in \mathbb{R}^{M \times M}$ is the current batch relation matrix, and $S \in \mathbb{R}^1$ is the sum of moving average fraction. So the online updating process can be defined as:

$$\begin{aligned} &R_{sb}(k+1) = \lambda R_{sb}(k) + R_b \\ &S(k+1) = \lambda S(k) + 1 \\ &R_g(k+1) = R_{sb}(k+1)/S(k+1) \end{aligned} \tag{1}$$

where the moving average fraction meets $\lambda \in (0, 1]$ and k denotes the k-th mini-batch iteration.

Drawing inspiration from the work [8], we propose a novel loss function called OBRL to utilize attribute relation supervision to achieve better prediction performance without extra samples. For further explanation, we define $y_{ij} \in \{0, 1\}$ as the ground truth of the j-th attribute in the i-th sample, and x_{ij} denotes the corresponding prediction result. The probability $p_{ij} = p(y_{ij} = 1)$ can be computed by a sigmoid function $p_{ij} = (1 + e^{-x_{ij}})^{-1}$ and OBRL L_B is defined as:

$$\begin{aligned} L_B = \frac{1}{2KM^2} \sum_{k=1}^{K} \sum_{i,j=1}^{M} & f_p(R_{ij}) \left\| p_{ki} - p_{kj} \right\|_2^2 \\ & + f_n(R_{ij})(1 - \left\| p_{ki} - p_{kj} \right\|_2^2)^2 \end{aligned} \tag{2}$$

where R_{ij} is the correlation coefficient between the i-th and the j-th attribute and K denotes the mini-batch size. Two threshold functions $f_p(x)$ and $f_n(x)$ are respectively defined as:

$$f_p(x) = \begin{cases} 1 & x >= r_p \\ 0 & others \end{cases} \text{ and } f_n(x) = \begin{cases} 1 & x <= r_n \\ 0 & others \end{cases} \tag{3}$$

where r_p and r_n are the threshold of positive and negative correlations respectively. We combine Cross Entropy Loss and Online Batch Relation Loss as the final loss function $L = L_E + \alpha L_B$, where α denotes loss weight and L_E stands for Cross Entropy Loss:

$$L_E = -\frac{1}{K} \sum_{k=1}^{K} \sum_{m=1}^{M} y_{km} \log(p_{km}) + (1 - y_{km}) \log(1 - p_{km}) \tag{4}$$

III. Experiments

To evaluate the performance of the proposed MTCNN framework and OBRL, we conducted extensive experiments on CelebA and LFWA with Caffe [16]. Throughout the experiments, we fix moving average fraction $\lambda = 0.99$, loss weight $\alpha = 0.5$, positive threshold $r_p = 0.8$ and negative threshold $r_n = 0.2$. The performance is measured by average accuracy of attribute prediction.

TABLE II
Average Predicting Accuracy on CelebA and LFWA with Different Fusing Face-level Features

Feature Fusion	OBRL ?	CelebA	LFWA
Net_0	×	89.21%	84.17%
Net_0	✓	90.51%	85.37%
Net_1	×	89.89%	84.72%
Net_1	✓	90.97%	85.69%
Net_2	×	89.56%	85.02%
Net_2	✓	91.08%	85.81%
Net_3	×	89.73%	84.94%
Net_3	✓	91.24%	85.90%
$Net_1 + Net_2$	×	90.09%	85.23%
$Net_1 + Net_2$	✓	91.12%	86.27%
$Net_1 + Net_3$	×	89.42%	85.32%
$Net_1 + Net_3$	✓	91.26%	86.19%
$Net_2 + Net_3$	×	89.54%	85.43%
$Net_2 + Net_3$	✓	90.55%	86.28%
$Net_1 + Net_2 + Net_3$	×	90.38%	85.97%
$Net_1 + Net_2 + Net_3$	✓	**91.47%**	**86.51%**

A. Face-level Representation Effectiveness

In this experiment, we evaluate the effectiveness of face-level representations from IdentityNet, AgeNet and RaceNet. To guarantee the validity of the followed experiment, we experiment with eight kinds of feature fusion on CelebA and LFWA. The results are shown in Table II. Net_1, Net_2 and Net_3 represent IdentityNet, AgeNet and RaceNet respectively. Net_0 is finetuned with face and no-face samples and its architecture is the same as Net_1, Net_2 and Net_3. Under the condition of utilizing the same loss supervision signal, the performance increases by fusing richer facial features. On average, the accuracy improves by 0.96% and 1.14% approximately on CelebA and LFWA in comparison with Net_0. It is clear that fusing various face-level features helps to discover semantic concepts of facial attributes.

B. OBRL Evaluation

In this section we evaluate the effectiveness of OBRL on CelebA and LFWA. As shown in Table II, the performance of MTCNN with the joint supervision of Cross Entropy Loss and OBRL is better than the model with only Cross Entropy Loss. So we demonstrate that attribute relationships can be employed to learn better attribute features effectively.

We also explore the sensitiveness of loss weight α on CelebA and LFWA. α is varied from 0 to 1 to obtain average accuracy while other hyper parameters are fixed. The results are shown in Figure 2. It is obvious that properly choosing α can improve prediction performance in our method. It also indicates that the performance of the proposed method maintains highly stable across a wide range of α.

C. Comparison with the State-of-the-art Methods

Now we compare our MTCNN model on CelebA and LFWA with seven advanced methods: FaceTracer [17],

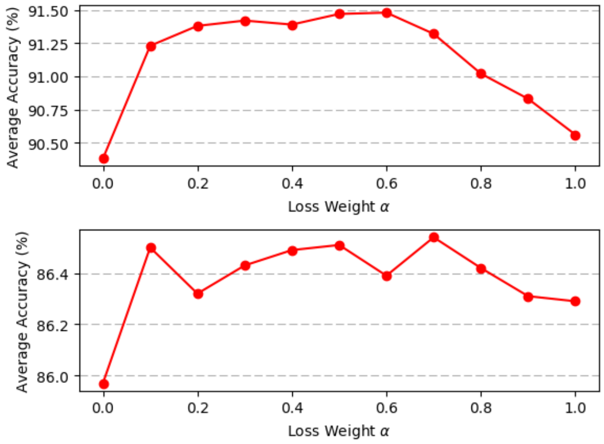

Fig. 2. Attribute prediction accuracies on CelebA (top) and LFWA (down)

PANDA-l [18], PANDA-w [18], LNet + ANet [6], Location + Weather [19], Off-the-shelf [20] and FaceSTN [21].

FaceTracer trains a SVM classifier with hand-crafted features (HOG + color histogram) on face regions. PANDA-l and PANDA-w predict attributes by embedding multiple facial part features. LNet + ANet cascades two CNNs, LNet and ANet, which can detect faces and predict attributes respectively. Weather and Location extracts informative representation from WeatherNet and LocationNet to learn discriminative attribute features. Off-the-shelf uses the existing architectures to extract different level features to predict attributes. FaceSTN implements a spatial transformer network (STN) to improve performance of face alignment and attributes prediction simultaneously. As shown in Table III, our approach achieves superior prediction performance over those state-of-the-art methods on CelebA and LFWA.

TABLE III
AVERAGE PERFORMANCE COMPARISON ON CELEBA AND LFWA

Method	CelebA	LFWA
FaceTracer [17]	81.13%	73.93%
PANDA-l [18]	85.1%	80.06%
PANDA-w [18]	85.43%	81.03%
LNet + ANet [6]	87.3%	83.5%
Location + Weather [19]	88.65%	86.6%
Off-the-shelf [20]	86.3%	84.5%
FaceSTN [21]	91.1%	86.0%
MTCNN (without OBRL)	90.38%	85.97%
MTCNN (with OBRL)	**91.47%**	**86.51%**

IV. CONCLUSION

In this paper, we propose a novel MTCNN framework for predicting facial attributes with the joint supervision of Cross Entropy Loss and OBRL. Different from previous methods, our approach can learn discriminative features for facial attributes by fusing various face-level features. The proposed MTCNN can leverage hidden correlation among attributes through OBRL without requiring the cost of extra samples, which outperforms the state-of-the-art methods across two challenging benchmarks, CelebA and LFWA.

ACKNOWLEDGMENT

This work is supported by Chinese National Natural Science Foundation (61372169, 61532018), and the Key Laboratory of Forensic Marks, Ministry of Public Security, Beijing, China.

REFERENCES

[1] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in *CVPR*, 2014.
[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *ICCV*, 2015.
[3] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, “Describable visual attributes for face verification and image search,” *TPAMI*, 2011.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.
[5] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015.
[6] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *ICCV*, 2015.
[7] X. Tan and B. Triggs, “Fusing gabor and lbp feature sets for kernel-based face recognition,” in *AMFG*, 2007.
[8] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, “Joint patch and multi-label learning for facial action unit detection,” in *CVPR*, 2015.
[9] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv:1502.03167*, 2015.
[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Tech. Rep., 2007.
[11] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *arXiv:1411.7923*, 2014.
[12] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *CVPR*. IEEE, 2010.
[13] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “Ms-celeb-1m: A dataset and benchmark for large-scale face recognition,” in *ECCV*. Springer, 2016.
[14] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *ECCV*. Springer, 2016.
[15] R. Rothe, R. Timofte, and L. Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *IJCV*, 2016.
[16] Y. Jia, E. Shelhamer, and Donahue, “Caffe: Convolutional architecture for fast feature embedding,” in *international conference on Multimedia*. ACM, 2014.
[17] N. Kumar, P. Belhumeur, and S. Nayar, “Facetracer: A search engine for large collections of images with faces,” in *ECCV*. Springer, 2008.
[18] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, “Panda: Pose aligned networks for deep attribute modeling,” in *CVPR*, 2014.
[19] J. Wang, Y. Cheng, and R. Schmidt Feris, “Walk and learn: Facial attribute representation learning from egocentric video and contextual data,” in *CVPR*, 2016.
[20] Y. Zhong, J. Sullivan, and H. Li, “Face attribute prediction using off-the-shelf cnn features,” in *ICB*. IEEE, 2016.
[21] L. Tan, Z. Li, and Q. Yu, “Deep face attributes recognition using spatial transformer network,” in *ICIA*. IEEE, 2016.