HIERARCHICAL MULTI-TASK NETWORK FOR RACE, GENDER AND FACIAL ATTRACTIVENESS RECOGNITION

Lu Xu[1], *Heng Fan*[2], *Jinhai Xiang*[1,*]

[1]College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China
[2]Department of Computer & Information Sciences, Temple University, Philadelphia, 19122, USA

ABSTRACT

Deep learning has powered many face related tasks and shown state-of-the-art performance. However, existing deep models are often trained separately for different problems, which results in heavy computational burden. To address this problem, we propose a novel multi-task network with fully convolutional architecture–Hierarchical Multi-task Network (HMTNet), that simultaneously recognizes a person's gender, race and facial attractiveness from a given portrait image. Aiming to improve the robustness to outliers in facial beauty prediction task, a novel loss is introduced into HMTNet. Compared to existing deep approaches, the proposed HMTNet achieves state-of-the-art performance on several datasets, and it can learn more discriminative feature representation through joint training and feature aggregation. Extensive experiments evidence the effectiveness of HMTNet.

Index Terms— Deep learning, multi-task learning, hierarchical multi-task neural network (HMTNet), facial beauty prediction, race recognition, gender recognition

1. INTRODUCTION

Facial beauty prediction (FBP) has been extensively studied in recent decades [1, 2, 3]. With the popularity of social networks services (SNS) (like Facebook and Instagram) and short video platforms (like TikTok and Musical.ly). FBP gains increased attention in both academic and industrial fields [4, 5, 6].

Previous works [7, 3, 1] indicate that data-driven models can be used to automatically learn facial attractiveness. Inspired by the success of deep convolutional neural networks (DCNN) [6, 8, 9] in computer vision (e.g., object detection [10, 11], visual tracking [12, 13], semantic segmentation [14, 15], etc.), face related tasks have witnessed many great achievements in recent years. However, due to the diverse poses, facial expression, low resolution and illumination problems, it remains challenging to develop an accurate model for precisely predicting facial attractiveness levels.

*Corresponding author. This work was primarily supported by National Key R&D Program of China(NO.2018YFC1604000) and Foundation Research Funds for the Central Universities (Program No. 2662017JC049).

Deep models are often data-hungry and parameter-heavy, which make them easy to stuck in overfitting. HMTNet can overcome these problems by learning from three relevant but different tasks, which greatly solves the dilemma of scarce training datasets for all related tasks. In addition, HMTNet is a kind of *fully convolutional neural network* [15], which is more light-weighted and needs less computational resource to train than its counterparts with fully connected layers. Furthermore, conventional regression models using mean squared error (MSE) loss are easily influenced by outliers. We introduce *Smooth Huber Loss* to effectively solve this problem.

HMTNet is mainly composed of three parts: *shared feature layers* for learning universal representations, *feature aggregator* to form more discriminative representations by fusing deep features from multi-level layers, *branched layers* for diverse recognition tasks (race, gender and facial beauty).

The main contributions of this paper are as follows: (1) We propose a novel fully convolutional network named HMTNet, to simultaneously recognize human gender, race and facial beauty with state-of-the-art performance on relevent benchmark datasets [16, 17]. (2) We introduce an effective loss function named "Smooth Huber Loss" in FBP task, which is more robust to outliers and yields better results than traditional MSE loss, L_1 loss and Smooth L_1 loss widely used in regression problems. (3) The transferability of the jointly learned features is studied during our experiments. (4) Through detailed analysis and visualization of deep features, we reveal the most significant elements for facial beauty perception.

2. RELATED WORKS

2.1. Facial Beauty, Race and Gender Recognition

Facial beauty prediction has attracted much attention recently [6, 8, 17, 7, 4]. Conventional methods extract geometry features, color features, and texture features to represent a face and train supervised models to predict facial beauty [7, 2]. However, due to the limited representation power of hand-crafted features, they fail to achieve satisfactory performance on larger datasets. Since deep learning has

boosted many computer vision tasks [18, 19, 20], researchers pay more attention to DCNN. Gray et al. [8] propose a CNN-based method and achieve a Pearson Correlation (PC) of 0.425 on the challenging Hot-or-Not dataset [8]. Xu et al. [6] achieve a PC of 0.87 on SCUT-FBP [17] dataset with their proposed PI-CNN. Rothe et al. [4] introduce a collaborative filtering method for preference prediction. Race and gender recognition are widely used among biometric applications [21]. But previous works pay little attention to the relation of these tasks and feature transferability.

2.2. Multi-task Learning

Despite the quite promising performance of DCNN, the models should be designed and trained separately according to variant tasks. Multi-task learning (MTL) [22, 23, 24] have aroused increased attention in recent years. HyperFace [22] can conduct pose estimation, face detection, facial landmarks localization, and gender recognition in one neural network with very promising results. MTCNN [23] can be used for facial landmarks localisation and face detection simultaneously.

Diffrent from previous MTL models, HMTNet is a MTL model with fully convolutional architecture. Instead of splitting the branches for individual subtasks in same layers [22, 23, 24], we split the branches from different level layers according to the learning difficulty of decision boundaries.

3. PROPOSED METHODS

Previous studies [24, 22] indicate that low-level features are more general and can be shared among other tasks, despite being guided by different loss functions. HMTNet is a MTL model with fully convolutional architecture. Compared with MTCNN [23], which needs to be trained sequentially. HMTNet follows an absolutely "end-to-end" fashion, and the optimization objects of individual sub-tasks are optimized jointly, which is more convenient for both training and deployment.

The overall architecture of HMTNet is illustrated in Fig. 1, and the detailed architecture of GNet and RNet are listed in Table 1. To the best of our knowledge, we are the first to adopt multi-task model with fully convolutional architecture in FBP task.

Table 1. Architecture of GNet and RNet. GNet is branched after *relu3*, and RNet is branched after *relu5* in HMTNet (see Figure 1). The input to GNet/RNet is $13 \times 13 \times 512$. BatchNorm and ReLU layers are not shown for simplicity.

Architecture of GNet & RNet
Conv (In=512, Out=256, Kernel=3, Stride=1, Pad=0)
Conv (In=256, Out=128, Kernel=3, Stride=1, Pad=0)
Max Pooling (3)
Conv (In=128, Out=2, Kernel=1, Stride=2)
Global Average Pooling

3.1. Feature Aggregator

Feature aggregator can form more discriminative representation by aggregating features from different layers for diverse recognition tasks. In this paper, we adopt two kinds of feature aggregation methods named *average aggregation* and *depth-wise concatenate aggregation*. The former constructs richer representation by averaging feature maps from different layers, which contain both low-level and high-level information. While the latter forms more informative features by concatenating feature maps with same dimensions. Equation 1 and Equation 2 describes "average aggregation" and "depth-wise concatenate aggregation", respectively. In our experiments, we find *average aggregation* yields better performance than *depth-wise concatenate aggregation* (0.8783 VS 0.8706).

$$f_{avg} = \frac{1}{\mathcal{C}} \sum_{i=1}^{\mathcal{C}} fm_i, \quad fm_i, f_{avg} \in \mathbb{R}^{w \times h \times c} \tag{1}$$

$$f_{concat} = fm_1 \otimes \cdots \otimes fm_{\mathcal{C}}, \quad f_{concat} \in \mathbb{R}^{w \times h \times c \times \mathcal{C}} \tag{2}$$

where $\mathcal{C}$ represents the number of feature map channels, fm_i represents the i-th feature map, w, h and c represent the width, height and channel of the feature map. $\otimes$ represents concatenate operator. f_{avg} stands for *averaged aggregation feature*, and f_{concat} stands for *concatenated aggregation feature*.

3.2. Branched Layers

HMTNet follows a MTL mode, which indicates that HMTNet can simultaneously perform multiple tasks with a single model. Existing multi-task models often reuse the features in the last layers directly [22, 24, 23]. However, we take a different fashion for branch strategy. Namely, the sub-networks for relative easier tasks are branched out in relative lower layers and embed coarser information, while the sub-networks for difficult tasks are branched out in relative higher layers. The advantages of this strategy is that we can not only form more informative and richer representations, but also reduce the computational burden as well.

3.3. Optimization Object

Gender Recognition. We adopt *GNet* for gender recognition. Gender classification is treated as a binary classification (male VS female) problem in our experiments. GNet is branched out from lower layers, cross entropy is used as the loss:

$$Loss_g = -g \cdot log(\hat{g}) - (1 - g) \cdot log(1 - \hat{g}) \tag{3}$$

where g denotes the groundtruth label, and $\hat{g}$ denotes the predicted value by GNet.

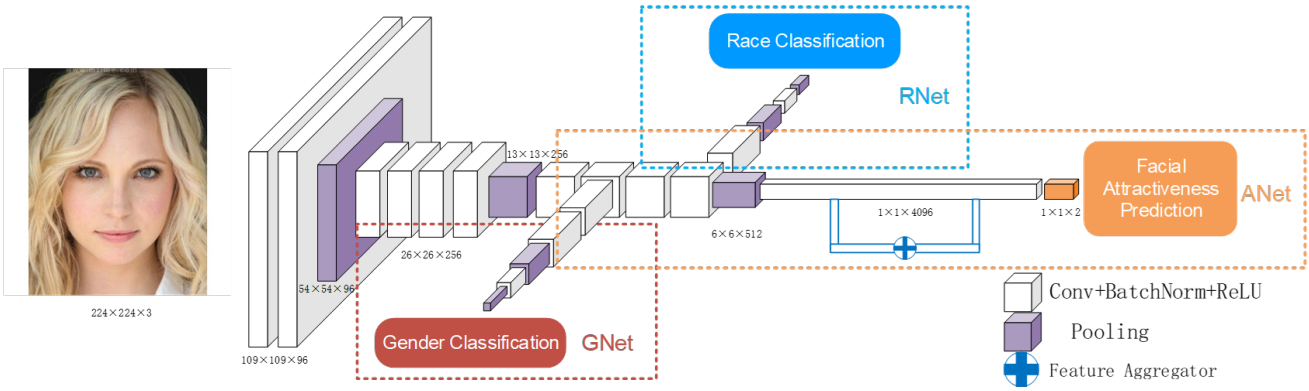

Fig. 1. Overall architecture of HMTNet. RNet (Race Network) and GNet (Gender Network) are used to recognize the race and gender, respectively. ANet (Attractive Net) is utilized to predict the facial attractiveness score. Lower layers can are shared among three sub-networks. All the layers are fully convolutional, and all three branched layers are trained jointly.

Race Recognition. We adopt *RNet* for race recognition task. Softmax loss is adopted as the loss:

$$Loss_r = -\sum_i r_i log(\hat{r}_i) \tag{4}$$

where r_i and $\hat{r}_i$ represent groundtruth label and predicted label of image i, respectively.

Facial Attractiveness Prediction. We adopt *ANet* for facial attractiveness prediction task. We introduce a new loss function for solving FBP task, which is called "Smooth Huber Loss". It follows a *Huber* fashion, but it's smoother when the loss is Smooth Huber Loss, and is more robust to outliers, which is discussed in details later in ablation analysis (4.5).

$$Loss_a = \begin{cases} \sum_i log(\frac{1}{2}(e^{a_i - \hat{a}_i} + e^{\hat{a}_i - a_i})) & if |a_i - \hat{a}_i| \leq \delta \\ \sum_i |a_i - \hat{a}_i| & otherwise \end{cases} \tag{5}$$

where $\hat{a}$ denotes the predicted value, a_i denotes the groundtruth facial beauty score of image x_i. We set $\delta = 0.6$ in our experiments.

Having defined specific loss functions for diverse subtasks. The total loss function is computed as the weighted sum of individual loss functions:

$$Loss_{all} = \sum_{t \in \{g,r,a\}} \alpha_t Loss_t \tag{6}$$

where t denotes t^{th} task in $T = \{g, r, a\}$, the hyper parameters α_t denotes the importance of each task t in the overall optimization object $Loss_{all}$. We set $\alpha_g = \alpha_r = 1$ and $\alpha_a = 2$ in our experiments.

4. EXPERIMENTS

4.1. Datasets and Performance Metric

We conduct experiments on the newly proposed SCUT-FBP5500 [16] and SCUT-FBP [17] datasets to verify the effectiveness of our proposed method. SCUT-FBP5500 [16] contains 5500 portraits with diverse attributes (race, gender and attractiveness scores within a range of $[1, 5]$). Each image is labeled by 60 volunteers, and the average score is used as the groundtruth to remove personal preference bias. In addition, we also perform experiments on SCUT-FBP [17] dataset to verify the feature transferability of multi-task training.

Mean absolute error (MAE), root mean squared error (RMSE) and Pearson Correlation (PC) are used as performance metrics on SCUT-FBP5500 [16]. PC is adopted as measurement on SCUT-FBP [17] dataset. Furthermore, to measure the performance of HMTNet on race and gender recognition tasks, we define Acc_r and Acc_g as the accuracy of race and gender recognition, respectively.

4.2. Implementation Details

We implement our method with PyTorch on NVIDIA P100 GPU. During our experiments, we randomly crop the 224×224 patches from an input image with a shorter size of 227. Besides, random rotation with a angle within $30°$, and color jitter are also applied for data augmentation [18]. We use batch normalization [25] to accelerate model training. HMTNet is trained with SGD algorithm for 170 epochs, weight decay and batch size are set as 0.05 and 32, respectively. The learning rate starts from 0.001 and is divided by 10 per 50 epochs.

4.3. Experimental Results

Table 2 shows the performance comparison with other models. Since SCUT-FBP5500 [16] is a newly released benchmark,

few models have been proposed yet. We compare HMTNet with baseline models reported in [16] and reimplemented CR-Net [26] in FBP with same deep learning framework and parameter settings for fair comparison. HMTNet achieves state-of-the-art performance.

Table 2. Performance comparison on SCUT-FBP5500.

Model	MAE	RMSE	PC
AlexNet [18, 16]	0.2938	0.3819	0.8298
ResNet-18 [19, 16]	0.2818	0.3703	0.8513
ResNeXt-50 [20, 16]	0.2518	0.3325	0.8777
CRNet [26]	0.2835	0.3677	0.8558
HMTNet (Ours)	**0.2501**	**0.3263**	**0.8783**

4.4. Ablation Analysis

Effects of Multi-task Joint Training. Table 3 shows performance on three tasks with or without jointly training, respectively. It demonstrates that multi-task joint training in HMTNet gains performance improvement over three tasks.

Table 3. Evaluation on joint training.

With Joint Training			Without Joint Training		
Acc_r	Acc_g	PC	Acc_r	Acc_g	PC
99.26%	**98.16%**	**0.8783**	98.62%	97.56%	0.8616

Effects of Smooth Huber Loss. We compare our Smooth Huber Loss with other widely used loss functions in FBP task (such as MSE, L_1 loss, Smooth L_1 loss). Other experimental settings are kept unchanged for fair comparison. *Smooth Huber Loss* achieves the best, which reveal the effectiveness of the newly adopted loss function.

Table 4. Evaluation on different loss functions.

Loss Function	MAE	RMSE	PC
MSE Loss	0.2556	0.3372	0.8693
L_1 Loss	**0.2500**	0.3299	0.8753
Smooth L_1 Loss	0.2531	0.3313	0.8738
Smooth Huber Loss	0.2501	**0.3263**	**0.8783**

Effects of Feature Transferability on Multi-task Training. In addition to verifying the effectiveness on SCUT-FBP5500 [16], we also conduct experiments on another dataset [17]. We treat the pretrained HMTNet as a feature extractor (after relu5), and train a simple ridge regressor, our transferred HMTNet achieves state-of-the-art performance on [17] (see Table 5). It indicates that our HMTNet does learn more discriminative representations via multi-task training.

Deep Feature Visualization. We list both precisely predicted and imprecisely predicted images in Fig. 2. Surprisingly and interestingly, HMTNet seems to show more bias on attractive faces since the predicted values of attractive faces

Table 5. Performance comparison on SCUT-FBP.

Methods	PC
Combined Features+Gaussian Reg [17]	0.6482
CNN-based [17]	0.8187
Liu et al. [27]	0.6938
KFME [28]	0.7988
RegionScatNet [5]	0.83
PI-CNN [6]	0.87
CRNet [26]	0.8723
Ours	**0.8977**

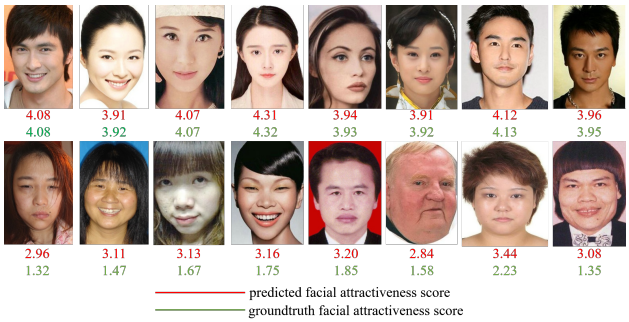

Fig. 2. Precisely predicted (the 1st row) and imprecisely predicted images (the 2nd row) by HMTNet.

are more accurate than those with unattractive faces. From Fig. 3 we can see that eyes play a significant role in facial beauty perception. The fashionable hairstyle also contribute to beauty impression.

Fig. 3. Deep feature visualization learned by HMTNet, lighter denotes higher intensity. Attractive faces (the 1st row); Unattractive faces (the 2nd row). Eyes play a significant part in facial beauty perception.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a novel multi-task network with fully convolutional architecture named HMTNet, to simultaneously recognize a person's gender, race and facial beauty score with very promising results. Detailed exploration about facial beauty perception and feature transferability via multi-task training are also discussed.

6. REFERENCES

[1] David I Perrett, Karen A May, and Sin Yoshikawa, "Facial shape and judgements of female attractiveness," *Nature*, vol. 368, no. 6468, pp. 239, 1994.

[2] Yael Eisenthal, Gideon Dror, and Eytan Ruppin, "Facial attractiveness: Beauty and the machine," *Neural Computation*, vol. 18, no. 1, pp. 119–142, 2006.

[3] Amit Kagian, Gideon Dror, Tommer Leyvand, Daniel Cohen-Or, and Eytan Ruppin, "A humanlike predictor of facial attractiveness," in *NIPS*, 2007.

[4] Rasmus Rothe, Radu Timofte, and Luc Van Gool, "Some like it hot-visual guidance for preference prediction," in *CVPR*, 2016.

[5] Shu Liu, Bo Li, Yang-Yu Fan, Zhe Quo, and Ashok Samal, "Facial attractiveness computation by label distribution learning with deep cnn and geometric features," in *ICME*, 2017.

[6] Jie Xu, Lianwen Jin, Lingyu Liang, Ziyong Feng, Duorui Xie, and Huiyun Mao, "Facial attractiveness prediction using psychologically inspired convolutional neural network (pi-cnn)," in *ICASSP*, 2017.

[7] David Zhang, Fangmei Chen, Yong Xu, et al., *Computer models for facial beauty analysis*, Springer, 2016.

[8] Douglas Gray, Kai Yu, Wei Xu, and Yihong Gong, "Predicting facial beauty without landmarks," *ECCV*, 2010.

[9] Samuel Albanie and Andrea Vedaldi, "Learning grimaces by watching tv," in *BMVC*, 2016.

[10] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[12] Heng Fan and Haibin Ling, "Sanet: Structure-aware network for visual tracking," in *CVPRW*, 2017.

[13] Heng Fan and Haibin Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," in *ICCV*, 2017.

[14] Heng Fan, Xue Mei, Danil Prokhorov, and Haibin Ling, "Multi-level contextual rnns with attention model for scene labeling," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 11, pp. 3475–3485, 2018.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[16] Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li, "Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction," in *ICPR*, 2018.

[17] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li, "Scut-fbp: A benchmark dataset for facial beauty perception," in *SMC*, 2015.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[20] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[21] Gil Levi and Tal Hassner, "Age and gender classification using convolutional neural networks," in *CVPRW*, 2015.

[22] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[23] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[24] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa, "An all-in-one convolutional neural network for face analysis," in *FG*, 2017.

[25] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[26] Lu Xu, Jinhai Xiang, and Xiaohui Yuan, "Crnet: Classification and regression neural network for facial beauty prediction," in *PCM*, 2018.

[27] Shu Liu, Yang-Yu Fan, Zhe Guo, Ashok Samal, and Afan Ali, "A landmark-based data-driven approach on 2.5 d facial attractiveness computation," *Neurocomputing*, vol. 238, pp. 168–178, 2017.

[28] Anne Elorza Deias, "Face beauty analysis via manifold based semi-supervised learning," 2017.