Gender and Smile Classification using Deep Convolutional Neural Networks

Kaipeng Zhang Lianzhi Tan Zhifeng Li Yu Qiao

Shenzhen key lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China

kp.zhang@siat.ac.cn, lz.tan@siat.ac.cn, zhifeng.li@siat.ac.cn, yu.qiao@siat.ac.cn

Abstract

Facial gender and smile classification in unconstrained environment is challenging due to the invertible and large variations of face images. In this paper, we propose a deep model composed of GNet and SNet for these two tasks. We leverage the multi-task learning and the general-to-specific fine-tuning scheme to enhance the performance of our model. Our strategies exploit the inherent correlation between face identity, smile, gender and other face attributes to relieve the problem of over-fitting on small training set and improve the classification performance. We also propose the tasks-aware face cropping scheme to extract attribute-specific regions. The experimental results on the ChaLearn'16 FotW dataset for gender and smile classification demonstrate the effectiveness of our proposed methods.

1. Introduction

Facial gender and smile classification attracts extensive research interests partly due to their increasing numbers of applications. The large visual variations of faces, such as occlusions, pose changes, and extreme lightings, impose great challenge for these tasks in real world applications. Several previous methods [7, 8] treat multiple attributes classification with a single deep network and solve them jointly. However, their models are not face attribute-specific hence their performance on a specific attribute (gender or smile) may be limited. Some methods [4, 5] classify different attributes independently, but they ignore the inherent correlation among smile, gender and other face attributes prediction tasks.

CelebA [7] is a prevalent and large-scale public dataset with forty kinds of face attribute annotations including gender, smile, and other thirty-eight attributes (e.g., wearing hat, wearing glasses and young). The weak visual variations of faces in this dataset are impeditive for training a powerful model directly. FotW is a more challenging dataset with a smaller size, provided by CVPR 2016 Looking at People Challenge [14]. Therefore, we pre-train the models on CelebA [7] and fine-tune on FotW. Besides, we use multi-task and general-to-specific fine-tuning scheme to obtain more discriminative description. Different from General-to-Specific Deep Transfer Learning [9], our method includes coarse-to-fine phases where exploits the inherent correlation between gender, smile and other face attributes.

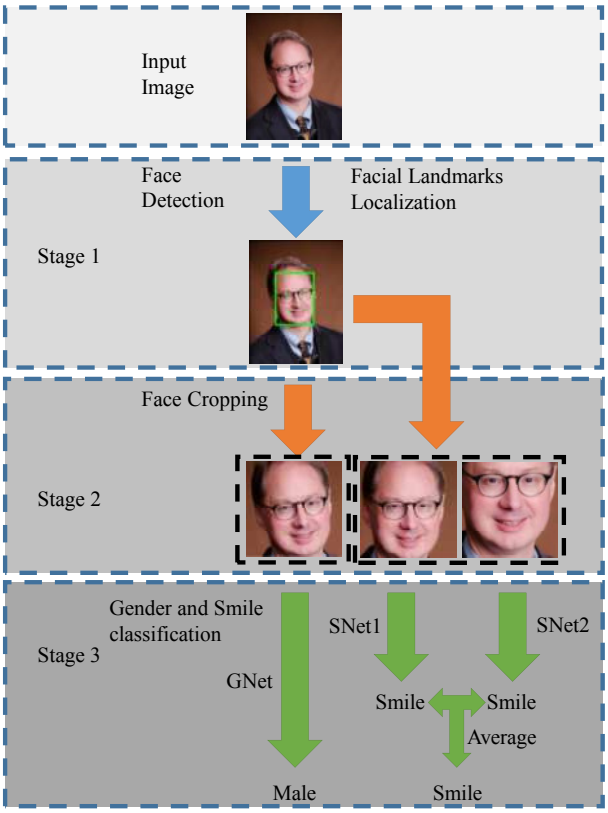

Figure 1: Testing pipeline of our framework. In the first stage, we detect face and facial landmarks from input image. Then we crop face using different cropping schemes. In the last stage, cropped faces are fed to GNet or SNet for gender or smile classification.

In this paper, we propose a deep architecture, composed of two convolutional neural networks (CNNs) GNet and SNet, for facial gender and smile classification tasks. The testing pipeline is showed in Figure 1. We use multi-task and general-to-specific fine-tuning scheme while training

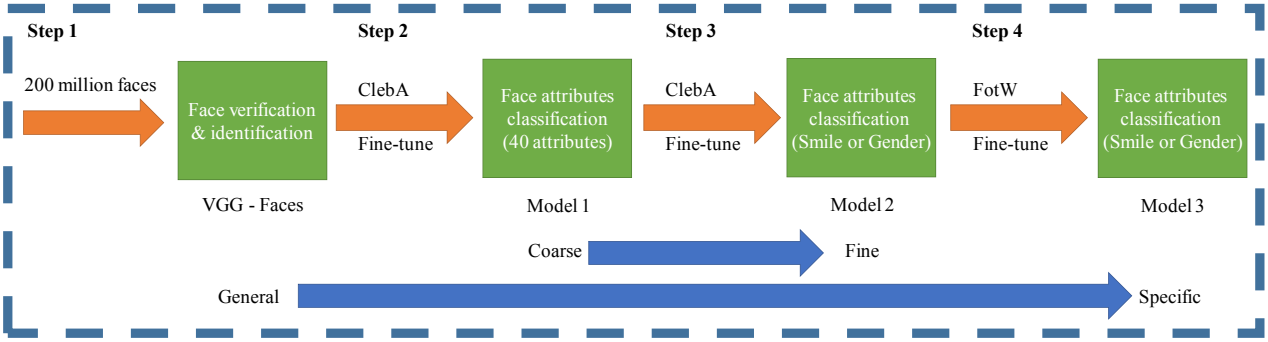

Figure 2. The pipeline of multi-task and general-to-specific fine-tuning scheme. After training a CNN for face identification, there are three steps of fine-tuning to adapt the CNN for recognizing attributes.

our models. Specifically, we use face identification and other face attributes classification supervisory signal for a more powerful CNN. After that, we fine-tune this CNN for a specific face attribute (smile or gender). Besides, we introduce tasks-aware face cropping scheme for better performance. The contributions of this work are summarized as below:

1) We propose a CNNs based framework using multi-task and general-to-specific fine-tuning scheme for facial smile and gender classification.
2) We use tasks-aware face cropping scheme to get better performance.

2. Related Works

In this section, we briefly review related methods for facial gender and smile classification.

Inspired by the good performance of CNNs in computer vision tasks, such as image classification [17] and face recognition [18], several CNNs based smile and gender classification approaches have been proposed in recent years.

For facial gender classification, Fudong Nian *et al.* [1] propose to use CNN for robust gender classification in unconstrained environment. They test their method on the LFWA [7] database and get the state-of-art performance of 98.8% for gender classification. Compared with object classification using a large-scale dataset, Imagenet [10], there are only few small-scale datasets collected from real-world for gender and smile classification. G Levi *et al.* [2] propose a CNN based approach that works well on small dataset.

Facial expression plays an important role in social communications and attracts extensive research interests. Gil Levi *et al.* [3] propose an emotion recognition method on the Emotion Recognition in the Wild Challenge (EmitionW 2015) [13]. They propose a mapped binary pattern and pre-train models on other datasets for different face analysis tasks, such as CASIA-WebFace [11] that leads the model to learn face identification related features. These features are useful for facial expression analysis. For smile classification, P. O. Glauner *et al.* [4] use CNN based method to achieve great a performance on the Denver Intensity of Spontaneous Facial Action (DISFA) [12] and exploit the smile-related region (mouth and cheek).

3. Our Approach

The overall pipeline of our approach is shown in Figure 1. It includes three stages as follows:

Stage 1: Given an image, we initially detect the central face and its five facial landmarks including left eye, right eye, nose, left mouth corner, and right mouth corner using the face locater developed in our previous work [15].

Stage 2: The face is globally cropped to 224 × 224 with three different cropping schemes (see Section 3.3).

Stage 3: One face is fed to GNet for gender classification and the other two faces are fed to two SNets for smile classification. We take the average of outputs (probabilities) form two SNets (SNet1 and SNet2), trained with different face cropping schemes, as final smile classification result.

3.1 Multi-task and general-to-specific fine-tuning scheme

To relieve the over-fitting problem, we propose multi-task and general-to-specific fine-tuning scheme. Figure 2 shows the pipeline of this scheme. Our CNNs' architecture is the same as VGG-Faces [6] shown in Table 1 but with different last fully connected layers (output layers). We train one GNet and two SNnets respectively using different cropping scheme.

Step 1: We adopt VGG-Faces [6] model which is pre-trained on a large-scale face identification dataset for face identification and face verification.

Step 2: We fine-tune VGG-Faces [6] on CelebA [7] with forty attribute annotations including smile and gender. It can exploit the inherent correlation between face identification and smile and gender classification. In addition, it is much better than random initialized parameters. The model

Table 1. The architecture of the VGG-Faces [6] deep convolutional neural network

Layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
Name		conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
Support	-	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
Filt dim	-	3	-	64	-	-	64	-	128	-	-	128	-	256	-	256	-	-	256
Num filts	-	64	-	64	-	-	128	-	128	-	-	256	-	256	-	256	-	-	512
Stride	-	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
Pad	-	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1
Layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
Type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmax
Name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
Support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
Filt dim	-	512	-	512	-	-	512	-	512	-	512	-	-	512	-	4096	-	4096	-
Num filts	-	512	-	512	-	-	512	-	512	-	512	-	-	4096	-	4096	-	2622	-
Stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
Pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

trained in this step is called Model 1.

Step 3: We fine-tune Model 1 on CelebA [7] with a specific attribute (smile or gender). It can exploit the inherent correlation between a specific attribute (smile or gender) and other attributes. Therefore, compared with VGG-Faces, Model 1 can get better performance. Its effectiveness is demonstrated in the Section 4. The model trained in this step is called Model 2.

Step 4: We fine-tune Model 2 on FotW with a specific attribute (smile or gender). FotW is a small-scale dataset but more challenging with large visual variations of faces. The model trained in this step called Model 3.

3.2 Deeply learned face attributes classifier

We use the cross-entropy loss for face attributes classification. For each sample x_i, the loss can be formulated as:

$$L_i^k(\theta) = (y_i^k \log(p_i^k) + (1 - y_i^k)(1 - \log(p_i^k))) \qquad (1)$$

where p_i^k is the probability of k^{th} attribute produced by the network, and θ denotes the parameters of the network. The notation y_i^k denotes the ground-truth label of the k^{th} attribute.

The learning target can be formulated as:

$$\min_\theta \sum_{i=1}^{N_2} \sum_{k=1}^{N_1} L_i^k(\theta) \qquad (2)$$

where N_1 is the number of attributes. N_2 is the number of training examples.

3.3 Tasks-aware face cropping scheme

For gender classification, it needs the global view on a face and face alignment is not necessary. We compared the performance of four kinds of cropping schemes for gender classification and the examples are showed in Figure 3 (b) to (e) where (b) has been aligned and others has been not aligned. In addition, Figure 3 (a) shows the results of face detection and landmarks localization.

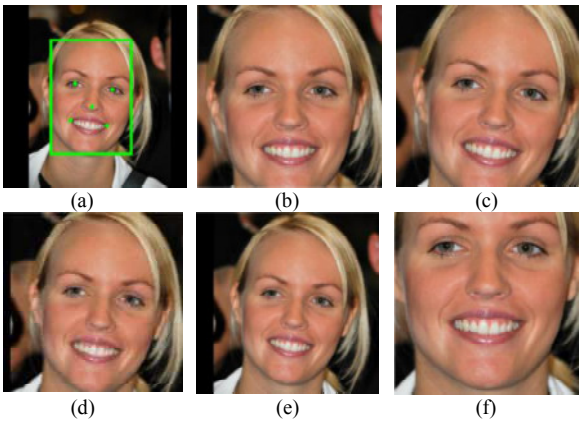

Figure 3. Examples of different cropping schemes for gender and smile classification

For smile classification, it needs not only the global view of a face but also the local regions around mouth and cheek. We compare three kinds of cropping schemes and their individual and ensemble performance. The examples are showed in Figure 3 (b), (c), and (f) where (c) has been not aligned and others have been aligned.

Their effectiveness is demonstrated in the Section 4. According to the results of experiments, cropping overall face without neck and alignment (Figure (d)) is better for gender classification. Besides, the ensemble of cropping overall face and around cheek with alignment (Figure (b) and (f)) is better for smile classification.

Table 2. Comparisons of different fine-tuning stages on the final performance (gender classification).

Training scheme	Gender accuracy
without step 1, 2 and 3	84.04%
without step 2, 3	91.07%
without step 2	91.19%
without step 3	91.30%
with all steps	91.66%

Table 3. Comparisons of different cropping schemes, "+" denotes the model is an ensemble mode. "-" denotes that we skip this experiment.

Cropping scheme	Gender accuracy	Smile accuracy
1	89.34%	88.79%
2	89.66%	88.66%
3	91.66%	-
4	91.21%	-
5	-	88.89%
1+5	-	89.34%
3+4	91.32%	-

3.4 Implementation details

All CNNs are trained using Caffe deep learning toolbox [16]. In every fine-tuning step, the momentum is set as 0.9 and the weight decay is set as 0.0005. The base learning rate is 0.001 in step 2 and it is 0.01 in step 3 and 4. The learning rate is reduced by polynomial with gamma value equals to 0.1. We set lr_multi as 0.1 on all convolutional layers while we set it as 0.5 on fc7 and it as 1 on the last fully connected layers. We set batch size as 120 and total iterations as 110K in all steps.

4. Experiment

In this section, we present the experimental evaluations of our proposed methods. For evaluating convincingly and accurately, we detect all faces from FotW validation set and choose the faces closest to the bounding box provided in each image. It can avoid detecting wrong faces. The collection of these faces is a validation set while training. The training data is collected in the same way. We also correct some definitely wrong annotations in training data. However, we choose the face closest to each image's center while testing. Therefore, the accuracies on FotW validation set while testing or training are different. First, we evaluate the effectiveness of multi-task and general-to-specific fi ne-tuning scheme and tasks-aware cropping scheme in training phase. Then, we evaluate our proposed approach on validation set in testing phase using overall testing pipeline.

4.1. The effectiveness of multi-task and general-to-specific fine-tuning scheme

To evaluate the contribution of multi-task and general-to-specific fine-tuning scheme, we firstly evaluate the performance of four kinds of Model 3 (GNet) as follows: (1) trained without step 1, 2, 3 using 0.01 base_lr. (2) trained without step 2 and 3. (3) trained without step 2. (4) trained without step 3. (5) trained with all steps. Table 2 suggests that the multi-task and general-to-specific fine-tuning scheme is useful for better performance.

4.2. The effectiveness of tasks-aware cropping scheme

We evaluate the performance of different face cropping schemes and their ensemble on FotW validation set in training phase. Cropping scheme 1-5 showed in Figure 3 (b) to (f). According to the results in table2, face alignment is useless for gender classification (scheme 1 and 2) and is useful for smile classification (scheme 1 and 2), so, we skip some experiments about above conclusion. Table 3 suggests that scheme 2 for gender classification and the ensemble of scheme 1 and 5 for smile classification achieve the best performance.

4.3. Evaluation on smile and gender classification

We evaluate our approach using overall testing pipeline on FotW validation set for gender and smile classification. We use the evaluation code provided by organizers and the mean square error is 263.5.

5. Conclusion and future works

In this paper, we proposed a deep convolutional neural network based approach for robust facial gender and smile classification. We propose a multi-task and general-to-specific fine-tuning scheme that exploits the inherent correlation between face identity, gender, smile and other face attributes. Besides, tasks-aware cropping scheme is proposed to further enhance the performance.

For future work, we will exploit the inherent correlation among more face attributes and find an automatic cropping scheme.

Acknowledgement

This work was supported by grants from Natural Science Foundation of Guangdong Province (2014A030313688), Shenzhen Research Program (JSGG20150925164740726, JCYJ20150925163005055, and CXZZ20150930104115529), Guangdong Research Program (2014B050505017 and 2015B010129013), the Key Laboratory of Hman-Machine Intelligence-Synergy Systems through the Chinese Academy of Sciences, and National Natural Science Foundation of China (61103164).

References

[1] F. Nian, L. Li, T. Li, "Robust gender classification on

unconstrained face images," in ACM International Conference on Internet Multimedia Computing and Service, 2015, pp. 77.
[2] G. Levi, T. Hassner, "Age and Gender Classification Using Convolutional Neural Network," in in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 34-42.
[3] G. Levi, T. Hassner, "Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns," in ACM International Conference on Multimodal Interaction, 2015, pp. 503-510.
[4] P. O. Glauner, "Deep Learning For Smile Recognition," arXiv preprint arXiv:1602.00172.
[5] G. Antipov, S. A. Berrani, J. L. Dugelay, "Minimalistic CNN-based ensemble model for gender prediction from face images," Pattern Recognition Letters, vol. 70, pp. 59-65,2015.
[6] O. M. Parkhi, A. Vedaldi, A. Zisserman, "Deep Face Recognition," in British Machine Vision Conference, 2015.
[7] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in IEEE International Conference on Computer Vision, 2015, pp. 3730-3738.
[8] Y. Zhong, J. Sullivan, H. Li, "Face Attribute Prediction Using Off-The-Shelf Deep Learning Networks", arXiv preprint arXiv: 1602.03935.
[9] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, X. Chen, "AgeNet: Deeply Learned Regressor and Classifier for Robust Apparent Age Estimation," in IEEE International Conference on Computer Vision Workshops, 2015, pp. 16-24.
[10] R. Olga, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, F. Li, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, 2015, vol. 115, no. 3, pp. 211-252.
[11] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv preprint arXiv: 1411.7923, 2014.
[12] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, J. F. Cohn, "DISFA: A Spontaneous Facial Action Intensity Database", IEEE Transactions on Affective Computing, vol.4, no. 2, pp. 151-160, 2013.
[13] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in ACM International Conference on Multimodal Interaction, 2015, pp. 423-426.
[14] S. Escalera, M. Torres, B. Mart`ınez, X. Baro, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri, and M. Valstar, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016.
[15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," arXiv preprint arXiv: 1604.02878, 2016.
[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in ACM International Conference on Multimedia, 2104, pp. 675–678.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.
[18] Y. Wen, Z. Li, Y. Qiao, "Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016.