Face Attributes Recognition via Deep Multi-Task Cascade

Ya Li
Guangzhou University
Guangzhou 510006, China
liya@gzhu.edu.cn

Qing Wang
Sun Yat-sen University
Guangzhou 510006, China
wangq79@mail.sysu.edu.cn

Lin Nie
Sun Yat-sen University
Guangzhou 510006, China
nie.lin@foxmail.com

Hui Cheng
Sun Yat-sen University
Guangzhou 510006, China
chenghui9@mail.sysu.edu.cn

ABSTRACT
Predicting face attributes in images is challenging due to complex face variations. We observe that the signals of some attributes are strong in roughly fixed facial area. In this paper, we propose a deep multi-task cascaded network for face attributes recognition, which is trained in an end-to-end way by the combination of attribute-associated region discovery and attribute prediction together. Associated region is mapped by the first sub-network, which is a fully convolutional network (FCN). Then the second sub-network takes the arbitrary-sized associated regions as input. We use RoI pooling to transform them into a uniform-sized region, and train the whole model by stochastic gradient descent (SGD). We also introduce several useful training strategies, including the batch sampling and the early stopping. We adopt sample augmentation for the balance of positive and negative training data in batch sampling. We involve the early stopping to terminate the iterations flexibly for each attribute in case of the model deviation caused by the overfitting of easy attributes. We evaluate our model on two widely used attributes recognition databases, and the experimental results demonstrate that the performance of our approach is better than other state-of-the-art methods on the most of face attributes.

Keywords
Face Attributes Recognition; Multi-Task Learning; Deep Neural Network; RoI Pooling

1. INTRODUCTION
Recognizing face attributes, such as age, gender, expression, and hair style, is very useful in many applications, such as face images retrieval, security surveillance monitoring, human-computer interactive system, and intelligent advertisement system. However, predicting face attributes in images is challenging, because of complex face variations, such as poses, scales, occlusions, and illuminations.

DMCIT '17, May 25-27, 2017, Phuket , Thailand

ACM ISBN 978-1-4503-5218-5/17/05…$15.00
http://dx.doi.org/10.1145/3089871.3089878

We observe that the signals of some attributes are strong in roughly fixed facial area. For example, the information of wearing hat on the top of head is stronger than other area. Only taking the region of top head into account is less noisy than take whole image, so the performance will be better. Therefore, localizing those associated regions of attributes may benefit the predicting. Some local region models also pay their attentions on this problem. For example, the FaceTracer [1] recognized face attributes by extracting hand-crafted features from ten hand-labeling face parts. Kumar et al. [2] predicted face attributes by concatenating low-level features of different face regions. Zhang et al. [3] inferred human attributes by combing part-based models and deep learning, which employed hundreds of poselets [4] for aligning of human body parts. However, these part-based methods are limited by their low-level features, hand- labeling regions, or constrained environment.

To this end, we propose a multi-task cascaded network for face attributes recognition, which discovers attribute associated region and predicts attributes at the same time. This model consists of two deep convolutional neural networks that are responsible for two tasks respectively. As shown in Figure 1, the first sub-network discovers attribute associated region via a weakly supervised learning, i.e. only using the existence information of attribute label, while the second sub-network takes the region of interest (RoI), i.e. the output region of the first sub-network as input, and predicts attribute on it. These networks are designed to share their convolutional features and trained in an end-to-end way. In Figure 1, the connected convolutional layers are designated "CONVs" for conciseness. Multitask Learning usually learns tasks in parallel using a shared representation and aim to help other tasks be learned better by what is learned from each task [5]. Our method is different from them. In our method, the second task depends on the output of the first one; two tasks are in a cascaded form.

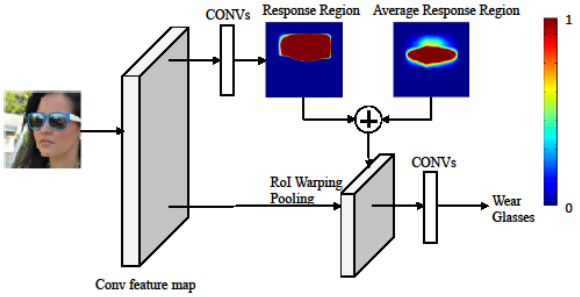

Figure 1. The proposed pipeline of multi-task cascaded attribute inference.

Associated region is mapped by a fully convolutional network (FCN) [6], which produces region response maps directly from raw images without relying on any pre-process. FCNs are designed to predict a category label for each pixel. The filtering value by convolution shows the associated degree, we call it response value. The larger the response value, the greater the likelihood of an attribute exist. We map the point which response values larger than certain threshold back to original image, and the minimal rectangle is looked as the associated region.

Then the second sub-network takes the arbitrary-sized output region of task one as input for further recognition, so it is necessary to transform the region into a uniform-sized. We use RoI pooling [7] to reshape the region, which produces a fix-sized patch by a bilinear interpolation function. We train the whole model by stochastic gradient descent (SGD) and use the Caffe library [8].

The main contributions of this work are summarized as follows. (1) We propose a novel deep multi-task cascade model for face attributes recognition. Two cascaded tasks are trained in an end-to-end way. To the best of our knowledge, it is the first work to recognize face attributes based on the learned associated region. We train the proposed model using a novel RoI pooling operation to transform the associated region into fixed size. And several carefully designed training strategies are introduced, such as batch sampling and early stopping mechanism. (2) We evaluate our model on LFWA and CelebA databases, and the experimental results demonstrate the effectiveness of our approach.

2. RELATED WORK

The attributes recognition is always the hot research issue in the vision literature, such as face attributes and human attributes. For face attributes recognition, the early researches focus on single or few attribute by hand-crafted features. Hand-crafted features like principal component analysis (PCA) [9], histogram of oriented gradient (HOG) [10], local binary pattern (LBP) [11], and Gabor wavelet, discrimination methods like support vector machine (SVM), Adaboost, linear discriminant analysis (LDA) and their variations are widely used in the face attributes recognition, such as gender [12], race [13], expression [14], and age [15]. Kumar et al. [2] extracted HOG-like features on different face regions for face attributes recognition. Guo et al. [16] estimated age, gender and race jointly by canonical correlation analysis (CCA). Bourdev et al. [17] extracted higher-level information by a three-level SVM system.

Instead of using hand-crafted features, deep CNN learns features from raw images directly, and it has made impressive progress in image classification, object detection, semantic segmentation, face recognition, and many other vision tasks. Recently, more and more researchers have involved in multi-attributes recognition by using the deep learning methods. Zhang et al. [3] showed the substantial improvement on attribute classification tasks by training pose-normalized CNNs. Luo et al. [18] proposed a sum-product network for fifteen attributes recognition. Li et al. [19] divided the face into 6 parts based on 36 landmarks for age and gender recognition. Zhao et al. [20] presented a peak-piloted deep network for facial expression recognition, which trained by peak gradient suppression, that is, only kept the gradients to the features of the non-peak expression. Liu et al. [21] predicted abundant face attributes by deep network without using face or landmark detector.

3. OUR APPROACH

3.1 Network Architecture

Our multi-task cascaded model consists of two sub-networks; they are responsible for associated region mapping and attributes recognition tasks. The detail is shown in Figure 2, which is realized using Caffe model. The upper and lower parts correspond to the realization of two tasks respectively. The first two convolutional layers are shared by all attributes. For task one, first five layers are similar to AlexNet [22], after that, two more convolutional layers are appended. The square in conv3 is the interesting region, which is mapped by conv7, the last response layer. Task two conducts the RoI pooling on the associated region first, and then employs a global max-pooling after conv7. At last, the outputs of these two tasks are concatenated together as a vector, and we finally obtain the attribute prediction by making the vector multiply the parameters and add the bias. The model parameters are shown in table 1. Pre-training the network of task one till the model is convergent before the jointly training with task two.

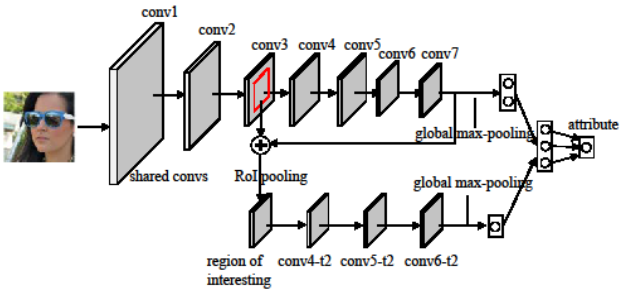

Figure 2. The network architecture.

Table 1. The network parameters

	Layer Type	Channel No.	Kernel Size	Stride	Padding	Output Size
All attributes shared	conv1	96	11	4	0	54
	pool1	96	3	2	0	27
	conv2	256	5	1	2	27
	pool2	256	3	2	0	13
Two task shared	conv3	384	3	1	1	13
Task one	conv4	384	3	1	1	13
	conv5	256	3	1	1	13
	conv6	64	3	2	0	6
	conv7	2	3	1	0	4
Task two	conv4-t2	128	3	1	1	5
	conv5-t2	64	3	1	1	5
	conv6-t2	1	1	1	1	5

3.2 Associated Region Mapping

The first sub-network produces region response map directly from raw images via FCNs, and the associated region is learned by only using the attribute label as weak supervised information due to the absence of ground truth region information. The destination is making the response values as large as possible if attribute label is 1, while making it as small as possible if attribute label is 0. Therefore we define the loss function as:

$$\ell = \min \sum_i \max_j (1(y_j = 1) r_j(x_i) + 1(y_j = 0)(1 - r_j(x_i))), \quad (1)$$

where $r_j(x_i)$ is the j-th output of conv7 for sample x_i. Suppose θ is all the network parameters to be optimized, and the associated region R is function of θ. Eqn. (1) can be simplified as:

$$L_1 = L_1(R(\theta)). \quad (2)$$

The associated region is obtained by mapping the output points, those response values of conv7 lager than threshold ε, back to the original image, that is, computing the receptive field. Figure 3 illustrates the relationship between the receptive fields and the outputs. For example, there are two convolution layers with two kernels of 5×5 and 7×7 receptively. The receptive field of a response 1×1 in layer 3 is a patch with size 7×7 in layer 2, and

$$L = \alpha L_1(R(\theta)) + \beta L_2(P(\theta) \mid R(\theta)), \quad (6)$$

the receptive field of the 7×7 patch in layer 2 is a patch with size 11×11 in layer 1. For convolutional layer and pooling layer, the receptive field can be obtained by $r_i = s_i \times (r_{i-1} - 1) + k_i$, where r_i denotes the receptive field in i-th layer, s_i is the stride, k_i is he kernel size. For ReLU layer, the receptive field is computed by $r_i = r_{i+1}$.

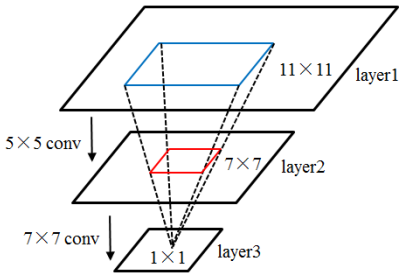

Figure 3. An example of receptive field.

The associated region is used as input for task two directly, if the region location is deviated from the truth, attribute recognition will be wrong. In order to avoid this problem and increase the reliability of location, we adopt the average associated region for revision. We train the first sub-network standalone, then input abundant aligned face images with same attribute into it, and the average of these outputs is used as the average associated region.

3.3 Region Attributes Recognition

The second sub-network takes the shared convolutional features and the associated region as input. Firstly, we reshape the input region into same size using the RoI pooling. It can be formulated as:

$$F^{RoI}(\theta) = G(R(\theta))F(\theta). \quad (3)$$

Here G represents the cropping and warping operations, which is the bilinear interpolation function, and can transform a proposed region from size of $w\times h$ to size of $w'\times h'$. A $R(\theta)$ is the region of interesting, $F(\theta)$ is the feature map and F^{RoI} is the RoI warping output. The RoI warping operation performs on each channel respectively. There is a max-pooling operation after the RoI warping, so we name this layer as RoI pooling layer. After RoI pooling, then we append 3 convolutional layers and a global max-pooling layer for attribute prediction. The cross-entropy loss function is used for attribute prediction. It is formulated as:

$$H(x, y) = \sum_i y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i)), \quad (4)$$

where (x_i, y_i) is the i-th training sample with label y, $p(x_i)$ represents the posterior probability of x_i having this attribute. And $p(x_i)$ is formulated as $p(x_i) = \frac{1}{1 + e^{-f(x_i)}}$, $f(x_i)$ denotes the output value of image x_i. Eqn. (4) can also be simplified as the function of the network parameters θ:

$$L_2 = L_2(P(\theta) \mid R(\theta)). \quad (5)$$

Two tasks in our model are associated, where the second task depends on the first task's output. And all of the operations are differentiable, therefore, we can train the model in an end-to-end way. The whole loss function is shown below.

where α and β are the weights of two tasks.

3.4 Model Training Strategies

There are several useful training strategies for our model, including the batch sampling and the early stopping.

3.4.1 Batch Sampling

The distributions of positive and negative samples for a same attribute are very different. For example, the number of male and female is almost same for gender. But for bald, goatee, and wearing hat/glasses, it is significant different. The number of negative samples is much larger than that of negative samples. If we directly used them for training, it will result in the deviation of model due to the scarcity of certain type samples. Therefore, it is necessary to balance the positive and negative samples in a batch. To this end, we adopt samples augmentation and proportion balancing. Samples augmentation is carried out on images having few instances for certain attribute. We crop patches with same size on original image. Suppose the original size is $H\times W$, we crop patches those taking the points that distance to the original center in the range of $[-0.05H, 0.05H]\times[-0.05W, 0.05W]$ as new center, and then reflect these patches horizontally for double augmentation. For proportion balancing, we fix the proportion of positive and negative numbers as 1:1 in a batch.

3.4.2 Early Stopping

If all of the attributes training terminated at the same iterations, it tends to result in over-fitting for easy attributes while still is not convergent for hard attributes. Furthermore, the over-training on some irrelevant attributes also makes the representation ability of shared features declined. Therefore, we involve the early stopping to terminate the iterations. Evaluate the loss per mini-batch, and find the threshold μ as the termination condition when the losses are vibrated in a certain range. The gradient doesn't pass back for propagation while the loss is lower than μ, and training on current attribute stop accordingly.

4. EXPERIMENTS

We evaluate the performance of our approach on two different databases: LFWA and CelebA [21]. LFWA has 13, 233 images of 5, 749 identities. CelebA contains 10 thousand identities, each of which has 20 images. There are 200 thousand images totally. Each image in LFWA and CelebA is annotated with 40 face attributes and 5 landmarks. Following the settings of paper [21], we divide the CelebA into 3 parts, 80% is training set, 10% is validation set for parameters determining, and the rest of 10% is testing set. For the LFWA, half images for training and halt for testing.

The proposed method is compared with FaceTracer [1], PANDA [3], and ANet+LNet [21]. PANDA has two settings, PANDA-w and PANDA-1 like paper [21] considered. The experimental results of FaceTracer and PANDA are come from paper [21]. ANet+LNet extracts learned-features by three cascaded CNNs and adopts SVM for attributes recognition.

We train the model using a PC with Core i7-3770K 3.50 GHz CPU, Nvidia GeForece Titan X GPU and 16GB memory.

We first show the performance of attribute prediction of our model. The prediction accuracies are reported in Table 2, where twenty attributes we think they are more meaningful than others are shown. On CelebA, the average accuracies of FaceTracer, PANDA-w, PANDA-1, ANet+LNet and our method are 81, 79, 85, 87 and 90.2 respectively, while the corresponding accuracies on LFWA are 74, 71, 81, 84 and 87.8 percent. Our method outperforms ANet+LNet by 3.7% and 4.5% on CelebA and LFWA respectively.

Then we demonstrate the effectiveness of our model. To show how the associated region improves the accuracies, we finish the experiment only using the whole face image without the associated region location sub-network. The average accuracies on CelebA and LFWA are 89 and 86.3 respectively. The associated region location improves the accuracies by 1.3% and 1.7% correspondingly.

Table 2. Performance comparison

		Attractive	Bald	Black Hair	Blond Hair	Brown Hair	B. Eyebrows	Eyeglasses	Goatee	Heavy Markup	Male	Mouth S. Open.	Mustache	Oval Face	Smiling	Straight Hair	Wavy Hair	Wear. Earrings	Wear. Hat	Wear. Necklace	Young	Average
CelebA	Face Tracer	78	89	70	80	60	80	98	93	85	91	87	91	64	89	63	73	73	89	68	80	81
	PANDA-w	77	92	74	81	69	76	94	86	84	93	82	83	62	89	67	76	72	91	67	77	79
	PANDA-1	81	96	85	93	77	86	98	93	90	97	93	93	65	92	69	77	78	96	67	84	85
	ANet+LNet	81	**98**	88	**95**	80	**90**	99	95	90	**98**	92	95	66	92	73	80	82	**99**	71	**87**	87
	Ours	**82**	**98**	**88.4**	94.9	**83.5**	89.6	**99.6**	**97**	**91.3**	97.8	**93.7**	**96.3**	**75.1**	**92.2**	**81.8**	**84.9**	**87.6**	**99**	**84.9**	86.1	**90.2**
LFWA	Face Tracer	71	77	76	88	62	67	90	69	88	84	77	83	66	78	67	62	88	75	81	80	74
	PANDA-w	70	82	78	87	65	63	84	65	86	86	74	77	64	77	68	63	85	78	79	76	71
	PANDA-1	81	84	87	94	74	79	89	75	93	92	78	87	72	89	73	75	92	82	86	82	81
	ANet+LNet	83	88	90	**97**	77	82	95	78	**95**	94	82	92	74	91	**76**	**76**	**94**	88	88	86	84
	Ours	**85.5**	**95.1**	**91.1**	95.9	**79.1**	**84.3**	**98.9**	**80**	94.9	**95.1**	**83.6**	**92.8**	**78.5**	**91.6**	73.5	75.7	92.9	**92.3**	**88.2**	**87.7**	**87.8**

Recall that we adopt the average associated region for better accuracy. The average associated region comes from the output of abundant aligned face images passing through the pre-trained network. We think the associated region is more reliable after the revision by attribute's average associated region. Figure 4 demonstrates the effectiveness of it on eyeglasses and male attributes. For figure 4, (a) is the original image; (b) is the response image without using average associated region; (c) is associated region mapping; (d) is the response image using average associated region; (e) is the revised associated region mapping. It is obvious that the location of revised associated region is more accurate.

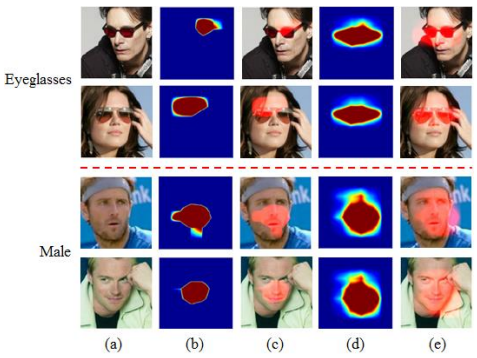

Figure 4. The effectiveness of average associated region.

5. CONCLUSION

This paper has proposed a novel deep multi-task cascaded network for face attributes recognition. It discovers the potential associated region of attribute and predicts attribute at the same time. Experiments demonstrate the effectiveness of this model. The performance of our approach is better than other state-of-the-art methods on most of 20 face attributes. We have also introduced some important training strategies on model learning. We believe that our method can be easily transformed into other recognition problems that are associated with certain region on computer vison area. And we will pay our attentions to it in future.

6. ACKNOWLEDGMENTS

This research is supported by the Research Project of Guangzhou Municipal Universities (No. 1201620302), the Science and Technology Planning Project of Guangdong Province (No. 2015B010128009, 2013B010406005). The authors would like to thank the reviewers for their comments and suggestions.

7. REFERENCES

[1] Kumar, N., Belhumeur, P. N., and Nayar, S. K. 2008. Facetracer: A search engine for large collections of images with faces. In *Proceedings of* 10th European Conference on Computer Vision (Marseille, France, Oct. 12-18). ECCV'08, Springer Verlag, Heidelberg, Germany, 340–353, 2008. DOI= https://dx.doi.org/10.1007/978-3-540-88693-8-25.

[2] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. 2009. Attribute and simile classifiers for face verification. In *Proceedings of the 12th IEEE International Conference on Computer Vision* (Kyoto, Japan, Sep. 29-Oct.02, 2009). ICCV'09. IEEE, NJ, USA, 365-372, 2009. DOI= https://dx.doi.org/10.1109/ICCV.2009.5459250.

[3] Zhang, N., Paluri , M., Ranzato, M., Darrell, T., and Bourdev , L. 2014. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the 27th IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Columbus, OH, United states, June 23-28). CVPR'14. IEEE, NJ, USA, 1637-1644, 2014. DOI= https://dx.doi.org/10.1109/CVPR.2014.212.

[4] Bourdev, L., Maji, S., and Malik, J. 2011. Describing people: A poselet-based approach to attribute classification. In *Proceedings of the 13th IEEE International Conference on Computer Vision* (Barcelona, Spain, Nov. 6-13). ICCV'11. IEEE, NJ, USA, 1543-1550, 2011. DOI= https://dx.doi.org/10.1109/ICCV.2011.6126413.

[5] Caruana, R. (1997). Multi-task learning. *Machine Learning*. 28, 1(July 1997), 41–75. DOI= http://dx.doi.org/10.1023/A:1007379606734.

[6] Long, J., Shelhamer, E., and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the 28th IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA, United states, June 7-12) CVPR'15. IEEE, NJ, USA, 3431-3440, 2015. DOI= https://dx.doi.org/10.1109/CVPR.2015.7298965.

[7] Dai, J., He, M., Sun, J. 2016. Instance-aware Semantic Segmentation via Multi-task Network Cascades. In *Proceedings of the 29th IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, United states, June 26-July 1). CVPR'16. IEEE, 3150-3158, 2016. DOI= https://dx.doi.org/10.1109/CVPR.2016.343.

[8] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of ACM International Conference on Multimedia* (Orlando, FL, United states, Nov. 3-7). MM'14. ACM, 675-678, 2014. DOI= https://dx.doi.org/10.1145/2647868.2654889.

[9] Turk, M., and Pentland, A.1991. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 1991. 3(1), 71–86.

[10] Dalal N and Triggs B. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Diego, CA, United states, June 20-25), CVPR'05. IEEE, 886-893, 2005. DOI= https://dx.doi.org/10.1109/CVPR.2005.177.

[11] Ahonen, T., Hadid, A., Pietikainen, M. 2006. Face description with local binary patterns: application to face recognition. *IEEE Trans Pattern Anal. Mach. Intell*. IEEE. 28(12), 2037-2041, 2006. DOI= https://dx.doi.org/10.1109/TPAMI.2006.244.

[12] Lian, H., and Lu, B. 2006. Multi-view gender classification using local binary patterns and support vector machines. In *Proceedings of 3rd International Symposium on Neural Networks* (Chengdu, China, May 28-June 1), ISNN'06. Springer, 202-209, 2006. DOI= https://dx.doi.org/10.1007/11760023_30.

[13] Xie, Y., Luu, K., and Savvides, M. A robust approach to facial ethnicity classification on large scale face databases. In *Proceedings of Biometrics: Theory, Applications and Systems* (Arlington, VA, United states, Sep. 23-27). BTAS'12. IEEE, Washington, DC, USA, 143-149, 2012. DOI= https://dx.doi.org/10.1109/BTAS.2012.6374569.

[14] Zeng, Z., Pantic, M., Roisman, G., Huang, TS. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans Pattern Anal. Mach. Intell*. IEEE. 31(1), 39-58, 2009. DOI= https://dx.doi.org/10.1109/TPAMI.2008.52.

[15] Geng, X., Zhou, ZH., and Smith-Miles, K. 2007. Automatic age estimation based on facial aging patterns. *IEEE Trans Pattern Anal. Mach. Intell*. IEEE. 29(12), 2234-2240, 2007. DOI= https://dx.doi.org/10.1109/TPAMI.2007.70733.

[16] Guo, G., and Mu, G. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *Proceedings of 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* (Shanghai, China, Apr. 22-26). FG'13, IEEE, Washington, DC, USA, 1-6, 2013. DOI= https://dx.doi.org/10.1109/FG.2013.6553737

[17] Bourdev, L., Maji, S., and Malik, J. 2011. Describing people: A poselet-based approach to attribute classification. In *Proceedings of the 13th IEEE International Conference on Computer Vision* (Barcelona, Spain, Nov. 6-13). ICCV'11. IEEE, NJ, USA, 1543-1550, 2011. DOI= https://dx.doi.org/10.1109/ICCV.2011.6126413.

[18] Luo, P., Wang, X., and Tang, X. 2013. A deep sum-product architecture for robust facial attributes analysis. In *Proceedings of the 14th IEEE International Conference on Computer Vision* (Sydney, Australia, Dec 1-8). ICCV'13 IEEE, NJ, USA, 2864-2871, 2013. DOI= https://dx.doi.org/10.1109/ICCV.2013.356.

[19] Li, S., Xing, J., Niu, Z., Shan, S., and Yan, S. 2015. Shape Driven Kernel Adaptation in Convolutional Neural Network for Robust Facial Trait Recognition. In *Proceedings of the 28th IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, MA, United states, June 7-12) CVPR'15. IEEE, NJ, USA, 222-230, 2015. DOI= https://dx.doi.org/10.1109/CVPR.2015.7298618.

[20] Zhao, X., Liang, X., Liu, L., Li, T., Han, Y., Vasconcelos, N., Yan, S. 2016. Peak-Piloted Deep Network for Facial Expression Recognition. In *Proceedings of* 14th European Conference on Computer Vision (Amsterdam, Netherlands, Oct. 8-16). ECCV'16. Springer, 425-442, 2016. DOI= https://dx.doi.org/10.1007/978-3-319-46475-6_27.

[21] Liu, Z., Luo, P., Wang, X., Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the 15th IEEE International Conference on Computer Vision* (Santiago, Chile, Dec. 11-18). ICCV'15. IEEE, NJ, USA, 3730-3738, 2015. DOI= https://dx.doi.org/10.1109/ICCV.2015.425.

[22] Krizhevsky, A., Sutskever, I., Hinton, G E. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems* (Lake Tahoe, USA, Dec. 3-6). NIPS'12. Canada, 1097-1105, 2012.