MULTI-LABEL NETWORKS FOR FACE ATTRIBUTES CLASSIFICATION

Sara Atito Aly *Berrin Yanikoglu*

Sabanci University
Istanbul Turkey 34956
{saraatito,berrin}@sabanciuniv.edu

ABSTRACT

Face attributes classification is drawing attention as a research topic with applications in multiple domains, such as video surveillance and social media analysis. In most attribute classification systems in literature, independent classifiers are trained separately for each attribute. In this work, we propose to train attributes in groups based on their localization (head, eyes, nose, cheek, mouth, shoulder, and general areas) in a multi-task learning scenario to speed up the training process and to prevent overfitting. We have evaluated the idea of using the location knowledge for a particular attribute group to speed up the network training. Attention is drawn to the area of interest by blurring training images outside the region of interest, fine-tuning the system and freezing the earlier layers before continuing training with original images. Several data augmentation techniques are also performed to reduce overfitting. Our approach outperforms the state-of-the-art of the attributes on the public LFWA dataset, with an average improvement of almost 0.7% points. The accuracy ranges from 78% (detecting oval face or shadow on the face) to 97.4% (detecting blond hair) across the attributes.

Index Terms— Face Attributes, Deep Learning, Transfer Learning, Multi-Label classification, Data Augmentation

1. INTRODUCTION

Detecting facial attributes, such as hair style, gender, and smile, is very beneficial in large scale applications [1] like face recognition and identification [2], face verification [3, 4], and image understanding [5]. However, being able to automatically describe face attributes from images is a challenging task, as real-life images have different illuminations, occlusions, poses and background variations.

Automatic recognition of face attributes became an active research topic, especially with the release of CELEBA and LFWA attribute datasets with more than 200,000 images, each with 40 attribute annotations, by Liu et al. [6].

The general pipeline of face attribute classification can be summarized as follows: (1) Face localization; (2) Feature extraction; (3) Attributes classification. Face localization is outside the scope of this paper, as we work on aligned images. Feature extraction and classification have been addressed separately in the past [7, 4], while newer approaches based on deep learning and especially Convolutional Neural Networks (CNNs) address both problems at once.

In spite of the fact that valuable information can be obtained from the correlation of attributes, most of the state-of-the-art methods are dealing with attributes independently. In this paper, we approached this task in a Multi-Task Learning (MTL) scenario by grouping attributes based on their localization and sharing weights of each group of attributes, also suggested in [8, 9]. Grouping attributes not only reduced number of needed classifiers to classify 40 different attributes, but also sharing weights helped reducing overfitting. We also speed up the training by indicating the area of interest for a group of attributes (e.g. mouth region for smile and wearing lipstick attributes, in a two-stage learning. The main contributions of this paper are as follows:

i) Proposing a state-of-the-art approach for face attribute classification, using the Multi-Task Learning framework and various forms of data augmentation in order to reduce overfitting. Our results are evaluated on a well known dataset (LFWA), obtaining an average improvement of almost 0.7% points and maximum relative improvement of 3.77% over the state-of-the-art.
ii) Suggesting a simple method for passing prior information about the general location of an attribute group, to direct network's attention in order to speed up convergence. We show that the two-stage training (with first blurred images and then original) is both faster and slightly more accurate (Fig. 4).

2. RELATED WORKS

Until recent years, facial attributes classification has been addressed with handcrafted representations, as in [7, 4, 10]. This kind of approaches may fail with unconstrained background and different variations of face images. More recently, researchers tackle this task using deep learning, which has resulted in huge performance leaps in several domains [11, 9, 6, 12, 13, 14, 15].

In Zhu et al. [12] and Razavian et al. [13], CNNs are

978-1-5386-1737-3/18/$31.00

used to extract features from landmarks to train independent classifiers for each attribute. This approach requires an accurate landmarks detection. Liu et al. [6] use two cascaded convolutional neural networks, for face localization (LNet) and attributes prediction (ANet), replacing the last fully connected layer with a support vector machine classifier. Each attribute classifier was trained separately. Similarly in Zhong et al. [11], attribute prediction is accomplished by leveraging different levels of CNNs. Hand and Chellapa's work [9] is the most similar to ours: they divide the attributes into nine groups and train a CNN consisting of three convolutional sub-networks and two multi-layer perceptrons. The first two convolutional sub-networks are shared for all of the classifiers (representing earlier and shared features) and the rest of the network is independent for each group. They also compare their results to the results of classifiers trained independently for each attribute and show the advantage of grouping attributes together.

3. METHOD

Most of the existing work on face attributes classification ignores the relationship between different facial attributes, and trains individual classifiers for each attribute separately. In this work, we propose to train attributes in groups based on their localization (head, eyes, nose, cheeks, mouth, shoulder, and general areas) in a multi-task learning scenario, to speed up the training process and to prevent overfitting. The area of interest for a particular attribute group is indicated by blurring the image outside the attribute group region, based on the mean image of the training set. In our case, 40 different attributes are considered and divided into 7 groups (Table 1).

3.1. Network Architecture and Training

Training a large deep learning network from scratch is time consuming and needs tremendous amount of training data. Therefore, our approach is based on fine-tuning a pre-trained model, namely the VGG19 network [16] which is the winning architecture of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014. VGG19 is trained on a dataset with 1.2 million hand-labeled images of 1,000 different object classes. Its architecture involves 16 convolution layers, five pooling layers and three fully-connected layers.

As we consider the problem as a multi-task learning problem, the output layer is changed to represent the labels in each attribute group and the loss function is replaced with a multi-label sigmoid Loss. For a single image I with A attributes, the cross-entropy error is denoted as shown in Equation 1:

$$E(I) = \sum_{a=1}^{A} -y_I[a] \times \hat{y}_I[a] + log(1 + exp(\hat{y}_I[a])) \quad (1)$$

where $y_I[a]$ and $\hat{y}_I[a]$ are the target and output of image I indexed by attribute a, respectively.

Group	Attributes
Head	Black Hair, Blond Hair, Brown Hair, Gray Hair, Bald, Bangs, Straight Hair, Wavy Hair, Receding Hairline, Hat
Eyes	Arched Eyebrows, Narrow Eyes, Bushy Eyebrows, Bags Under Eyes, Eyeglasses
Nose	Big Nose, Pointy Nose
Cheek	5 O-clock Shadow, Rosy Cheeks, Goatee, High Cheekbones, No Beard, Sideburns
Mouth	Big Lips, Smiling, Mustache, Wearing Lipstick, Mouth Slightly Open
Shoulder	Double Chin, Wearing Necklace, Wearing Necktie
General	Attractive, Blurry, Chubby, Young, Male Pale Skin, Oval Face, Heavy Makeup, Earrings

Table 1: Grouping attributes based on their relative location.

Multi-Task learning has already shown a significant success in different applications like face detection, facial landmarks annotation, pose estimation, and traffic flow prediction [17, 18, 19, 20]. MTL is mainly applied by sharing all of the hidden layers between the given tasks but with different output layer for each task. As shown in [21], sharing weights for multiple tasks acts as a regularizer that help reducing the risk of overfitting. Intuitively, the model is forced to learn a general representation that captures all of the specified tasks which less the chance of overfitting.

We used the VGGNet models provided in the CAFFE deep learning framework [22]. Throughout this work, we set the batch size equal to 20 with iteration size equal to 2 and the initial learning rate as 10^{-3} with a total of $1K$ iterations for stage 1 and $10K$ iterations for stage 2.

In order to speed up the training and concentrate the feature extraction process into a local region, the training process of each group of attributes is completed in two stages: (1) directing the attention of the network to the area of interest by first training with blurred images outside the area of interest (Sec. 3.2); and (2) and freezing early layer weights and fine-tuning the system using the original dataset (Sec. 3.3).

3.2. Stage 1: Directing Attention

Training a huge convolutional neural network with a small dataset, especially if ground-truth labels are noisy, requires thousands of iterations to obtain a good representation from the region of interest (ROI). Automatic attention mechanisms have attracted interest in recent years, with the goal of focusing on a small part of the input or attending to past input in a recurrent network [23]. Our goal is simply to direct attention by indicating a small amount of prior information to the network, in order to speed up the convergence. We indicate the location information of a group of attributes to the network

by blurring the images outside the ROI, so as to extract most of the features within the desired region. The early weights learned in this stage are then fixed in the next stage.

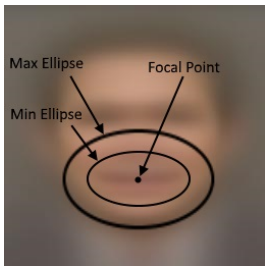
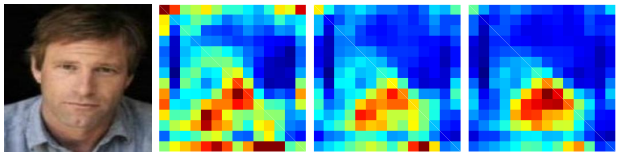
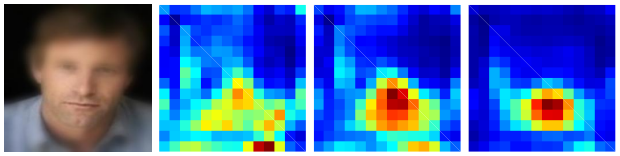


Fig. 1: Stage 1 for the mouth region: The region outside the ROI is blurred, as defined by the min and max ellipses whose center is detected on the mean training image.



Original Image 1000 iterations 3000 iterations 5000 iterations

(a) Extracted features using original training images.



Apply Attention 100 iterations 500 iterations 1000 iterations

(b) Extracted features using attention mechanism.

Fig. 2: Comparison of training the network a) directly with original training images or (b) by directing attention with blurred images.

Training images are pre-processed by convolution with an elliptical 2D Gaussian kernel centered on the region of interest, outside the ROI itself, as shown in Figures 1. The center and the size around the ROI are defined based on the mean image of the dataset. Furthermore, dataset augmentation is also achieved by changing the strength of blur and the size of the ellipse between pre-defined minimum and maximum, as shown in Figure 1.

The system input is an image resized to 256×256 and blurred as described above. Then, it undergoes internal data augmentation and gets cropped to 224×224, according to the input layer size of VGG19. The pre-trained VGG19 network is then fine-tuned using the blurred images for 1,000 iterations.

Figure 2 shows the summation of the last convolutional layer outputs after different number of iterations by training the network with original images directly (Figure 2a) and training the network with pre-processed images by focusing on the region of interest (Figure 2b). Neural activations show that the focus of the network is tuned mostly to the region of interest by the end of Stage 1.

In the second stage, we freeze the early layer weights from this stage and fine-tune the rest of the network using original images. In Section 4 we compare this approach to fine-tuning with original or blurred images in one stage. Our results show that the network learns much faster in our case, as well as having a slightly higher accuracy.

3.3. Stage 2: Fine Tuning

In this stage, the VGG19 network is fine-tuned by continuing the back-propagation starting from the trained model coming from Stage 1, but by freezing the weights of low-level portion of the network (10 convolutional layers) and using the original images. The learning rate of the rest convolution layers are reduced by factor of 10 to keep learning but sustaining the extracted features from stage 1. Thus, the features that lie outside of the region of interest but might be helpful in classifying the current group of attributes (e.g. eye features being used in smile detection) can be considered.

For data augmentation, we used both internal and external augmentation. For external augmentation, all augmented data are generated before training where several augmentation techniques are used as shown in Section 4.2. For internal augmentation, each input image is augmented by random cropping and random horizontal flipping, provided optionally in the CAFFE framework [22].

4. EXPERIMENTS

4.1. Dataset

The LFW [24] dataset is used to assess our proposed method. Originally, the dataset is constructed for face identification and verification, while recently, it is annotated with 40 different binary attributes [6]. The annotated dataset (LFWA) is publicly available where it contains 13,143 images of 5,749 different identities. The dataset has a designated training set portion of 6,263 images, while the rest is reserved for testing. LFWA is one of the challenging datasets with large variations in pose, contrast, illumination and image quality.

4.2. Data Augmentation

In deep learning, data augmentation plays an important role in avoiding overfitting, specially with smaller datasets. Recently, several advanced methods for face data augmentation have been developed. In this paper, simple but effective data augmentation techniques are used: (1) Rotation: training images are rotated using a random rotation angle between [-5, +5] around the origin. (2) Scaling: images are scaled up and down with a random scale factor up to a quarter of the image

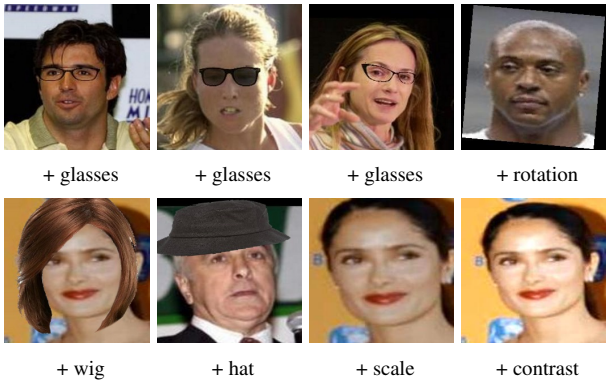


Fig. 3: Data augmentation with accessories.

size. (3) Contrast: by converting the color space of the images from RGB to HSV and randomly multiplying the S and V channels with a factor range between [0.5, 1.5]. In addition, blurring with two different filter size (3x3 and 5x5) and histogram equalization are performed. Furthermore, some techniques are applied in combination (e.g. rotation and scaling, or rotation and blurring).

We also add another type of augmentation by superimposing accessories such as glasses, hats, and wigs, on the training images. For this, the annotation of eyes locations are used to properly scale and rotate the added accessory. Random samples from the embedded items and generated augmented data are shown on Figure 3. In total, we generated 22 images per training sample, which corresponds to expanding our training set to 137,786 samples.

4.3. Results and Evaluation

We compare our work to the results obtained by three state-of-the-art methods, along with the baseline of choosing the most frequent label for each attribute. The performance comparison reported in Table 2 shows that our average accuracy compared to the best system (MCNN-AUX [9]) is almost 0.7% higher and outperforms it for 33 of 40 of the attributes. The state-of-art on this dataset has shown a relative increase of 2.46 on average in more than two years. Considering the results, we see that both our approach and the MCNN-AUX approach performs better compared to each attribute being trained individually. Thus our results confirm that grouping attributes in a MTL framework is useful.

As for the small but consistent improvements over the state-of-the-art, we believe that there are two reasons: First, we used several data augmentation techniques, whereas the augmentation is done by only jittering the original dataset in [9]. Second, [9] uses a small network that consists of three convolutional stages and two hidden layers and shares weights among different attributes. We believe that using a larger network and sharing the weights only within a regional group allows for a more powerful network, which is then constrained by way of data augmentation to reduce overfitting.

Finally, in order to see the benefits of directing the attention and the two-stage training, we trained 3 systems: System1) applying only the second stage, which corresponds to training in a MTL scenario using the original images without directing attention; System2) using blurring in both stages rather than using original images in Stage 2; and System3) the proposed method.

As can be seen in Figure 4, the error drops fastest in the proposed scheme where the system is given a little information about the rough feature location (proposed approach is better than System1), but only enough to direct the attention (proposed approach is better than System2). Training with original images eventually catches up and even surpasses training with blurred images and also comes close to the proposed method. This is in fact expected, since blurring loses some information.

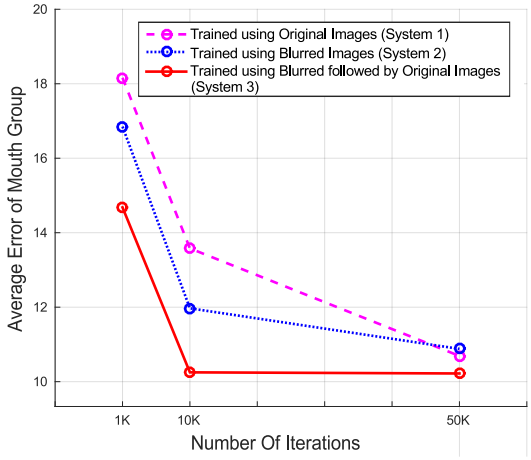


Fig. 4: Average error of mouth group for 3 systems.

5. CONCLUSION

We presented a multi-task framework for face attribute classification based on feature locality. The grouping of the attributes reduces overfitting, in addition to speeding up the learning process. We also show that by using a little amount of domain knowledge about attributes' locality on the face, the network learns much faster and even slightly increases accuracy. With the use of several data augmentation techniques, the system obtains state-of-art results.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge NVIDIA Corporation with the donation of the Titan X Pascal GPU used in this research.

7. REFERENCES

[1] Antitza Dantcheva, Petros Elia, and Arun Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2016.

[2] Ohil K Manyam, Neeraj Kumar, Peter Belhumeur, and David Kriegman, "Two faces are better than one: Face recognition in group photographs," in *IJCB, 2011*. IEEE, 2011, pp. 1–8.

[3] Thomas Berg and Peter N Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *CVPR*, 2013, pp. 955–962.

[4] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar, "Attribute and simile classifiers for face verification," in *ICCV*. IEEE, 2009, pp. 365–372.

[5] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.

[7] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Describing people: A poselet-based approach to attribute classification," in *ICCV*, 2011, pp. 1543–1550.

[8] Rogerio Schmidt Feris, Christoph Lampert, and Devi Parikh, *Visual Attributes*, Springer, 2017.

[9] Emily M Hand and Rama Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification.," in *AAAI*, 2017, pp. 4068–4074.

[10] Yan Li, Ruiping Wang, Haomiao Liu, Huajie Jiang, Shiguang Shan, and Xilin Chen, "Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction," in *ICCV*, 2015, pp. 3819–3827.

[11] Yang Zhong, Josephine Sullivan, and Haibo Li, "Face attribute prediction using off-the-shelf cnn features," in *ICB*. IEEE, 2016, pp. 1–7.

[12] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *NIPS*, 2014, pp. 217–225.

[13] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *CVPR workshops*. IEEE, 2014, pp. 806–813.

[14] Fengyi Song, Xiaoyang Tan, and Songcan Chen, "Exploiting relationship between attributes for improved face verification," *Computer Vision and Image Understanding*, vol. 122, pp. 143–154, 2014.

[15] Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult, "Are facial attributes adversarially robust?," in *ICPR*. IEEE, 2016, pp. 3121–3127.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[17] 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," *arXiv:1603.01249*, 2016.

[18] Sihua Yi, Nan Jiang, Bin Feng, Xinggang Wang, and Wenyu Liu, "Online similarity learning for visual tracking," *Information Sciences*, vol. 364, pp. 33–50, 2016.

[19] Wenhao Huang, Guojie Song, Haikun Hong, and Kunqing Xie, "Deep architecture for traffic flow prediction: deep belief networks with multitask learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014.

[20] Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and Stephen J Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 523–536, 2013.

[21] Jonathan Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine learning*, vol. 28, no. 1, pp. 7–39, 1997.

[22] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Int. Conf. on Multimedia*. ACM, 2014, pp. 675–678.

[23] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[24] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., 07-49, University of Massachusetts, Amherst, 2007.

[25] Hu Han, Anil K Jain, Shiguang Shan, and Xilin Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

#	Attribute	Baseline	[6]	[11]	[25]	Independent [9]	MCNN-AUX [9]	Ours
				Head				
1	Black Hair	87.63	90	91	83	91.84	92.63	**92.79**
2	Blond Hair	95.74	97	97	92	97.23	**97.41**	**97.41**
3	Brown Hair	64.56	77	76	97	80.84	80.85	**81.09**
4	Gray Hair	84.25	84	87	**89**	88.98	88.93	88.92
5	Bald	89.37	88	91	**93**	91.51	91.94	92.09
6	Bangs	83.59	88	91	77	90.47	90.08	**91.05**
7	Straight Hair	64.44	76	77	79	81.54	78.53	**82.30**
8	Wavy Hair	55.49	76	77	**94**	81.58	81.61	81.89
9	Reced. Hairline	59.84	85	86	85	86.00	86.26	**86.89**
10	Wear. Hat	85.52	88	90	**92**	89.79	90.07	91.50
				Eyes				
11	Arch. Eyebrows	74.88	82	83	**86**	81.40	81.78	84.01
12	Narrow Eyes	65.50	81	81	82	82.48	82.86	**83.26**
13	Bushy Eyebrows	53.70	82	83	82	84.79	84.97	**85.94**
14	Bags Under Eyes	58.29	83	83	**92**	83.24	83.48	83.01
15	Eyeglasses	81.99	**95**	91	86	92.15	91.30	92.54
				Nose				
16	Big Nose	68.59	81	83	80	84.43	**84.98**	84.80
17	Pointy Nose	71.10	80	83	84	**84.41**	84.14	84.40
				Mouth				
18	Big Lips	62.86	75	78	81	79.06	79.24	**82.24**
19	Smiling	60.50	91	90	92	**92.22**	91.83	92.14
20	Mustache	86.62	92	94	**95**	93.69	93.43	94.14
21	Wear. Lipstick	85.53	95	95	93	94.68	**95.04**	94.46
22	Mouth S. O.	58.70	82	81	**86**	82.41	83.51	85.75
				Cheek				
23	5 O-clock Shadow	58.64	84	77	**80**	77.39	77.06	78.01
24	Rosy Cheeks	79.65	78	82	86	**89.46**	87.92	88.90
25	Goatee	74.68	78	83	**88**	83.34	82.97	82.50
26	H. Cheekbones	67.74	88	88	**89**	88.02	88.38	88.49
27	No Beard	70.05	79	80	81	81.45	82.15	**83.39**
28	Sideburns	68.72	77	82	80	81.70	83.13	**83.49**
				Shoulder				
29	Double Chin	62.44	78	80	**92**	82.00	81.52	81.92
30	Wear. Necklace	80.49	88	90	**91**	89.98	89.94	90.77
31	Wear. Necktie	64.09	79	81	81	80.34	80.66	**81.19**
				General				
32	Attractive	62.87	83	79	**84**	80.20	80.31	80.96
33	Blurry	84.02	74	88	75	86.71	85.23	**86.82**
34	Chubby	63.92	73	75	**78**	75.85	76.86	76.93
35	Young	79.60	86	86	**87**	85.11	85.84	86.06
36	Male	78.77	94	94	93	93.27	94.02	**94.20**
37	Pale Skin	52.09	84	73	91	94.31	93.32	**94.38**
38	Oval Face	51.49	74	75	75	77.06	77.39	**78.01**
39	Heavy Makeup	89.20	95	95	95	95.63	**95.85**	95.47
40	Wear. Earrings	86.86	94	95	80	94.73	94.95	**95.04**
	Average	71.85	83.85	84.78	86.15	86.28	86.31	**86.98**

Table 2: State-of-the-art accuracies compared with the results obtained in this work. Bold figures indicate the best results.