

FACE RECOGNITION BY FACIAL ATTRIBUTE ASSISTED NETWORK

Jui-Shan Chan, Gee-Sern (Jison) Hsu, Hung-Cheng Shie, Yan-Xiang Chen

National Taiwan University of Science and Technology

ABSTRACT

We propose the Facial Attribute Assistant Network (FAAN) for face recognition in the wild. The FAAN is developed to imitate some general human description of a face using facial attributes, such as gender, ethnicity, hair color and with or without eyeglasses. Given a face as the input, the FAAN renders an identity descriptor (I-descriptor) and an attribute descriptor (A-descriptor) in the output. These descriptors are exploited with the Minimum Distance Subset scheme, proposed in this study, for handling image-set based recognition. When searching for a match in a gallery set, only the subjects with similar attributes as of the probe, measured by the A-descriptors, are selected for matching by comparing the I-descriptors. The FAAN is built on the Residual Network (ResNet), which is first trained for face identification and then *fine tuned* for attribute identification. This study shows that attributes can effectively reduce the search space and make the search more efficient and accurate. Compared with other state-of-the-art methods, the FAAN demonstrates a competitive performance on the MPIE and IJB-A benchmarks.

1. INTRODUCTION

Approaches for face recognition may be split into two categories, one with handcrafted features and the other with deeply-learned features. The difference is made by the deep learning networks employed in the latter but not in the former. Many methods developed in the last couple decades belong to the first category. The Local Binary Pattern (LBP) is proposed in [1]. The high-dimensional multiscale LBP features extracted from patches around facial landmarks are exploited in [2]. In [3], the local patterns of Gabor magnitude and phase are combined as an effective feature. Cross-pose performance has been a central concern in face recognition, and much of the progress has been made in recent years. Asthana et al. [4] exploits landmark detection and support vector regression for pose normalized recognition. A query face with the landmark-based estimated pose is aligned to a 3D face model, and the aligned face model is rotated to the frontal view to match against the gallery faces. The Generic Elastic Model (GEM) has been used in several latest studies. The method in [5] extracts sparse features by using subspace modeling and l_1 -minimization to induce pose-tolerance. This approach enables the synthesis of an equivalent frontal face

for a given query. The approach in [6] uses the GEM and extracts the Walsh local binary patterns from the periocular region as features. However, most of these methods only work for poses within 60° in yaw, and degrade significantly for profile and extreme poses.

Some great progress has been made by deep learning approaches in recent years. Trained on 4 million faces taken from more than 4,000 subjects, the DeepFace reaches 97.35% in accuracy on the Labeled Faces in the Wild (LFW) dataset [7], comparable to human performance at 97.35%. By increasing the dimension of hidden representations and adding supervision to early convolutional layers, DeepID2+ achieves a better-than-human accuracy at 99.47% [8]. As nearly perfect performance has been achieved on the LFW, the IARPA Janus Benchmark A (IJB-A) database is made and introduced to the community [9]. The IJB-A offers the following properties: 1) full pose variation, 2) protocols supporting open-set identification (1:N search) and verification (1:1 comparison), 3) wider geographic variation of subjects and 4) ground truth eye and nose locations. Using multiple pose specific models and rendered face images, the Pose-Aware Model (PAM) [10] reports TAR 65.2% at FAR=0.001 and recognition rate 84.0% at Rank 1 on this new benchmark. Similar to the PAM, the features considered by Chen et al. [11] are also extracted from a deep network trained on the CASIA-WebFace database [12]. However, they exploit the joint Bayesian metric learning and attain TAR 83.8% at FAR=0.001 and recognition rate 90.3% at Rank 1.

We propose the Facial Attribute Assisted Network (FAAN) that takes in a face and renders two feature vectors, called I-descriptor and A-descriptor. The I-descriptor characterizes the identity of the face and the A-descriptor characterizes the attributes of the face. The FAAN is built on the Residual Network (ResNet) [13] as it outperforms many state-of-the-art networks in the ILSVRC competition with comparably low computational complexity. It is trained on a large face database, and fine tuned on facial attribute databases. *Fine tune* refers to freeze the weights of the majority of the network and only allow a small part of the weights to change when retraining the network for a different target task. The facial attributes refer to semantic observables of one's face, such as gender, ethnicity, hair color and others. When searching for a match in a gallery set to a probe, the A-descriptors are called up to reduce the size of the search pool. When comparing

a pair of faces, the A-descriptors are used to measure the generic similarity in between.

Although the contribution of this study can be summarized as the proposed FAAN experimentally validated as effective for face recognition, it can be decomposed into the following: 1) Facial attributes proven to facilitate recognition, 2) the proposed Minimum distance pairing validated as an efficient means to measure set-to-set similarity, and 3) the FAAN shows a comparable performance to the state of the art on the IJB-A benchmark.

2. FACIAL ATTRIBUTE ASSISTED NETWORK

The proposed Facial Attribute Assisted Network (FAAN) is built on the ResNet-101 [13], trained on a face database for person identification, and then fine tuned on the attribute databases for attribute identification. The Minimum Distance Pairing is proposed to handle the set-to-set match.

2.1. Training Databases and Feature Extraction

The ResNet-101 has an input convolution layer, 33 triple-convolution-layer blocks, and a fully-connected (FC) layer. For identity recognition, three databases were used in our experiments, the MPIE, the CASIA-WebFace and the IJB-A. The MPIE offers more than 750,000 images of 337 individuals taken in four sessions held at different time. The facial images were taken under 15 viewpoints, 19 illumination conditions and with 6 different expressions. It is used as one of the training database.

The CASIA-WebFace is one of the publicly accessible in-the-wild databases, and it offers 494,414 facial images of 10,575 subjects collected in the wild. Most subjects have 10~60 facial images available. As the images were collected in a semi-automatic way, part of the images are of poor quality or mislabeled. We manually remove most of the poor-quality and mislabeled images and the images of the subjects who were also collected in the IJB-A database. The remaining are 426,975 images of 9,168 subjects. Similar to the works in [10, 11], the CASIA-WebFace is used as one training database. However, we compare the performances of the I-descriptors extracted from the FAAN trained on 1) MPIE [14], 2) CASIA-WebFace and 3) both databases combined to address the issues of training set. Performance evaluation is conducted on the IJB-A database and reported in Sec.3.

The IJB-A is one of the latest public in-the-wild databases. It contains 500 subjects with 5,397 images and 2,042 videos split into 20,412 frames, 11.4 images and 4.2 videos per subject. It offers a large collection of challenging samples and a properly designed evaluation protocol [15, 16]. Different protocols are defined for search and comparison with the same data specifications. All 500 IJB-A subjects are randomly split into training and testing sets. For each split, 333 subjects are randomly sampled and placed in the training set, and the remaining 167 subjects are placed in the testing set.

The performance is average of ten such splits. Additional imagery may be used for training under the strict condition that no such imagery contain the same subjects that are in the test set. Bootstrap samples of training and testing sets are performed instead of cross validation to increase the number of testing subjects. For each split, every testing subject has his/her imagery randomly sampled into either the probe set or the gallery set. The protocol includes single and multiple probe samples for each subject, simulating different scenarios for querying a face recognition system.

The IJB-A search protocol measures the accuracy of open-set and closed-set search on the gallery templates using probe templates. 55 randomly selected subjects in each random split have templates/imagery removed from the gallery set. Every probe template in a given split is to be searched against the set of gallery templates. Each search run contains 112 gallery templates and 1763 probe templates, which include 1,187 genuine probe templates and 576 impostor probe templates. Each comparison (verification) run contains 11,748 pairs of templates, which include 1,756 positive and 9,992 negative pairs on average. Ten random splits of training and testing sets are specified in the protocol. For each search, the identities of the 30 closest matching gallery templates, and the corresponding match scores are recorded. Using these results, the rank-1 and rank-5 accuracy are reported with the miss rate corresponding to False Alarm Rates (FARs) at 0.1 and 0.01. The Cumulative Match Characteristic (CMC) gives the percentage of probes identified within a given rank. The accuracy for verification is measured using the Receiver Operating Characteristic (ROC), which shows the True Accept Rate (TAR) versus the False Accept Rate (FAR) for different thresholds.

When fine tuning the ID-trained FAAN for attribute identification, we only allow the parameters on the output FC layer to change and freeze the rest of the network connections. The attributes considered include gender, ethnicity (Asian, African and Caucasian), hair color (black, blond, brown, white), eyeglasses and baldness. We use the CelebA [17] as the major database for attribute training. It contains 10,000 subjects with 20 images per subject, and each image is annotated with 40 facial attributes. Since the ethnic distribution in the CelebA is biased to European, we augment the training set by adding in 10,000 Europeans from the PubFig [18], 10,000 Asians from the CAS-PEAL [18] and 10,000 African from the Morph [19].

2.2. Minimum Distance Pairing for Matching

The protocol considered is template based, and one template contains a single or multiple images that belong to the same subject. When comparing two templates, T_1 with N_{T_1} images included and T_2 with N_{T_2} images included, there are $N_{T_1} N_{T_2}$ pairing distances in between. We propose the Minimum Distance Pairing (MDP) to measure the template-to-template similarity. Given a set of images of the same subject

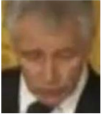

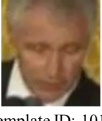

















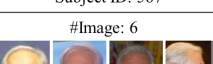


Probe Template		Rank-1	Rank-2	Rank-4	Rank-6	Rank-8
   Template ID: 1011 Subject ID: 347	Without Att.	#Image: 42  Template ID: 1191 Subject ID: 395	#Image: 42  Template ID: 2600 Subject ID: 391	#Image: 79  Template ID: 2656 Subject ID: 374	#Image: 40  Template ID: 4089 Subject ID: 507	#Image: 25  Template ID: 986 Subject ID: 347
	E	#Image: 42  Template ID: 1191 Subject ID: 395	#Image: 42  Template ID: 2600 Subject ID: 391	#Image: 95  Template ID: 5762 Subject ID: 746	#Image: 25  Template ID: 986 Subject ID: 347	#Image: 104  Template ID: 1811 Subject ID: 14
	E Hc	#Image: 64  Template ID: 4208 Subject ID: 3535	#Image: 25  Template ID: 986 Subject ID: 347	#Image: 32  Template ID: 2859 Subject ID: 567	#Image: 21  Template ID: 3664 Subject ID: 750	#Image: 42  Template ID: 3902 Subject ID: 1826
	E Hc Eg	#Image: 25  Template ID: 986 Subject ID: 347	#Image: 104  Template ID: 1811 Subject ID: 14	#Image: 6  Template ID: 1378 Subject ID: 485	#Image: 14  Template ID: 4232 Subject ID: 889	#Image: 38  Template ID: 2069 Subject ID: 17

Fig. 1. Face search with and without facial attributes

and each image with a different level of difficulty to be recognized, we tend to ignore those with poor recognizability and focus on the ones with better quality to be recognized. The MDP is proposed to reflect this tendency and postulate that the distance between two sets \mathbf{X} and \mathbf{Y} is given by the minimum distance with all cross-set data pairs, as described as follows.

$$d(\mathbf{X}, \mathbf{Y}) = \min (d(x_i, y_j), x_i \in \mathbf{X}, y_j \in \mathbf{Y}, \forall i, j) \quad (1)$$

where $d(\cdot)$ is a metric, x_i and y_j are an arbitrary datum in \mathbf{X} and \mathbf{Y} , respectively. The metric for measuring the difference between a face pair (d_i^g, d_j^p) is the Euclidean distance and Hamming distance for the attribute pair (a_i^g, a_j^p), where d_i^g and d_j^p denote the I-descriptors extracted from face x_i in the gallery and x_j in the probe set, and a_i^g and a_j^p are the associated A-descriptors, respectively. Note that a_i^g and a_j^p are 11-dimensional, however, the 3-class ethnicity and 4-class hair color are considered mutually exclusive. For example, one is identified as an Asian, the 3-class ethnicity gives $[1, 0, 0]$ and there cannot be more than one 1 in these three labels. The hair color and baldness are also exclusive, i.e., when the baldness is identified as 1, the 4-class hair color must be all 0. In our experiments, we compare the identification results with and without using the attribute distance for search pool reduction. We impose a threshold on the Hamming distance between (a_i^g, a_j^p) to adjust the attributes to be considered.

We collect the MDP distances by applying (1) to the IJB-A training set. The IJB-A training set can be used to make intra-pairs that belong to the same subjects, and extra-pairs that belong to two different subjects. We compose 3,200 intra-pairs and 9,600 extra-pairs in each training session. Given this

collection of the MDP distances, we build an SVM classifier to maximize the margin between the intra-pairs and extra-pairs. At testing or runtime phase, the MDP distances between a probe template and all gallery templates are computed and classified by the SVM classifier.

3. EXPERIMENTAL EVALUATION

All face images are scaled to 256×256 using the landmarks obtained by the Regressive Tree Structured Model (RTSM) [20]. Each training session is run with data randomly partitioned into 80% for training, 10% for validation and another 10% for testing. The training mostly follows the way reported in [13]. Batch normalization is performed after each convolution and before (Relu) activation. Steepest gradient descend with a mini-batch size of 256 is employed. The learning rate starts at 0.1 and is divided by 10 when the error plateaus, and the models are trained for up to 200k iterations. It is first trained on a face database with FC-ID (fully-connected for identity) layer, and fine tuned on attribute databases with FC-AT layer. We extract the feature from the output of the last triple-convolution-layer block as the I-descriptor or A-descriptor, depending on whether it is from the FC-ID layer or the FC-AT layer.

Fig.1 shows a case with the benefit of incorporating attributes in the FAAN framework for face search. Row 1 shows the result with I-descriptors only, and the target appears at Rank-8. When imposing the E (Ethnicity) constraint, as shown in Row 2, the subject 507 in Row 1 is removed, and the target subject is moved to Rank-6. Note that the subject 746 in Row 2, Rank-4 was in Row 1, Rank-7, but not shown because of limited width. When additionally imposing the Hc

Methods ↓	IJB-A Verification (TAR)		IJB-A Identification (Rec. Rate)		
Metrics →	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5	@Rank-10
PAM [10]	0.826±0.018	0.652±0.037	0.840±0.012	0.925±0.008	0.946±0.007
DCNN _{fusion} [11]	0.838±0.042	-	0.903±0.012	0.965±0.008	0.977±0.007
FAAN	0.912±0.030	0.850±0.021	0.890±0.015	0.963±0.009	0.982±0.005

Table 1. The verification in FAR=0.01 and 0.001 and identification performances in Rank 1, 5 and 10 on IJB-A.

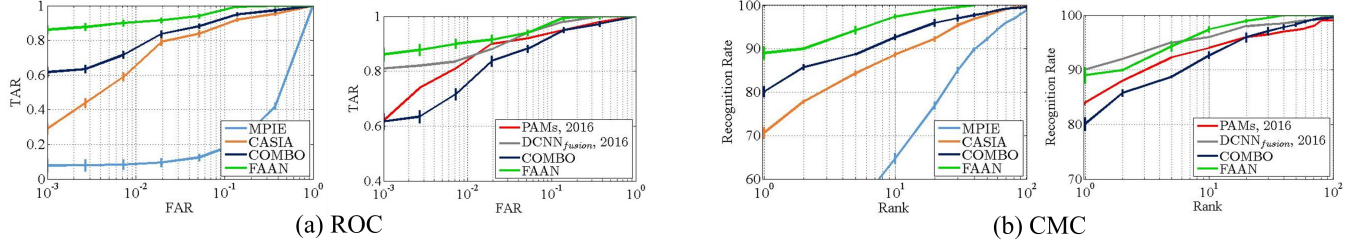


Fig. 2. The CMC (Cumulative Match Characteristic) and ROC (Receiver Operating Characteristic) of the FAAN compared with different setups and state-of-the-art approaches [10, 11].

(Hair Color), the subject 395 in Row 2, Rank-1 was removed from Row 3, and the target is promoted to Rank-2. Row 4 shows the result of imposing E, Hc and Eg (Eyeglasses).

The ROCs and CMCs of the FAAN compared with different setups and state-of-the-art approaches [10, 11] are shown in Fig.2. Note that the vertical scales are made different in the plots for better visibility. The MPIE denotes for the network trained on the MPIE only, the CASIA denotes for training on the CASIA-WebFace only, and the COMBO denotes for training on the combination of MPIE and CASIA-WebFace. All these three setups are made for identity recognition, without considering of attributes. The FAAN is the COMBO fine-tuned for the A-descriptor extraction. The left one in (a) shows that the FAAN outperforms all, while the MPIE-trained performs the worst, highlighting the fact that in-the-wild training set is a requirement for achieving good performance for in-the-wild face recognition. It also deserves special attention that the COMBO-trained outperforms both the MPIE-trained and CASIA-trained, and performs closely to [10, 11] with 8% lower TAR at FAR=0.01 and a fraction lower at Rank-10 recognition rate.

It is plausibly arguable that some of the attributes are not considered permanent to a person and cannot be used for characterizing a subject. However, it is a common practice for a security personnel to inquire those and other attributes when searching for a particular suspect. With more semantic attributes considered, the search can be faster and more efficient than without. In some cases when the search time is critical, the proposed FAAN can be a powerful solution. Table 2 shows the time consumption for cases with and without attributes. For a gallery size of 112 subjects, the FAAN takes 84ms for searching through all 112 subjects, but it only takes 14ms to search through the 20 subjects selected by using three

Feat. Ext.	Whole Gallery (112)	Attr. Matched (20)
13 ms	84 ms	14 ms

Table 2. Time needed for feature extraction (I-descriptor), for search thru whole gallery with 112 subjects, and for search thru the 20 subjects with matched attributes

matched attributes. Note that the flexibility of adjusting the attributes considered in the FAAN is provided in the framework. We only have to turn on or off the selected attributes, as what it shows in Fig.1.

4. CONCLUSION

We propose the Facial Attribute Assisted Network for tackling face recognition in the wild. The FAAN aims to imitate the generic description for a face using facial attributes, e.g., gender, ethnicity, hair color and others. The network is trained on a large face database for identity identification, and retrained on attribute databases for attribute identification, resulting in a network that can render an identity descriptor and an attribute descriptor for a face. We have demonstrated the following results in this study: 1) In-the-wild training database is required to ensure the performance for in-the-wild recognition; however, additional database may boost the performance. 2) Facial attributes offer a coarse dimension to reduce the search space and improve recognition, and the FAAN framework provides an easy-to-use mechanism to turn on or offer any particular attribute to alter the search space. 3) The Minimum Distance Pairing scheme is an efficient solution for measuring the similarity between homogeneous datasets.

5. REFERENCES

- [1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen, "Face description with local binary patterns: Application to face recognition," *TPAMI*, vol. 28, no. 12, pp. 2037–2041, 2006.
- [2] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*, 2013, pp. 3025–3032.
- [3] Shufu Xie, Shiguang Shan, Xilin Chen, and Jie Chen, "Fusing local patterns of gabor magnitude and phase for face recognition," *TIP*, vol. 19, no. 5, pp. 1349–1361, 2010.
- [4] Akshay Asthana, Tim K. Marks, Michael J. Jones, Kinh H. Tieu, and M. V. Rohith, "Fully automatic pose-invariant face recognition via 3d pose normalization," in *ICCV*, 2011, pp. 937–944.
- [5] Ramzi Abiantun, Utsav Prabhu, and Marios Savvides, "Sparse feature extraction for pose-tolerant face recognition," *TPAMI*, vol. 36, no. 10, pp. 2061–2073, Oct. 2014.
- [6] Felix Juefei-Xu, Khoa Luu, and Marios Savvides, "Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios," *TIP*, vol. 24, no. 12, pp. 4780–4795, Dec. 2015.
- [7] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.
- [8] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," in *CVPR*, 2015, pp. 2892–2900.
- [9] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *CVPR. IEEE*, 2015, pp. 1931–1939.
- [10] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan, "Pose-aware face recognition in the wild," in *CVPR*, 2016, pp. 4838–4846.
- [11] Jun-Cheng Chen, Vishal M Patel, and Rama Chellappa, "Unconstrained face verification using deep cnn features," in *CVPR*, 2016, pp. 1–9.
- [12] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li, "Learning face representation from scratch," *arXiv:1411.7923*, 2014.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
- [14] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker, "Multi-pie," *IVC*, vol. 28, pp. 807–813, May 2010.
- [15] J. C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep cnn features," in *WACV*, March 2016, pp. 1–9.
- [16] Iacopo Masi, Stephen Rawls, Gerard Medioni, and Prem Natarajan, "Pose-aware face recognition in the wild," in *CVPR*, 2016, pp. 4838–4846.
- [17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.
- [18] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, 2009, pp. 365–372.
- [19] Karl Ricanek and Tamirat Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *FG*, 2006, pp. 341–345.
- [20] Gee-Sern Hsu, Kai-Hsiang Chang, and Shih-Chieh Huang, "Regressive tree structured model for facial landmark localization," in *ICCV*, 2015, pp. 3855–3861.