A Face Recognition Signature Combining Patch-based Features with Soft Facial Attributes

L. Zhang, P. Dou, I.A. Kakadiaris

Computational Biomedicine Lab, 4849 Calhoun Rd, Rm 373, Houston, TX 77204

Abstract

This paper focuses on improving face recognition performance with a new signature combining implicit facial features with explicit soft facial attributes. This signature has two components: the existing patch-based features and the soft facial attributes. A deep convolutional neural network adapted from state-of-the-art networks is used to learn the soft facial attributes. Then, a signature matcher is introduced that merges the contributions of both patch-based features and the facial attributes. In this matcher, the matching scores computed from patch-based features and the facial attributes are combined to obtain a final matching score. The matcher is also extended so that different weights are assigned to different facial attributes. The proposed signature and matcher have been evaluated with the UR2D system on the UHDB31 and IJB-A datasets. The experimental results indicate that the proposed signature achieve better performance than using only patch-based features. The Rank-1 accuracy is improved significantly by 4% and 0.37% on the two datasets when compared with the UR2D system.

Keywords: Face recognition, convolutional neural network, facial attribute

1. Introduction

Face recognition is one of the major visual recognition tasks in the fields of biometrics, computer vision, image processing and machine learning. In recent years, most of the significant advances in visual recognition have been achieved by deep learning models, especially deep Convolutional Neural Networks (CNNs) [1, 2, 3]. CNN was first proposed in the late 1990s by LeCun *et al.* [4, 5]. It

Email addresses: lzhang@34@uh.edu (L. Zhang), pdou@central.uh.edu (P. Dou), ioannisk@uh.edu (I.A. Kakadiaris)

was quickly overwhelmed by the combination of other shallow descriptors (such as SIFT, HOG, bag of words) with Support Vector Machines (SVMs). With the increase of image recognition data size and computation power, CNN has become more and more popular and dominant in the last five years. Krizhevsky et al. [6] proposed the classic eight-layer CNN model (AlexNet) with five convolutional and three fully connected layers. The model is trained via back-propagation through layers and performs extremely well in domains with a large amount of training data. Since then, many new CNN models have been constructed with larger sizes and different architectures to improve performance. Simonyan et al. [7] explored the influence of CNN depth by an architecture with small convolutional filters (3 \times 3). They achieved a significant improvement by pushing the depth to 16-19 layers in a VGG model. Szegedy et al. [8] introduced GoogLeNet as a 22-layer Inception network, which achieved impressive results in both image classification and object detection tasks. He et al. [9] proposed Residual Networks (ResNet) with a depth of up to 152 layers, which set new records for many image recognition tasks. Furthermore, He et al. [10] proposed a residual network of 1,000 layers with identity mappings that makes training easier and improves generalization.

Recently, many CNNs have been introduced in face recognition and have achieved a series of breakthroughs. Similar to image recognition, effective CNNs require a larger amount of training images and larger network sizes. Yaniv et al. [11] trained the DeepFace system with a standard eight-layer CNN using 4.4M labeled face images. Sun et al. [12, 13, 14] proposed the Deep-ID systems with more elaborate network architectures and fewer training face images, which achieved better performance when compared with the DeepFace system. FaceNet [15] was introduced with 22 layers based on the Inception network [8, 16]. It was trained on 200M face images and achieved further improvement. Parkhi et al. [17] introduced the VGG-Face network with up to 19 layers adapted from Simonyan et al. [7], which was trained on 2.6M images. This network also achieved comparable results and has been extended to other applications. To overcome the massive request of labeled training data, Masi et al. [18] proposed to use domain specific data augmentation, which generates synthesis images for CASIA Web-Face collection [19] based on different facial appearance variations. Their results trained with ResNet match the state-of-the-art results reported by the networks trained on millions of images. Recently, Xu et al. [20] presented the evaluation of a pose-invariant 3D-aided 2D face recognition system (UR2D) that is robust to pose variations as large as 90°. Different CNNs are integrated in face detection, landmark detection, 3D reconstruction and signature generation.

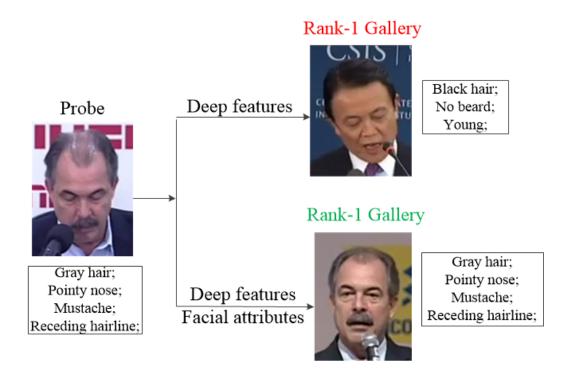


Figure 1: Depicted a matching comparison based on implicit deep features and the combination of both implicit deep features and explicit soft facial attributes on the UR2D system. The upper gallery indicates the matching with only implicit deep features while the bottom gallery indicates the matching with both implicit deep features and explicit soft facial attributes.

Overall, previous deep face networks have two limitations: (a) the learned deep features is implicit; there is no human-readable information and the discriminative information is encoded in high dimensional features. (b) explicit soft facial attribute features are underestimated, which can be used to improve recognition performance.

This work overcomes both limitations and improves face recognition performance by combining the implicit facial features with the new explicit facial attribute features. Due to the uncontrolled environments, the existing facial features are sensitive and more likely be corrupted. Thus, only using implicit facial features may lead to incorrect matching. On the other hand, the soft facial attribute features related to hair, eyebrow, nose and mouth are less sensitive to the change of environments and occlusions. By combining both facial features and soft facial attributes, more robust matching result can be obtained. Using the IJB-A face

recognition dataset [21] as an example, Figure 1 depicts the intuition of the proposed method. It can be observed that, due to similar background and poses, the probe face is mis-matched based on implicit deep features in the upper matching gallery. However, the probe face does not share any facial attributes with the incorrect matching gallery image. Using the proposed signature, however, both implicit deep features and explicit facial attribute features are taken into account, which leads to a correct matching gallery image in the bottom matching result. It can be observed that these human-describable facial attributes can be used to fix the shortage of deep feature based face recognition system.

The contributions of this paper are improving face recognition performance and robustness by the following new signature and matcher: (i) a facial signature with two components of patch-based features and soft facial attributes. The facial attribute component is extracted by a CNN adapted from state-of-art networks. The contribution of facial attribute information is not well explored in deep feature based face recognition systems. (ii) a signature matcher combining the soft facial attribute component with the existing patch-based features in the UR2D system. The signature matcher is also extended to assign different weights to different facial attributes.

The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 and Section 4 describe the proposed signature and matcher, respectively. The experimental design, results, and analysis are presented in Section 5. Section 6 concludes the paper.

2. Related Work

In face recognition, both global and local feature based methods have been proposed. Global methods learn discriminative information from the whole face image, such as subspace methods [22, 23], Sparse Representation based Classification (SRC) [24, 25] and Collaborative Representation based Classification (CRC) [26, 27] and CNN based methods [17, 18]. Although global methods have achieved great success in controlled environments, they are sensitive to the variations of facial expression, illumination and occlusion in uncontrolled real-world scenarios. On the other hand, local methods extract features from local regions. The classic local features include Local Binary Patterns (LBP) [28, 29], Gabor features [30, 31], Scale-Invariant Feature Transform (SIFT) [32, 33], gray values and so on. In local methods, most efforts focus on patch (block) based methods, which usually involve steps of local patch partition, local feature extraction, and local matching combination. With intelligent combination, these methods weaken

the influence of variant-prone or occluded patches and ensemble the matching of invariant or unoccluded patches [34, 31, 35, 36]. The drawback of most patch based methods is that they still rely on implicit features. The discriminative information is embedded in human-unreadable features.

Soft Attributes, also known as high-level semantic features, have drawn a lot of attention in the past years in domains from image recognition to face recognition. In the face recognition domain, soft facial attributes like gender, race, age, hair color, and facial hair are very intuitive and they provide more human-understandable descriptions of a subject. Humans also rely on these intuitive attributes to remember and identify different persons.

Many methods have been developed for improving the accuracy of attribute prediction. Based on the success of deep learning, most of them rely on CNNs. Yi et al. [37] proposed a multi-scale network that uses multiple local image patches as input to estimate age, gender, and ethnicity. The multi-scale image patches are cropped based on landmark locations. Liu et al. [38] developed the first CNN based facial attribute prediction framework to estimate 40 facial attributes. LNet and Anet are cascaded for face localization and attribute predication, respectively. Kang et al. [39] developed a face attribute classification method based on attribute-aware correlation maps and gated CNNs. Each correlation map of an attribute provides information about regions where the relevant features should be extracted. The CNN trained for each region is gated so that the classification errors of less relevant attributes contribute less in the learning process of back propagation. Rudd et al. [40] proposed a mixed objective optimization network based on joint optimization over all the attributes. The tasks of multi-label classification and domain adaptation are optimized under one unified objective function. Zhong et al. [41] proposed to use off-the-shelf CNN architectures to extract features for attribute prediction. Considering the diversity of different attributes, the features are extracted from different levels in CNNs. Kalayeh et al. [42] developed a semantic segmentation based network for facial attribute prediction. The localization cues learned by the semantic segmentation are used to guide the attention of the attribute prediction. Overall, these methods all focus on improving the prediction of facial attributes as a multi-label classification problem. In this paper, a further step is taken, which is using deep facial attributes to improve face recognition performance. Extracting high level semantic features from face recognition has been a goal for a long time. The facial attributes are both human-readable and describable. However, only using soft facial attributes may neglect the nondescribable features, which can be captured effectively by CNNs. Therefore, in this paper, a signature is proposed to combine both patch-based features and facial attribute features to boost face recognition performance.

Multi-label classification [43, 44, 45, 46, 47] is also related to the proposed method. In these methods, a classification model is learned to estimate multiple labels. These labels are used as output directly. On the other hand in the proposed method, the multi-labels are applied as input features to learn the matching identity.

3. Signature

In this section, one novel signature is introduced with two components: patch-based features extracted from the UR2D system; the proposed soft facial attributes extracted with one adapted state-of-the-art CNN to learn 40 facial attributes.

3.1. Patch-based feature component: \mathbb{S}^P

To integrated with the UR2D system [20], the same pipeline is followed for pre-processing and extracting the patch-based feature component: Pose Robust Face Signature (PRFS) component [48] and Deep Pose Robust Face Signature (DPRFS) component.

Given an input face image, the pipeline of UR2D follows: face detection, land-mark detection, pose estimation, 3D reconstruction, texture lifting, and signature extraction. Here only the signature extraction part is introduced, please refer to [20] for more details. Both PRFS and DPRFS are extracted from texture lifted images. Facial texture lifting is a technique [49] that lifts the pixel values from the original 2D images to a UV map. Given an original image I, a 3D-2D projection matrix J, 3D AFM model M, it first generates the geometry image G, each pixel of which captures the information of an existing or interpolated vertex on the 3D AFM surface. With G, a set of 2D coordinates referring to the pixels on an original 2D facial image is computed. Thus, the facial appearance is lifted and represented into a new texture image T. A 3D model M and Z-Buffer technique are applied to estimate the occlusion status for each pixel. This process also generates an occlusion mask M.

Both PRFS and DPRFS are patch-based features. In PRFS, the facial texture T and the self-occlusion mask Z are first divided into 64 non-overlapping local patches. Then, on each local patch, the discriminative DFD features [50] are extracted. In DPRFS, the facial texture T and the self-occlusion mask Z are first divided into eight partially-overlapping local patches. Then, a DPRFS model is trained for each patch based on softmax loss and center loss. Each patch-based feature component contains two part: feature matrix and occlusion encoding. Let

 $F=\{f_{ij}\}^{n\times m}$ represent a feature matrix, where each value f_{ij} represent the i^{th} feature of the j^{th} patch while n and m represent the number of features and the number of patches, respectively. The occlusion encoding is represented by $O=\{o_1,o_2,...,o_m\}$, where o_j is a binary value indicating whether the j^{th} patch is non-occluded. Based on the occlusion encoding of each patch, all the features are combined selectively during matching. Let $\mathbb{S}^P=\{F,O\}$ represent the patch-based feature component based on texture-lifted image. The \mathbb{S}^P size for PRFS and DPRFS are $64\times 1024+64$, $8\times 512+8$, respectively. Benefit from on CNN features, DPRFS performs better than PRFS.

3.2. Soft facial attribute component: \mathbb{S}^A

Given the original 2D image I, a CNN is built to extract facial attributes. One state-of-the-art network, such as VGG-Face [17] or ResNet [18] is adapted to learn the facial attribute features. First, the last fully connected layer is removed from the network. Then, a new fully connected facial attribute layer is added which outputs 40 facial attributes, listed in Table 1. Then, the sigmoid cross-entropy loss is applied to compute the attribute loss over N training images. The network architecture of the proposed facial attribute signature component is depicted in Figure 2 with VGG-Face as example. Let $A = \{a_1, a_2, ..., a_{40}\}$ represent the output of facial attribute layer for image I. With the sigmoid function, the probability of each facial attribute is obtained, which is denoted as $P = \{p_1, p_2, ..., p_{40}\}$, so that:

$$p_i = \frac{1}{1 + e^{-a_i}}. (1)$$

Thus, the explicit feature value for each facial attribute is obtained. For each 2D face image, the probability of each facial attribute is computed. This provides us the confidence score of each attribute. By setting a threshold 0.5 on P, a binary attribute vector $B = \{b_1, b_2, ... b_{40}\}$ is obtained, that indicates the valid attributes of each facial image. During the matching, the facial attributes that are contributing to improve the performance of face recognition can be directly observed as in Figure 1. Let $\mathbb{S}^A = \{A, B\}$ represent the facial attribute signature component. The component size of \mathbb{S}^A is 40×2 . The proposed new signature is represented by $\mathbb{S} = \{\mathbb{S}^D, \mathbb{S}^A\}$. The procedure of signature generation for \mathbb{S} is summarized in Algorithm 1.

Table 1: The 40 soft facial attributes are depicted below.

Male
Mouth Slightly Open
Mustache
Narrow Eyes
No Beard
Oval Face
Pale Skin
Pointy Nose
Receding Hairline
Rosy Cheeks
Sideburns
Smiling
Straight Hair
Wavy Hair
Wearing Earrings
Wearing Hat
Wearing Lipstick
Wearing Necklace
Wearing Necktie
Young

Algorithm 1: Signature: $\mathbb{S} = \{\mathbb{S}^P, \mathbb{S}^A\}$

Input: 2D image I and 3D AFM model M

Output: $\mathbb{S} = {\mathbb{S}^P, \mathbb{S}^A}$

- 1 Face detection and landmark detection
- ${f 2}$ Pose estimation and 3D reconstruction to obtain J
- $\mathbf{3}$ Generate geometry image G
- 4 Compute texture lifted image T and occlusion mask M
- 5 Compute feature matrix F and occlusion encoding O
- ${\bf 6}\,$ Compute facial attribute information A and B
- 7 return $\{\mathbb{S}=\{\mathbb{S}^P,\mathbb{S}^A\},\,\mathbb{S}^P=\{F,O\},\,\mathbb{S}^A=\{A,B\}\};$

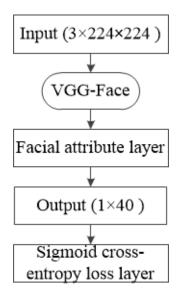


Figure 2: Depicted is the VGG-Face based network architecture of the soft facial attribute component.

4. Signature Matching

4.1. Patch-based feature component matching

In this subsection, the matching score is computed based on the component of patch-based features. Following the UR2D system, the cosine score is used to measure the similarity between each pair of signature components computed from gallery and probe image.

Let I^g and I^p represent a pair of gallery and probe image. In their patch-based signature component \mathbb{S}^{Pg} and \mathbb{S}^{Pp} , the feature matrix and occlusion encoding are represented by F^g , F^p and O^g , O^p , respectively. Let s^p represent the score of the patch-based feature component. The features are patch-based, and only non-occluded patches contribute to s^p . The signature matching score of patch-based feature component s^p is computed as:

$$s^p = \frac{1}{k} \sum_{j=1}^m (o_j^g \& o_j^p) \times cosine(F_j^g, F_j^p), \tag{2}$$

where k represents the number of non-occluded patch pairs.

4.2. Soft facial attribute component matching

In the facial attribute component, the attribute vectors of I^g and I^p are represented by A^g and A^p , respectively. Let s^a represent the score of the facial attribute component, which can be computed directly from

$$s^a = cosine(A^g, A^p). (3)$$

The final matching score s is computed as the sum of the scores of both components,

$$s = s^p + \lambda s^a, (4)$$

where λ represents the weight of the facial attribute component. With this matcher, the contributions are combined of both patch-based features and facial attribute features. λ is used to control the contribution of facial attribute features.

The problem of the previous matching is that all the facial attributes are treated equally. The difference between them are overlooked. However, these different attributes may have different weights. For example, the weight of "Bags under eyes" should be larger than that of "Eyeglasses". Also, the weight of "Receding hairline" should be larger than that of "Black hair" or 'Blond hair". Let $W = \{w_1, w_2, ..., w_{40}\}$ represent the weight vector of each attribute. These weights are introduced using the weighted cosine similarity. For A^g and A^p , the weighted similarity is computed as:

$$cosine_{w}(A^{g}, A^{p}, W) = \frac{\sum_{i=1}^{n} w_{i} a_{i}^{g} a_{i}^{p}}{\sqrt{\sum_{i=1}^{n} w_{i} a_{i}^{g^{2}}} \sqrt{\sum_{i=1}^{n} w_{i} a_{i}^{p^{2}}}}.$$
 (5)

Thus, the weighted attribute matching score can be computed as:

$$s_w^a = cosine_w(A^g, A^p, W). (6)$$

The final signature matching score with weighted attribute is:

$$s_w = s^p + \lambda s_w^a. (7)$$

With this weighted attribute matcher, different weights can be applied to different facial attributes. Note that, if binary weights are applied to facial attributes, the process has the effect of attribute selection. Only the attributes with non-zero weights will be selected in signature matching. The procedure of the signature matching is summarized in Algorithm 2.

Algorithm 2: Signature matching

```
Input: Gallery image signature \mathbb{S}^g = \{\mathbb{S}^{Pg} : \{F^g, O^g\}, \mathbb{S}^{Ag} : \{A^g, B^g\}\}, probe image signature \mathbb{S}^p = \{\mathbb{S}^{Pp} : \{F^p, O^p\}, \mathbb{S}^{Ap} : \{A^p, B^p\}\}, \lambda and W
```

Output: final siganture matching score s_w

- 1 Compute signature component matching score s^p based on Eq.(2)
- 2 Compute signature component matching score s_w^a based on Eq.(6)
- 3 Compute final matching score s_w based on Eq.(7)
- 4 return $\{s_w\}$;

5. Experiments

This section presents the evaluation of the proposed signature and matcher on two types of face recognition scenarios: constrained environment and unconstrained environment. The datasets used for testing are the UHDB31 dataset [51] and the IJB-A dataset [21]. The latest UR2D is used as a baseline pipeline. Following Xu et al. [20], the results are also compared with VGG-Face, FaceNet, and COTS v1.9. To demonstrate that the proposed signature can work with different facial features, two different facial features are used: PRFS and DPRFS. The facial attribute networks are trained on the CelebA dataset [38]. The weights of pre-trained models are used to fine-tune VGG-Face and ResNet. Both networks are trained for 50,000 iterations with Caffe [52]. The proposed signature with facial attribute is represented as UR2D-A. The λ for the signature matcher is set to 0.1, which is learned from a third dataset CASIA WebFace in the range of $\{0.1, 0.2, ..., 1\}$. The weight vector of the Weighted attribute matcher (UR2D-A-W) is decided by the training accuracy of each attribute. A baseline matcher is also created where different probe images are assigned with different weights for different attributes. The weight vector of the weighted Probe attribute matcher (UR2D-A-P) is decided by the attribute confidence scores of each probe image. Rank-1 accuracy is used as performance measurement.

5.1. Constrained face recognition

The UHDB31 dataset [51, 53] contains 29,106 color face images of 77 subjects with 21 poses and 18 illuminations. To exclude the illumination changes, a subset with nature illumination is selected. To evaluate the performance of cross pose face recognition, the front pose (pose-11) face images are used as gallery and

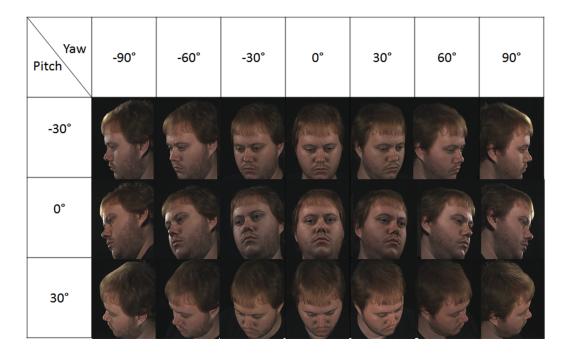


Figure 3: Depicted are image examples of different poses in the UHDB31 dataset.

the remaining images from 20 poses are used as probe. Figure 3 shows the example images from different poses. The performance of different methods under different facial features is shown in Tables 2 and 3.

From Table 2 and 3, it can be observed that with PRFS feature, the proposed signature can improve accuracy under nine poses, especially some large poses like pose-1 to pose-4 and pose-18 to pose-21. The accuracy improvements range from 1% to 8%. At the same time, the excellent performance of close frontal poses is retained. Under DPRFS, the proposed signature also achieves better performance. Overall, the proposed signatures achieve the best results on most of poses. The accuracy improvements range from 1% to 4%. It can be observed that the performance of facial attributes based on VGG-Face and ResNet are comparable. In addition, the weighted attribute matching and the weighted probe matching achieve comparable results on all the poses.

5.2. Unconstrained face recognition

The IJB-A dataset [21] contains images and videos from 500 subjects captured from "in the wild" environment. This dataset merges images and frames and

Table 2: Rank-1 performance of different methods computed on the UHDB31 dataset (I)(%). The methods in row-first order are VGG-Face, COTS v1.9, FaceNet, UR2D-PRFS, UR2D-A-PRFS-VGG-Face, UR2D-A-W-PRFS-VGG-Face, UR2D-A-P-PRFS-VGG-Face, UR2D-A-PRFS-ResNet, UR2D-A-W-PRFS-ResNet and UR2D-A-P-PRFS-ResNet.

Yaw Pitch		-90°		-60°		-30°			0°			+30°			+60°			+90°			
	14	11	58	69	32	95	94	90	100	99	100	100	95	93	99	79	38	92	19	7	60
+30°	48	51	48	90	94	94	100	100	100	100	100	100	100	100	100	95	96	96	47	55	51
	49	52	52	93	94	94	100	100	100	100	100	100	100	100	100	97	96	96	52	51	49
	48			94			100			100			100			96			51		
	22	9	84	88	52	99	100	99	100				100	100	100	94	73	99	27	10	91
00	79	82	82	100	100	100	100	100	100				100	100	100	100	100	100	84	87	87
0°	82	84	84	100	100	100	100	100	100		-		100	100	100	100	100	100	86	88	87
	82			100			100						100			100			87		
	8	0	44	2	19	80	91	90	99	96	99	99	96	98	97	52	15	90	9	3	35
-30°	43	45	46	90	92	92	99	99	99	100	100	100	99	99	99	95	97	97	58	57	57
	47	44	47	92	92	92	99	99	99	100	100	100	99	99	99	97	97	97	57	58	58
	46			92			99			100			99			97			57		

Table 3: Rank-1 performance of different methods computed on the UHDB31 dataset (II)(%). The methods in row-first order are VGG-Face, COTS v1.9, FaceNet, UR2D-DPRFS, UR2D-A-DPRFS-VGG-Face, UR2D-A-W-DPRFS-VGG-Face, UR2D-A-P-DPRFS-VGG-Face, UR2D-A-DPRFS-ResNet, UR2D-A-W-DPRFS-ResNet and UR2D-A-P-DPRFS-ResNet.

Yaw Pitch	-90°		-60°		-30°			0°			+30°			+60°			+90°				
	14	11	58	69	32	95	94	90	100	99	100	100	95	93	99	79	38	92	19	7	60
+30°	82	82	83	99	99	99	100	100	100	100	100	100	99	99	99	99	97	99	75	74	75
+30	82	82	83	99	99	99	100	100	100	100	100	100	99	99	99	97	97	99	75	78	77
	83			99			100			100			99			99			77		
	22	9	84	88	52	99	100	99	100				100	100	100	94	73	99	27	10	91
0°	96	97	97	100	100	100	100	100	100				100	100	100	100	100	100	96	96	96
	97	97	96	100	100	100	100	100	100		-		100	100	100	100	100	100	96	96	96
	96			100			100						100			100			96		
	8	0	44	2	19	80	91	90	99	96	99	99	96	98	97	52	15	90	9	3	35
200	75	78	77	97	99	99	100	100	100	100	100	100	100	100	100	96	96	96	79	83	83
-30°	76	76	76	99	99	99	100	100	100	100	100	100	100	100	100	96	96	95	83	82	83
	76			99			100			100			100			95			83		

Table 4: Rank-1 performance of different methods computed on the IJB-A dataset (%).

Methods	split-1	split-2	split-3	split-4	split-5	split-6	split-7	split-8	split-9	split-10	Average
VGG-Face	76.18	74.37	24.33	47.67	52.07	47.11	58.31	54.31	47.98	49.06	53.16
COTS v1.9	75.68	76.57	73.66	76.73	76.31	77.21	76.27	74.50	72.52	77.88	75.73
UR2D-PRFS	49.01	49.57	48.22	47.75	48.85	44.46	52.46	48.22	43.48	48.79	48.08
UR2D-A-PRFS-VGG-Face	52.77	52.83	51.16	51.08	51.21	47.42	56.47	51.26	46.24	52.23	51.27
UR2D-A-W-PRFS-VGG-Face	52.85	52.80	51.16	51.12	51.42	47.38	56.78	51.50	46.11	52.13	51.33
UR2D-A-P-PRFS-VGG-Face	52.13	52.27	50.46	50.14	50.79	46.70	55.62	50.86	45.41	51.23	50.56
UR2D-A-PRFS-ResNet	52.95	53.57	50.97	51.05	51.22	47.75	56.38	51.16	46.22	52.41	51.37
UR2D-A-W-PRFS-ResNet	53.05	53.50	51.05	51.08	51.32	47.83	56.08	51.43	46.07	52.33	51.38
UR2D-A-P-PRFS-ResNet	51.95	52.59	50.66	50.33	50.65	46.82	55.38	50.80	45.25	51.16	50.56
UR2D-DPRFS	78.78	77.60	77.94	79.88	78.44	80.57	81.78	79.00	75.94	79.22	78.92
UR2D-A-DPRFS-VGG-Face	79.16	77.80	78.38	80.04	78.51	80.81	81.96	79.14	76.16	79.36	79.13
UR2D-A-W-DPRFS-VGG-Face	79.19	77.82	78.53	80.04	78.76	80.95	81.78	79.37	76.50	79.51	79.29
UR2D-A-P-DPRFS-VGG-Face	79.09	77.88	78.42	80.17	78.72	80.95	81.70	79.10	76.35	79.54	79.19
UR2D-A-DPRFS-ResNet	78.95	77.92	78.23	80.50	78.54	80.87	82.16	79.54	76.24	79.51	79.25
UR2D-A-W-DPRFS-ResNet	79.06	77.92	78.34	80.44	78.62	80.91	82.03	79.61	76.24	79.47	79.26
UR2D-A-P-DPRFS-ResNet	78.89	77.88	78.53	80.17	78.62	80.74	82.12	79.41	76.24	79.36	79.20

provides evaluations on the template level. A template contains one or several images/frames of one subject. According to the IJB-A protocol, it splits galleries and probes into 10 splits. In the experiment, the same modification as [20] is followed for use it in close-set face recognition. The performance of different methods under different global signatures is shown in Table 4.

From Table 4, it can be observed that with PRFS signature, the proposed UR2D-A signature can improve the accuracy under all the splits. The average accuracy is improved by 3.19% and 3.29% with UR2D-A-VGG-Face and UR2D-A-ResNet, respectively. Under DPRFS, the proposed UR2D-A signature also achieves better performance. The average accuracy is improved by 0.21% and 0.33% with VGG-Face and ResNet, respectively. In addition, the weighed attribute matcher and the weighted probe attribute matcher achieve comparable results on all the splits. Overall, the proposed signature achieves the best results on all the splits compared to previous methods. The best performance is achieved by UR2D-A-W-DPRFS-VGG-Face. Figure 4 depicts more matching examples of the proposed signature. It can be observed that the proposed signature can be used to correct the matching error of implicit facial features. The reason behind this is that the proposed signature is more robust to facial attribute information. The facial attribute information is well captured to improve the performance while this information is overlooked in previous implicit facial feature based methods.

5.3. Sensitivity Analysis

In this section, the sensitivity of λ is evaluated with different values in the range of $\{0.1, 0.2, ..., 1\}$ for different matchers. The results of UR2D-A, UR2D-

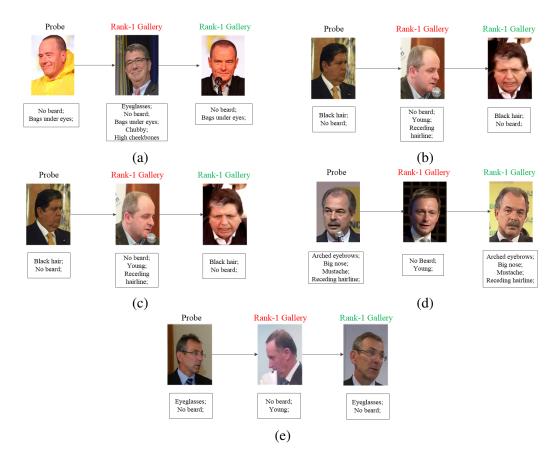


Figure 4: Depicted are the matching examples of the proposed signature. The three images in each example represent probe, incorrect gallery image matched with patch-based features only, and correct gallery image matched with the proposed signature, respectively.

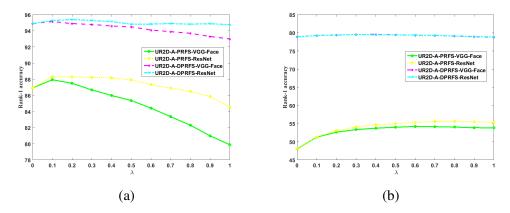


Figure 5: The performance of UR2D-A computed with different λ values. (a) UHDB31. (b) IJB-A.

A-W and UR2D-A-P are shown in Figures 5-7, respectively.

It can be observed that UR2D-A, UR2D-A-W and UR2D-A-P perform similarly on the two datasets. Different methods achieve the best result with different λ values. Also, the performance of DPRFS is less sensitive to λ than that of PRFS.

5.4. Statistical Analysis

In this section, statistical analysis is performed for the baseline and the best version of the proposed signature and matcher (UR2D-DPRFS and UR2D-A-W-DPRFS-VGG-Face) over the 30 data splits (20 from UHDB31 and 10 from IJB-A). From Demšar *et al.* [54], the Friedman test [55, 56] and the two tailed Bonferroni-Dunn test [57] are used to compare multiple methods over multiple datasets. Let r_i^j represent the rank of the j^{th} of k algorithm on the i^{th} of N datasets. The Friedman test compares the average ranks of different methods, by $R_j = \frac{1}{N} \sum_i r_i^j$. The null-hypothesis states that all the methods are equal, so their ranks R_j should be equivalent. The original Friedman statistic [55, 56],

$$\mathcal{X}_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right],\tag{8}$$

is distributed according to \mathcal{X}_F^2 with k-1 degrees of freedom. Due to its undesirable conservative property, Iman *et al.* [58] derived a better statistic

$$F_F = \frac{(N-1)\mathcal{X}_F^2}{N(k-1) - \mathcal{X}_F^2},\tag{9}$$

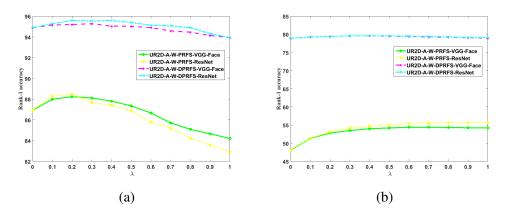


Figure 6: The performance of UR2D-A-W computed with different λ values. (a) UHDB31. (b) IJB-A.

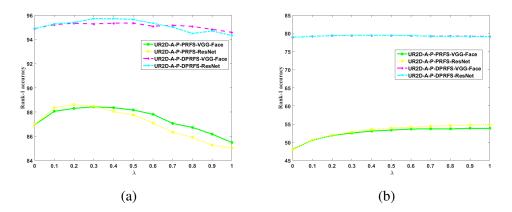


Figure 7: The performance of UR2D-A-P computed with different λ values. (a) UHDB31. (b) IJB-A.

which is distributed according to the F-distribution with k-1 and $(k-1)\times(N-1)$ degrees of freedom. First the average ranks for UR2D-DPRFS and UR2D-A-W-DPRFS-VGG-Face are computed as 1.73 and 1.27, respectively. The F_F statistical values of Rank-1 accuracy based on (9) are computed as 7.78. With two methods and 30 data splits, F_F is distributed with 2-1 and $(2-1)\times(30-1)=29$ degrees of freedom. The critical value of F(1,29) for $\alpha=0.10$ is 2.88<7.78, so the null-hypothesis is rejected. Then, the two tailed Bonferroni-Dunn test is applied to compare the two methods by the critical difference:

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}},\tag{10}$$

where q_{α} is the critical values. If the average rank between two methods is larger than critical difference, the two methods are significantly different. According to Table 5 in [54], the critical value of two methods when p=0.10 is 1.65. The critical difference is computed as $CD=1.65\sqrt{\frac{2\times 3}{6\times 30}}=0.30$. In conclusion, under Rank-1 accuracy, UR2D-A-W-DPRFS-VGG-Face performs significantly better than UR2D-DPRFS (the difference between ranks are 1.73-1.27=0.46>0.30).

6. Conclusion

This paper proposed a facial signature that contains both implicit facial features and explicit facial attribute features. Explicit soft facial attribute information is extracted to improve the performance of face recognition system that only uses implicit facial features. The experimental results confirmed the assumptions that facial attribute features explore more local discriminative information and can be used to improve matching performance. Comparing with the UR2D system, the Rank-1 accuracy is improved significantly by 4% and 0.37% for the UHDB31 dataset and the IJB-A dataset, respectively.

Acknowledgements

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH001. This grant is awarded to the Borders, Trade, and Immigration (BTI) Institute: A DHS Center of Excellence led by the University of Houston, and includes support for the project "Image and Video Person Identification in an Operational Environment: Phase I" awarded to the University of Houston. The views and conclusions contained in this document are those of the authors and should not be interpreted as

necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

References

- [1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.
- [2] R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (1) (2016) 142–158.
- [3] C. Szegedy, A. Toshev, D. Erhan, Deep neural networks for object detection, in: Proc. Advances in Neural Information Processing Systems, Lake Tahoe, NV, 2013, pp. 2553–2561.
- [4] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (4) (1989) 541–551.
- [5] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, L. Jackel, Handwritten digit recognition with a back-propagation network, in: Proc. Neural Information Processing Systems, Denver, CO, 1990, pp. 396–404.
- [6] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proc. Neural Information Processing Systems, Lake Tahoe, NV, 2012, pp. 1097–1105.
- [7] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. International Conference on Learning Representations, San Diego, CA, 2015, pp. 1–14.
- [8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. Computer Vision and Pattern Recognition, Boston, MA, 2015, pp. 1–9.

- [9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 770–778.
- [10] K. He, X. Zhang, S. Ren, J. S., Identity mappings in deep residual networks, in: Proc. European Conference on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 630–645.
- [11] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: Closing the gap to human-level performance in face verification, in: Proc. Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1701–1708.
- [12] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proc. Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1891–1898.
- [13] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proc. Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 1988–1996.
- [14] Y. Sun, D. Liang, X. Wang, X. Tang, DeepID3: Face recognition with very deep neural networks, arXiv preprint arXiv:1502.00873.
- [15] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: Proc. Computer Vision and Pattern Recognition, Boston, MA, 2015, pp. 815–823.
- [16] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proc. European Conference on Computer Vision, Zurich, Switzerland, 2014, pp. 818–833.
- [17] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proc. British Machine Vision Conference, Vol. 1, Swansea, UK, 2015, pp. 1–12.
- [18] I. Masi, A. Tran, T. Hassner, J. T. Leksut, G. Medioni, Do we really need to collect millions of faces for effective face recognition?, in: Proc. European Conference on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 579–596.
- [19] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923.

- [20] X. Xu, H. Le, P. Dou, Y. Wu, I. A. Kakadiaris, Evaluation of a 3D-aided pose invariant 2D face recognition system, in: Proc. International Joint Conference on Biometrics, Denver, CO, 2017, pp. 446–455.
- [21] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, A. K. Jain, Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A, in: Proc. Computer Vision and Pattern Recognition, Boston, MA, 2015, pp. 1931–1939.
- [22] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1) (1991) 71–86.
- [23] P. N. Belhumeur, J. P. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 711–720.
- [24] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210–227.
- [25] M. Yang, D. Zhang, J. Yang, Robust sparse coding for face recognition, in: Proc. Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011, pp. 625–632.
- [26] P. Zhu, L. Zhang, Q. Hu, S. C. Shiu, Multi-scale patch based collaborative representation for face recognition with margin distribution optimization, in: Proc. European Conference on Computer Vision, Florence, Italy, 2012, pp. 822–835.
- [27] D. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition?, in: Proc. International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 471–478.
- [28] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Application to face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 2037–2041.
- [29] S. Liao, X. Zhu, Z. Lei, L. Zhang, S. Z. Li, Learning multi-scale block local binary patterns for face recognition, in: Proc. International Conference on Biometrics, Seoul, Korea, 2007, pp. 828–837.

- [30] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang, Local gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition, in: Proc. International Conference on Computer Vision, Beijing, China, 2005, pp. 786–791.
- [31] Y. Su, S. Shan, X. Chen, W. Gao, Hierarchical ensemble of global and local classifiers for face recognition, IEEE Transactions on Image Processing 18 (8) (2009) 1885–1896.
- [32] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, B. Lu, Person-specific SIFT features for face recognition, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, Honolulu, HI, 2007, pp. 593–596.
- [33] M. Bicego, A. Lagorio, E. Grosso, M. Tistarelli, On the use of SIFT features for face authentication, in: Proc. Computer Vision and Pattern Recognition Workshop, New York City, NY, 2006, pp. 1–7.
- [34] A. M. Martínez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (6) (2002) 748–763.
- [35] J. S. Yuk, K. K. Wong, R. H. Chung, A multi-level supporting scheme for face recognition under partial occlusions and disguise, in: Proc. Asian Conference on Computer Vision, Queenstown, New Zealand, 2010, pp. 690–701.
- [36] L. Zhang, S. Shah, I. Kakadiaris, Hierarchical multi-label framework for robust face recognition, in: Proc. International Conference on Biometrics, Phuket, Thailand, 2015, pp. 127–134.
- [37] D. Yi, Z. Lei, S. Z. Li, Age estimation by multi-scale convolutional network, in: Proc. Asian Conference on Computer Vision, Singapore, 2014, pp. 144–158.
- [38] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proc. International Conference on Computer Vision, Santiago, Chile, 2015, pp. 3730–3738.
- [39] S. Kang, D. Lee, C. D. Yoo, Face attribute classification using attribute-aware correlation map and gated convolutional neural networks, in: Proc.

- International Conference on Image Processing, Quebec City, Canada, 2015, pp. 4922–4926.
- [40] E. M. Rudd, M. Günther, T. E. Boult, Moon: A mixed objective optimization network for the recognition of facial attributes, in: Proc. European Conference on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 19–35.
- [41] Y. Zhong, J. Sullivan, H. Li, Face attribute prediction using off-the-shelf cnn features, in: Proc. International Conference on Biometrics, Halmstad, Sweden, 2016, pp. 1–7.
- [42] M. M. Kalayeh, B. Gong, M. Shah, Improving facial attribute prediction using semantic segmentation, in: Proc. Computer Vision and Pattern Recognition, Honolulu, HI, 2017, pp. 1–9.
- [43] S. Ji, L. Tang, S. Yu, J. Ye, A shared-subspace learning framework for multilabel classification, ACM Transactions on Knowledge Discovery from Data (TKDD) 4 (2) (2010) 1–29.
- [44] X. Zhu, X. Li, S. Zhang, Block-row sparse multiview multilabel learning for image classification, IEEE Transactions on Cybernetics 46 (2) (2016) 450–461.
- [45] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, CNN-RNN: A unified framework for multi-label image classification, in: Proc. Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 2285–2294.
- [46] L. Zhang, S. K. Shah, I. A. Kakadiaris, Fully associative ensemble learning for hierarchical multi-label classification, in: Proc. British Machine Vision Conference, Nottingham, UK, 2014, pp. 1–12.
- [47] L. Zhang, S. K. Shah, I. A. Kakadiaris, Hierarchical multi-label classification using fully associative ensemble learning, Pattern Recognition 70 (2017) 89–103.
- [48] P. Dou, L. Zhang, Y. Wu, S. K. Shah, I. A. Kakadiaris, Pose-robust face signature for multi-view face recognition, in: Proc. Biometrics Theory, Applications and Systems, Arlington, VA, 2015, pp. 1–8.

- [49] I. A. Kakadiaris, G. Toderici, G. Evangelopoulos, G. Passalis, D. Chu, X. Zhao, S. K. Shah, T. Theoharis, 3D-2D face recognition with pose and illumination normalization, Computer Vision and Image Understanding 154 (2017) 137–151.
- [50] Z. Lei, M. Pietikäinen, S. Z. Li, Learning discriminant face descriptor, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (2) (2014) 289–302.
- [51] H. Le, I. A. Kakadiaris, UHDB31: A dataset for better understanding face recognition across pose and illumination variation, in: Proc. IEEE International Conference on Computer Vision Workshops, Venice, Italy, 2017 (In Press).
- [52] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Prc. International Conference on Multimedia, Orlando, FL, 2014, pp. 675–678.
- [53] L. Zhang, I. A. Kakadiaris, Local classifier chains for deep face recognition, in: Proc. International Joint Conference on Biometrics, Denver, CO, 2017, pp. 158–167.
- [54] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.
- [55] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (200) (1937) 675–701.
- [56] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, The Annals of Mathematical Statistics 11 (1) (1940) 86–92.
- [57] O. J. Dunn, Multiple comparisons among means, Journal of the American Statistical Association 56 (293) (1961) 52–64.
- [58] R. L. Iman, J. M. Davenport, Approximations of the critical region of the fbietkan statistic, Communications in Statistics-Theory and Methods 9 (6) (1980) 571–595.