Deep Sketch-Photo Face Recognition Assisted by Facial Attributes

Seyed Mehdi Iranmanesh, Hadi Kazemi, Sobhan Soleymani, Ali Dabouei, Nasser M. Nasrabadi
West Virginia University
{seiranmanesh, hakazemi, ssoleyma, ad0046}@mix.wvu.edu, {nasser.nasrabadi}@mail.wvu.edu

Abstract

In this paper, we present a deep coupled framework to address the problem of matching sketch image against a gallery of mugshots. Face sketches have the essential information about the spatial topology and geometric details of faces while missing some important facial attributes such as ethnicity, hair, eye, and skin color. We propose a coupled deep neural network architecture which utilizes facial attributes in order to improve the sketch-photo recognition performance. The proposed Attribute-Assisted Deep Convolutional Neural Network (AADCNN) method exploits the facial attributes and leverages the loss functions from the facial attributes identification and face verification tasks in order to learn rich discriminative features in a common embedding subspace. The facial attribute identification task increases the inter-personal variations by pushing apart the embedded features extracted from individuals with different facial attributes, while the verification task reduces the intra-personal variations by pulling together all the features that are related to one person. The learned discriminative features can be well generalized to new identities not seen in the training data. The proposed architecture is able to make full use of the sketch and complementary facial attribute information to train a deep model compared to the conventional sketch-photo recognition methods. Extensive experiments are performed on composite (E-PRIP) and semi-forensic (IIIT-D semi-forensic) datasets. The results show the superiority of our method compared to the state-of-the-art models in sketch-photo recognition algorithms.

1. Introduction

Face sketch recognition is an important problem when the photo of a suspect is not available or is captured with very poor quality. A face sketch is usually drawn by a forensic artist [38] or facial software [14] based on the information provided by a victim, or an eye-witness. Therefore, the generated sketch using the provided description of the victim is the only clue to identify the victim. An automatic matching method is necessary to identify a suspect accurately via searching the law enforcement face database or surveillance cameras using only the sketch of the suspect. The sketch recognition problem has been extensively studied in recent years [17]. Existing approaches can be divided in four different categories; hand-drawn viewed sketches, hand-drawn semi-forensic sketch, hand-drawn forensic sketch and software-generated composite sketch [24]. Due to the large phenomenological gap between sketch and photo domains, sketch recognition problem still remains a challenging task.

Forensic or composite sketches contain limited information such as a rough spatial topology of the suspect face and lack of some complementary information such as skin color, ethnicity, or hair color are noticeable. In addition, sketch recognition problems mainly focus on single sketch which can be unreliable in real-world situations. This unreliability can lead to a false identification [30]. In forensics investigation multiple sources of information such as verbal description of multiple witnesses or the verbal description and poor video surveillance can be utilized to enhance the performance of suspect identification [3, 13].

In general there are two classical ways to solve the sketch recognition problem. First approach namely generative methods transfer one of the modalities (either sketch or photo) to the other before matching [28, 39]. In the second approach, the discriminative methods utilize feature descriptors such as the scale-invariant feature transform (SIFT) [18], Weber's local descriptor (WLD) [5], and multi-scale local binary pattern (MLBP) [11]. The main drawbacks of these feature descriptors is that they might not be the optimal features for the task of sketch-photo recognition. To compensate for this, some other methods in the literature propose to extract modality-invariant features [20, 15].

Recently, deep learning methods have been widely utilized in face recognition and other classification problems [32, 26, 8, 34, 35, 16, 37] instead of classical methods [6, 2]. These methods, can also be employed for the task of sketch-photo recognition problem by learning the relationship between the two modalities. However, the problem of sketch recognition is more challenging compared to

the classical face recognition problem from the deep learning point of view. The reason behind this lies not only in the heterogeneous nature of sketch and photo modalities but also the lack of large databases in order to avoid over-fitting and local minima. For example, most of the datasets contain only one sketch per subject which makes it very challenging for a deep model to learn the robust features [12]. To avoid this, many deep techniques utilize relatively shallow model or train the network only on the photo modality [25].

Recently, in the literature soft biometric traits have been utilized jointly with hard biometrics (face photo) for different tasks such as person identification or face recognition [9]. In fact, using facial attributes in conjunction with sketch would be more advantageous since some attributes such as eye color, hair color, skin color, and ethnicity do not exist in sketch and could be considered as the complementary information. Moreover, some attributes such as wearing a hat or eyeglasses can be utilized as an auxiliary information to narrow down the suspect in the databases more accurately. Recent approaches on sketch recognition problem have mainly focused on closing the gap between the two domains of sketch and photo and use of soft biometrics has not been investigated adequately. In [21] an approach was proposed to directly use facial attributes in suspect identification without using the sketch. [19] used race and gender to narrow down the galley of mugshots for faster and more accurate matching. Mittal et. al. [24] fused multiple sketches of a suspect to increase the accuracy of their algorithm. They also employed some soft biometric traits such as gender, ethnicity, and skin color to reorder the ranked list of the suspects. Ouyang et al. [27] introduced a framework to combine the facial attributes with low-level features to fill the gap between sketch and photo modalities.

In this paper, we propose an attribute-assisted sketch recognition framework which uses relevant facial attributes, provided by a victim, to enhance the performance of our deep sketch recognition method. Our approach simultaneously learns a common embedding features of sketch and photo image by minimizing two supervisory loss functions, namely the facial attributes identification and sketch-photo verification loss functions (tasks). Attribute identification task classifies photo and attribute assisted sketch images into a set of facial attributes, while verification task is to classify a pair of sketch-photo as belonging to the same person or not. The attribute identification loss is trying to pull the common features of photo and attribute assisted sketch closer in the shared latent subspace if they belong to the same set of attributes and push them apart if they belong to two different sets of attributes. Therefore, the learned features contain rich variations and can classify the photos and sketches to the classes containing the same sets of attributes in a latent feature subspace.

In summary, the main contributions of this paper include the following:

1- We propose a novel deep learning approach utilizing the facial attributes to improve sketch-photo recognition performance.

2- We introduce a joint loss function which is based on an identification-verification framework in which the identification part is responsible for the facial attribute classification and the verification part is responsible for creating a common embedding subspace between the sketch and photo modalities. This loss function helps the proposed coupled deep architecture to produce a more discriminative embedding subspace which leads to a better sketch-photo recognition performance.

3- The proposed method is able to fuse textural information of forensic sketches and complementary facial attributes such as skin color and hair color implicitly.

The remainder of this paper is organized as follows: Section 2 introduces the motivation and presents the methodology for the proposed sketch-photo recognition framework. Section 3 gives comprehensive experimental results and analysis. Finally, in Section 4, conclusions are drawn.

2. Methodology

The network parameters in the proposed framework are learned by minimizing two supervisory loss functions namely the losses due to the sketch-photo verification and facial attribute identification tasks. In the following we describe these two supervisory loss functions in details:

2.1. Sketch-Photo Verification Task:

Sketch-photo verification is the final objective of the proposed model which is identification of the suspect sketch in a gallery of mugshots. For this reason, we coupled two VGG-16 like networks one dedicated to the photo image domain (P-DCNN) and the other one to the sketch and complementary facial attributes modalities (SA-DCNN). Each DCNN performs a non-linear transformation on the input. The ultimate goal of our proposed attribute-assisted deep convolutional neural network, as shown in Figure 1, is to find the global deep features representing the relationship between sketches and their corresponding images. In order to find the common embedding space between these two different modalities we coupled two VGG-16 structured networks (P-DCNN and SA-DCNN) via a contrastive loss function [7]. This function (ℓ_{cont}) pulls the genuine pairs (i.e., a face photo image with its own corresponding sketch image) towards each other into a common latent feature subspace and push the impostor pairs (i.e., a photo image with sketch image from another subject) apart from each other (see Fig. 2). Similar to [7], the contrastive loss is of

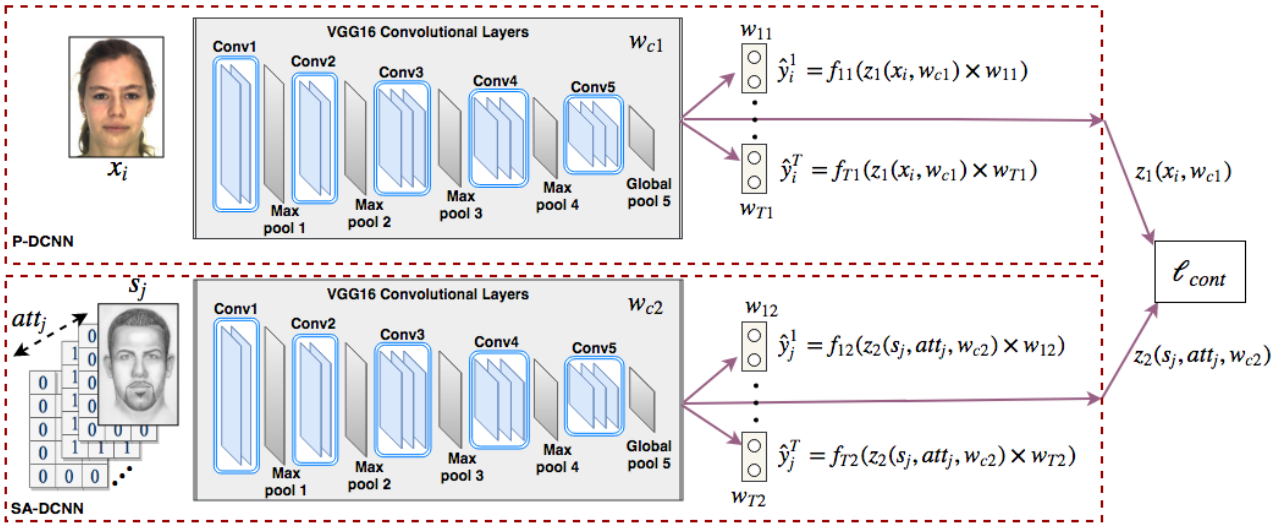

Figure 1. Attribute-assisted deep convolutional neural network. P-DCNN (upper network) and SA-DCNN (lower network) embed the photos and a pair of (sketch, attribute) into a common latent subspace.

the form:

$$\ell_{cont}(z_1(x_i), z_2(s_j, att_j), y_{cont}) = \quad (1)$$
$$(1 - y_{cont}) L_{gen}(D(z_1(x_i), z_2(s_j, att_j)) +$$
$$y_{cont} L_{imp}(D(z_1(x_i), z_2(s_j, att_j)) \ ,$$

where x_i is the input for the P-DCNN (i.e., a photo image), and (s_j, att_j) is the input for the SA-DCNN (i.e., an sketch image with its corresponding attributes provided by the eye witness). y_{cont} is a binary label, L_{gen} and L_{imp} represent the partial loss functions for the genuine and impostor pairs, respectively. z_1 and z_2 are the DNN-based embedding functions, which transform x_i and (s_j, att_j) into a common latent embedding subspace, respectively, and $D(z_1(x_i), z_2(s_j, att_j))$ indicates the Euclidean distance between the embedded data in the common feature subspace. The binary label, y_{cont}, is assigned a value of 0 when both modalities, i.e., photo and sketch, form a genuine pair, or, equivalently, the inputs are from the same subject. On the contrary, when the inputs are from different identities, which means they form an impostor pair, y_{cont} is equal to 1. In addition, L_{gen} and L_{imp} are defined as follows:

$$L_{gen}(D(z_1(x_i), z_2(s_j, att_j))) = \frac{1}{2} D(z_1(x_i), z_2(s_j, att_j))^2 \quad \text{for} \quad y_i = y_j \ , \quad (2)$$

$$L_{imp}(D(z_1(x_i), z_2(s_j, att_j))) = \quad (3)$$
$$\frac{1}{2}\, max(0, m - D(z_1(x_i), z_2(s_j, att_j)))^2 \quad \text{for} \quad y_i \neq y_j \ .$$

Therefore, the total loss function for the training dataset can be written as:

$$L_1 = 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \ell_{cont}(z_1(x_i), z_2(s_j, att_j), y_{cont}) \ , \quad (4)$$

where N is the number of samples. It should be noted that the contrastive loss function [7] considers the subjects' labels inherently. Therefore, it has the ability to find a discriminative embedding space by employing the data labels in contrast to some other metrics such as the Euclidean distance. This discriminative embedding space would be useful in identifying an sketch probe against a gallery of mugshots. However, in our framework we incorporate the facial attribute identification task in addition to the contrastive function to make the embedding space more discriminative. The facial attributes identification task assigns each sketch or image domain to a set of attributes. The attributes are predicted using both the P-DCNN and SA-DCNN networks in a multi-tasking manner. In the following subsection, we describe the multi-tasking problem in the context of attribute prediction.

2.2. Multi-Attribute Prediction and Identification Task:

The objective of this model is to predict a set of attributes using a face photo or an sketch. Therefore, in this architecture a face photo (face sketch) is presented to the network as an input and a set of attributes are predicted. Suppose the input is an image $x_i \in X$, and its class label is $y_i \in Y$ for $i = 1, \ldots, N$ where N is the number of the training

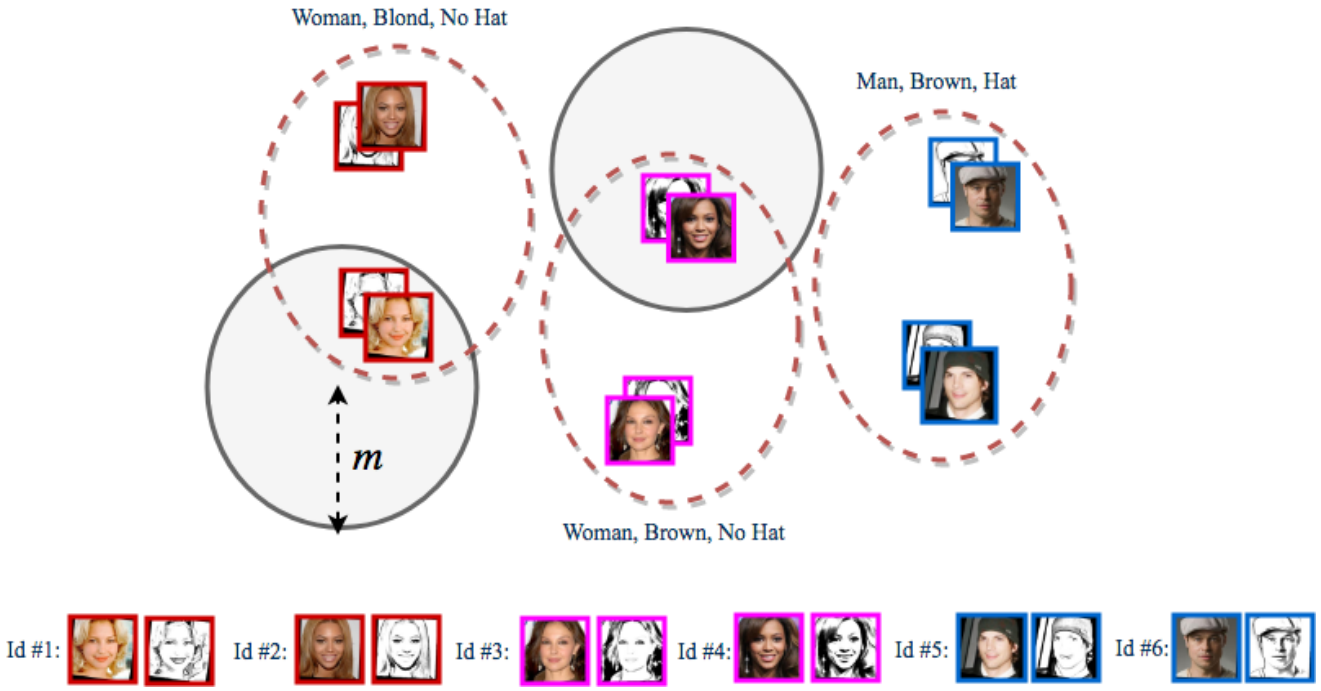

Figure 2. Visualization of the common latent subspace by leveraging facial attributes classification and verification loss functions simultaneously. Solid circles represent the contrastive margin in the embedding domain and the dashed circles depict the attributes classification. For the sake of clarity the contrastive margin is depicted for two Ids out of six Ids.

samples. Soft biometric traits, contain T different facial attributes or binary class labels provided by the eye witness. Therefore, in this framework we denote them as y^t for $t = 1, \ldots, T$. The loss function is defined as:

$$L_2 = 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(f_{t1}(z_1(x_i)), y_i^t) \ , \qquad (5)$$

where ℓ is a proper loss function (e.g., cross entropy) and $f_{t1}(z_1(x_i))$ is a binary classifier for the attribute t operated on the output of P-DCNN. Learning multiple CNNs separately is not optimal since different tasks may have some hidden relationships with each other and may share some common features. This is supported by [41] where they train a CNN features for the face recognition task and they used it directly for the face attribute estimation. Therefore, our network shares a big portion of its parameters among different tasks in order to enhance the performance of the recognition task. Thus, the loss function (5) can be reformulated as follows:

$$L_2 = 1/N \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(f_{t1}(z_1(x_i, w_{c1}) \times w_{t1}), y_i^t) \ , \qquad (6)$$

where ℓ is the cross entropy loss function. w_{c1} is the shared network parameters between all the tasks and w_{t1} represents the remaining parameters which are assigned separately for each facial attribute task.

The same procedure is performed in the other network (SA-DCNN) with an sketch as input. However, there are some attributes such as hair color and skin color which do not exist in the sketch modality while they inherently exist in the RGB images. These are the soft biometric traits which is provided by the eye witness description. Therefore, these complementary soft biometric traits are given to the SA-DCNN network which is dedicated to sketch modality. The SA-DCNN network is also responsible to estimate a set of soft biometric attributes. It should be noted that although some of the attributes in the output are given to the network from the beginning, but this attributes are fused with sketch information through the network layers. Therefore, it is worth to estimate them accurately. Also, the set of attributes which are given to the network are not necessarily the same as the set of attributes predicted by the network.

Suppose the input is an sketch $s_j \in S$, and its class label $y_j \in Y$ for $j = 1, \ldots, N$ where N is the number of the training samples. The facial attributes provided by the eye witness are also given to the network as an input, denoted as att for the sake of clarity. The loss function will be defined as:

$$L_3 = 1/N \sum_{j=1}^{N} \sum_{t=1}^{T} \ell(f_{t2}(z_2(s_j, att_j)), y_j^t) \ , \qquad (7)$$

where ℓ is the cross entropy loss function and $f_{t2}(z_2(s_j, att_j))$ is a binary classifier for the attribute t operated on the output of SA-DCNN. Here, as in P-DCNN network , we share a big portion of the network parameters among different tasks in order to enhance the performance of the recognition task. Therefore, the loss function (7) can be reformulated as follows:

$$L_3 = 1/N \sum_{j=1}^{N} \sum_{t=1}^{T} \ell(f_{t2}(z_2(s_j, att_j, w_{c2}) \times w_{t2}), y_j^t) \ , \qquad (8)$$

where w_{c2} is the shared features between all the tasks. w_{t2} represents the remaining features which are assigned separately for each soft biometric prediction task.

2.3. Total Loss Function:

The total loss function L_T for the whole framework can be written as (See Fig. 1) :

$$\begin{aligned} L_T = L_1 + \lambda_1 L_2 + \lambda_2 L_3 = \\ 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} \ell_{cont}(z_1(x_i), z_2(s_j, att_j), y_{cont}) \\ + \lambda_1 / N \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(f_{t1}(z_1(x_i)), y_i^t) \\ + \lambda_2 / N \sum_{j=1}^{N} \sum_{t=1}^{T} \ell(f_{t2}(z_2(s_j, att_j)), y_j^t) \ , \end{aligned} \qquad (9)$$

where the first term is the sketh-photo verification and the second and the third terms are the facial attributes classification loss for the P-DCNN network and SA-DCNN network, respectively. λ_1 and λ_2 are the hyper-parameters which weight facial identification cost functions of the P-DCNN network and SA-DCNN network, respectively. As it was mentioned earlier, the contrastive loss function has the ability to find a discriminative embedding space by employing the data labels. However, due to loss functions from the facial attributes classification term for photo (5) and for sketch (7), minimizing L_T will boost the discrimination in the common embedding domain. In another words, using just the contrastive loss it does not consider whether two subjects share similar facial attributes or not. Using the facial attribute classification, it enables the embedding space to be more discriminative from the attributes point of view.

Consider a subject sketch with $Id\#1$ (see Fig. 2). The contrastive loss function causes the corresponding photos from $Id\#1$ to move closer to $Id\#1's$ sketch and other Ids' photos to move farther away. Now, using the contrastive loss function in conjunction with the attribute classification makes $Id\#2$ to move closer to $Id\#1$ since they share the same set of attributes (see Fig. 2). In other words, it differentiates between different impostors of $Id\#1$. The same procedure is performed for the other identities during the training process. Figure 2 visualizes the overall concept of our joint loss function. As it is depicted, jointly training the model based on verification and facial identification will lead to a more discriminative embedding subspace which considers both the facial attributes and the geometrical relationship between the forensic sketches and photos.

During the testing phase, given a test probe with its facial attributes (s_t, att_t), the proposed AADCNN method transforms it to the common latent embedding domain, $z_2(s_t, att_t)$. In fact, after training our deep coupled network model, it has the ability to transform the photo and sketch images into a common discriminative embedding space. Therefore, the galley of the photo images is transformed to the mentioned embedding space. Eventually, the sketch image is identified, by calculating the minimum Euclidean distance between the transformed sketch prob and gallery of mugshots as follows:

$$\begin{aligned} x_i^* = \underset{x_i}{\text{argmin}} \quad D(z_1(x_i), z_2(s_t, att_t)) \\ \text{for} \qquad i = 1, 2, .., M \ , \end{aligned} \qquad (10)$$

where (s_t, att_t) is an sketch probe with its facial attribute provided by the eye witness and x_i^* is the selected matching person within the gallery of mugshots of size M.

3. Experiments and Evaluations

3.1. Implementation Details and Data Description

In this paper, we used a VGG-16 like network [33] in our sketch-photo recognition framework. The VGG-16 neural network comprised of five major convolutional components which are connected in series. The first two components, $Conv1-64$ and $Conv2-128$ consists of the following layers: a convolutional layer, a rectified linear unit layer, a second convolutional layer, a second rectified linear unit layer, and a max pooling layer. The remaining three components contain one additional convolutional layer and a rectified linear unit layer. The only difference between our CNN network and the VGG-16 is in the last component, where the last three convolutional layers of VGG16 with the size of 512, are replaced with two convolutional layers of 256, and one convolutional layer of 64 respectively for the sake of parameter reduction. Also the network uses global pooling instead of the max pooling in the last component which results

in a feature vector of size 64. The network which is dedicated to photo domain (P-DCNN) takes an RGB photo as an input and the other sketch-attribute network (SA-DCNN) gets an input consisting of multiple channels as shown in Fig. 1. The first channel is dedicated to a gray-scale sketch and the other channels are filled with 0 or 1 depending on the presence or absence of the attribute in the subject.

Experiment is performed using four main datasets, namely CUHK Face Sketch dataset (CUFS) [36] (containing 311 pairs), IIIT-D sketch dataset [4] (containing 238 viewed pairs, 140 semi-forensic pairs, and 190 forensic pairs), PRIP Viewed Software Generated Composite database (PRIP-VSGC) [14] (containing composite sketch and digital image pairs), extended-PRIP dataset (e-PRIP) [24], and unviewed Memory Gap Database (MGDB) [28] (containing 100 pairs). Since we are using the facial attribute classification in our proposed method, we utilized the CelebFaces Attributes dataset (CelebA) [22] (consisting of 200 k face images along with their attribute vectors of 40 attributes such as gender, face characteristics, skin color, hair color, etc.) to initialize the network. Since the CelebA dataset does not contain the sketch images we generated a synthetic sketch by employing xDOG [40] filter on each image. Twelve facial attributes namely bald, black hair, blond hair, brown hair, gray hair, male, Asian, Indian, White, Black, eye glasses and pale skin out of 40 attributes were selected. Since none of the sketch datasets used in this paper have any facial attribute annotation, we utilized MOON [31], which is a well-known method in facial attributes recognition, in order to annotate them.

We pre-trained our deep coupled architecture using synthetic sketch-photo pair from the CelebA dataset. We used the final weights to initialize the network in all of our training scenarios. Since our coupled DNN has a large number of parameters and the size of the sketch datasets is relatively small it is prone to overfitting. In order to avoid the overfitting problem, we utilized multiple augmentation techniques namely deformation, scale and crop, and flipping. In the following we explain each method in details (see Fig. 3).

1- Deformation: Deforms the sketch and photos to compensate for the problem of geometrically mismatching between the sketch image and its corresponding photo. The deformation is performed by translating 25 preselected points with random magnitude and direction.

2- Scale and crop: One of the main mismatch problems between sketch images and their corresponding images is the ratio deformation. To address this problem, this method upscale the sketches and photos to several random sizes, and then cut a 250×200 crop from the center of the scaled image.

3- Flipping: In this method, the images are randomly flipped horizontally.

During the training phase, instead of picking the impostor pairs randomely, we considered an strategy to select them. For each genuine pair, we considered four impostor pairs. Two of the impostors were selected among the subjects which are sharing the same set of facial attributes and the other two were selected among the subjects which have different sets of attributes. This selection technique made the framework to see more variant impostors and also helped to avoid the overfitting problem.

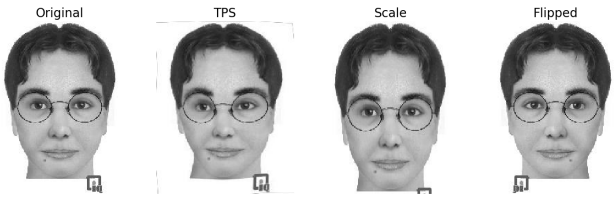

Figure 3. A sample of different augmentation techniques.

3.2. Performance Evaluation:

Our proposed framework identify a person of interest in the galley of mugshots utilizing a sketch probe and a set of facial attributes provided. In this section, we compare our approach with several state-of-the-art methods which are using both sketch and attributes and some other methods which are just using the sketch without using any attributes.

In order to evaluate performance of our method and compare with other methods, three different experiments are performed. For the sake of fair comparison, the first two experiment setups are adopted from [24]. In the first experimental setup which is the baseline (S1), the database is partitioned into two parts: training which is performed on 40% of the data and the remaining portion of the data is used for testing. e-PRIP dataset containing 123 subjects is used in this setup. Therefore, 48 identities are used in the training set and 75 subjects are considered for the testing phase. Only two out of the four different composite sketch datasets utilized in [24] are public at the time of writing this paper. These two public datasets were created by Identi-Kit tool, and FACES tool and were used by Asian and Indian artists respectively. In the second experimental setup, called S2, the gallery is extended to 1500 subjects. In this paper, the gallery is expanded utilizing WVU Multi-Modal [1], IIIT-D sketch, Multiple Encounter Dataset (MEDS) [10], and CUFS datasets. The facial attributes of the extended galley are obtained using MOON [31]. The training and probe datasets are the same as S1. The purpose of this experiment is to assess the robustness of the proposed method with a relatively large number of subject candidates. Finally, the method is evaluated on an unseen dataset. In this experimental setup (S3), we trained the network on IIIT-D Viewed, CUFS, and e-PRIP datasets and then tested it on IIIT-D Semi-forensic pairs and MGDB Unviewed. This

Table 1. Experimental Setup

Setup Name	Testing Dataset	Training Dataset	Train Size	Gallery Size	Prob Size
S1	e-PRIP	e-PRIP	48	75	75
S2	e-PRIP	e-PRIP	48	1500	75
S3	IIIT-D Semi-forensic	CUFS, IIIT-D Viewed, CUFSF, e-PRIP	1968	1500	135
	MGDB Unviewed				100

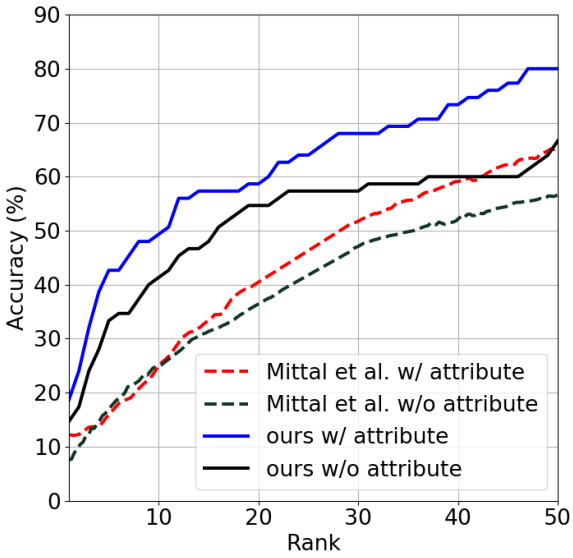

Figure 4. CMC curves of our proposed framework versus Mittal et al. algorithms [24] in the extended gallery experimental setup (S2) for the Indian dataset

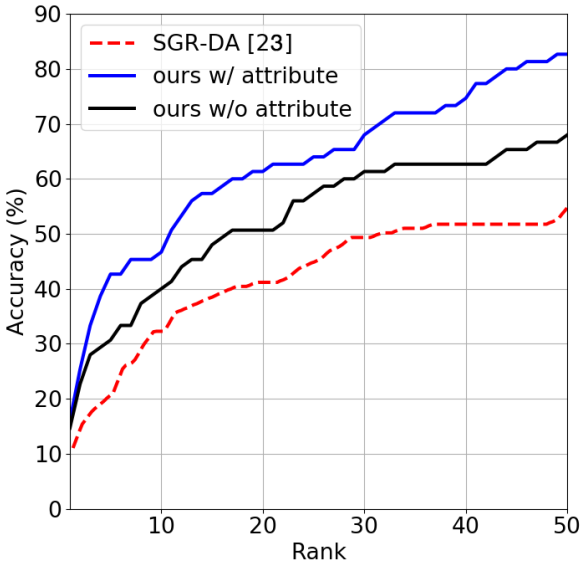

Figure 5. CMC curves of our proposed framework versus SGR-DA algorithm [29] in the extended gallery experimental setup (S2) for the Identi-Kit dataset

setup represents the level of dependency of the network on the sketch styles in the training datasets. Table 1 shows different scenarios and the size of training set, prob and gallery for each scenario.

In the experiments, different values were selected for λ_1 and λ_2. We report our best results which belong to λ_1=λ_2=1. The evaluation performance is validated using ten fold random cross validation and the results are compared with the state-of-the-art approaches.

3.3. Results:

In [24], they propose an approach called attribute feedback to study the effect of facial attributes on their recognition system. They reported the rank 10 accuracy of 58.4% and 53.1% for the prob sketches generated by the Indian (Faces) and Asian (Identi-Kit) artists, respectively. Another approach called SGR-DA [29] utilizes the sketch modality without using the facial attribute information. They reported the rank 10 accuracy of 70% on the Identi-Kit dataset. On the other hand, our proposed approach accuracy was 76.4% and 72.3% on the Faces and Identi-Kit, respectively. We also consider a baseline version of our proposed method which is only based on the contrastive loss function and does not consider the facial attributes. This way, we could observe the benefit of utilizing the facial attributes in our framework. The baseline network has an accuracy of 69.1% and 67.6%, on Faces and Identi-Kit datasets, respectively. The results demonstrate that our method outperforms all the previous methods in the literature and also express the effectiveness of our framework in utilizing the facial attributes compare to the baseline. It should be noted that, the baseline framework outperform the state-of-the-art methods except SGR-DA [29] which support the superiority of deep models over the shallow models (see Table 2).

To evaluate the effectiveness of our proposed method by using a relatively large galley of mugshots, the same experiments were performed on the extended experimental setup (S2). Figure 4 shows the results of our method as well as the other methods for the extended galley of 1500 subjects. The results depicts that our approach outperforms the method in [24] by nearly 14% for rank 50 which shows the robustness of our algorithm utilizing the facial at-

Table 2. Rank-10 identification accuracy (%) on the e-PRIP composite sketch database (S1 experimental setup).

Algorithm	Faces (In)	IdentiKit (As)
Mittal et al. [23]	53.3 ± 1.4	45.3 ± 1.5
Mittal et al. [25]	60.2 ± 2.9	52.0 ± 2.4
Mittal et al. [24]	58.4 ± 1.1	53.1 ± 1.0
SGR-DA [29]	-	70
Ours without attributes	69.1 ± 1.5	67.6 ± 1.9
Ours with attributes	$\mathbf{76.4 \pm 1.2}$	$\mathbf{72.3 \pm 0.8}$

tributes. We compared our method with SGR-DA [29] for the Identi-Kit dataset, since [24] does not provide the results on this dataset. Figure 5 shows the superiority of our proposed method compared to SGR-DA. As shown in Fig. 5, although SGR-DA outperformed our baseline network in S1 scenario (see Table 2), its results were not as promising as our proposed method in the extended experimental setup (S2). Also, our attribute-assisted method outperformed our baseline method to support the effectiveness of utilizing the attributes in relatively large gallery of mugshots as well.

Eventually, we evaluated the robustness of our proposed method in S3 experimental setup in which the network is trained on more than 1900 sketch-photo pairs and is tested on two unseen datasets, namely MGDB Unviewed and IIIT-D Semi-forensic datasets. In this scenario the gallery of mugshots was also extended to 1500. We repeated this experimental scenario for our baseline method which is not utilizing the facial attributes. As shown in Fig. 6, the proposed method showed a better performance in this scenario on both datasets compared to the baseline method indicating the advantage of facial attributes in the proposed method on unseen datasets.

4. Conclusion

We have introduced a novel approach to exploit facial attributes information for the purpose of sketch-photo recognition. The proposed network is capable of transforming the photo and sketch modalities into a common discriminative embedding subspace. We have proposed to use coupled deep neural network with facial attributes provided by eye witnesses. We simultaneously minimize the cost functions due to the facial attribute identification as well as the sketch-photo verification in order to increase inter-personal variations between different subjects with different sets of facial attributes and reducing intra-personal variations in the latent feature subspace. The combination of the two cost functions leads to a significantly more discriminative embedding subspace compared to the subspace that is created by either one of them. We compared our method with state-of-the-art sketch-photo recognition methods and showed the superiority of our method over them.

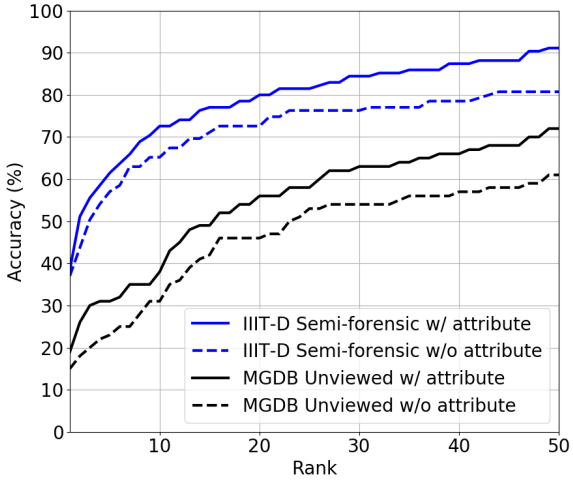

Figure 6. CMC curves of our proposed framework versus our baseline framework (without using attributes) for experimental setup (S3). The results support the robustness of our approach to different sketch styles.

References

[1] Biometrics and identification innovation center, wvu multimodal dataset. available at http://biic.wvu.edu/,.

[2] D. Alhelal, K. A. Aboalayon, M. Daneshzand, and M. Faezipour. Fpga-based denoising and beat detection of the ecg signal. In *Systems, Applications and Technology Conference (LISAT), 2015 IEEE Long Island*, pages 1–5. IEEE, 2015.

[3] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, 2014.

[4] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Memetic approach for matching sketches with digital face images. Technical report, 2012.

[5] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Memetically optimized mcwld for matching sketches with digital face images. *IEEE Transactions on Information Forensics and Security*, 7(5):1522–1535, 2012.

[6] A. Broumand, M. S. Esfahani, B.-J. Yoon, and E. R. Dougherty. Discrete optimal bayesian classification with error-conditioned sequential sampling. *Pattern Recognition*, 48(11):3766–3782, 2015.

[7] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:539–546, 2005.

[8] A. Dabouei, H. Kazemi, S. M. Iranmanesh, J. Dawson, and N. M. Nasrabadi. Fingerprint distortion rectification using deep convolutional neural networks. *arXiv preprint arXiv:1801.01198*, 2018.

[9] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. volume 11, pages 441–467. IEEE, 2016.

[10] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson. Nist special databse 32-multiple encounter dataset ii (meds-ii). Technical report, 2011.

[11] C. Galea and R. A. Farrugia. Face photo-sketch recognition using local and global texture descriptors. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 2240–2244. IEEE, 2016.

[12] C. Galea and R. A. Farrugia. Forensic face photo-sketch recognition using a deep learning-based architecture. *IEEE Signal Processing Letters*, 24(11):1586–1590, 2017.

[13] L. Gibson. *Forensic art essentials: a manual for law enforcement artists*. Academic Press, 2010.

[14] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Transactions on Information Forensics and Security*, 8(1):191–204, 2013.

[15] D.-A. Huang and Y.-C. F. Wang. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2496–2503. IEEE, 2013.

[16] S. M. Iranmanesh, A. Dabouei, H. Kazemi, and N. M. Nasrabadi. Deep cross polarimetric thermal-to-visible face recognition. *arXiv preprint arXiv:1801.01486*, 2018.

[17] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, and N. M. Nasrabadi. Attribute-centered loss for soft-biometrics guided face sketch-photo recognition. *arXiv preprint arXiv:1804.03082*, 2018.

[18] B. Klare and A. K. Jain. Sketch-to-photo matching: a feature-based approach. In *Biometric Technology for Human Identification VII*, volume 7667, page 766702. International Society for Optics and Photonics, 2010.

[19] B. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011.

[20] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1410–1422, 2013.

[21] B. F. Klare, S. Klum, J. C. Klontz, E. Taborsky, T. Akgul, and A. K. Jain. Suspect identification based on descriptive facial attributes. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–8. IEEE, 2014.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[23] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa. Recognizing composite sketches with digital face images via ssd dictionary. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, pages 1–6. IEEE, 2014.

[24] P. Mittal, A. Jain, G. Goswami, M. Vatsa, and R. Singh. Composite sketch recognition using saliency and attribute feedback. *Information Fusion*, 33:86–99, 2017.

[25] P. Mittal, M. Vatsa, and R. Singh. Composite sketch recognition via deep network-a transfer learning approach. In *Biometrics (ICB), 2015 International Conference on*, pages 251–256. IEEE, 2015.

[26] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 6670–6680, 2017.

[27] S. Ouyang, T. Hospedales, Y.-Z. Song, and X. Li. Cross-modal face matching: beyond viewed sketches. In *Asian Conference on Computer Vision*, pages 210–225. Springer, 2014.

[28] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li. Forgetmenot: Memory-aware forensic facial sketch matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5571–5579, 2016.

[29] C. Peng, X. Gao, N. Wang, and J. Li. Sparse graphical representation based discriminant analysis for heterogeneous face recognition. *arXiv preprint arXiv:1607.00137*, 2016.

[30] C. Peng, N. Wang, X. Gao, and J. Li. Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation. In *European Conference on Computer Vision*, pages 3–18. Springer, 2016.

[31] E. M. Rudd, M. Günther, and T. E. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, and N. M. Nasrabadi. Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification. In *24th International Conference on Pattern Recognition (ICPR)*, 2018.

[35] S. Soleymani, A. Torfi, J. Dawson, and N. M. Nasrabadi. Generalized bilinear deep convolutional neural networks for multimodal biometric identification. In *IEEE International Conference on Image Processing (ICIP)*, 2018.

[36] X. Tang and X. Wang. Face sketch synthesis and recognition. In *Computer vision, 2003. proceedings. ninth ieee international conference on*, pages 687–694. IEEE, 2003.

[37] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson. 3d convolutional neural networks for cross audio-visual matching recognition. *IEEE Access*, 5:22081–22091, 2017.

[38] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009.

[39] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2009.

[40] H. Winnemöller, J. E. Kyprianidis, and S. C. Olsen. Xdog: an extended difference-of-gaussians compendium including advanced image stylization. *Computers & Graphics*, 36(6):740–753, 2012.

[41] Y. Zhong, J. Sullivan, and H. Li. Face attribute prediction using off-the-shelf cnn features. In *Biometrics (ICB), 2016 International Conference on*, pages 1–7. IEEE, 2016.