IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 27, 2019, accepted May 19, 2019, date of publication June 6, 2019, date of current version September 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921382

SP-Net: A Novel Framework to Identify Composite Sketch

HAMID CHERAGHI[1] AND HYO JONG LEE[1,2]

[1]Department of Computer Science and Engineering, Chonbuk National University, Jeonju 54896, South Korea
[2]Center for Advanced Image and Information Technology, Chonbuk National University, Jeonju 54896, South Korea

Corresponding author: Hyo Jong Lee (hlee@chonbuk.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant GR2016R1D1A3B03931911.

ABSTRACT Identifying composite sketches with digital face photos is an important and challenging task for law enforcement agencies. It has attracted a wide research interest in the face recognition area. In this paper, we present a novel framework that identifies the photo corresponding to a given composite face sketch. A coupled deep convolutional neural network, named Sketch-Photo Net (SP-Net) is proposed, which is fed with a positive or negative photo-sketch pair. In the proposed SP-Net, the customized VGG-Face network is adopted as base model and is followed by two branches, namely S-Net and P-Net, for sketch and photo, respectively. The S-Net and the P-Net are able to learn discriminative features between the sketches and the photos, regardless of the appearance gap by introducing the concept of elastic learning. In other words, to extract the most important features from the input, the network needs to learn the relevant features along with the irrelevant ones. To do so, higher dimension layers are used after the three 512 layers from VGG-FaceNet. Since the network learns representative features, we decrease the dimension of the layers to produce the most representative features. In addition, contrastive loss is employed to discover the coherent visual structures between sketch and photo. Experimental results on E-PRIP face sketch dataset indicate that the proposed network significantly outperforms the state-of-the-art composite sketch identification methods.

INDEX TERMS Composite sketch, hand-drawn sketches, convolutional neural network, contrastive loss.

I. INTRODUCTION

Face sketch identification is a challenging computer vision task in criminal investigations. Due to the unavailability of information such as images or video recording in some circumstances, law enforcement agencies are in urgent need of an alternative method to catch suspects [1], [2]. Accordingly, an automatic algorithm is one of the police urgent needs for searching face archives to identify criminal suspects and arrest them within a short period of time. In comparison with captured faces using digital cameras that contain facial information accurately, sketches only include basic information such as shape, alongside some salient facial characteristics.

The veracity of a drawn sketch will mainly depend on the oral descriptions given as input to the artist in addition to the toolkits used by the artist for producing sketches. According to the synthesis method, the sketch can be classified into three categories: viewed sketch, semi-forensic sketch and forensic sketch.

- **Viewed Sketches:** In this modality of drawing sketch, artists can look at a photo of the subject or the subject directly. Therefore, these types of sketches contain rich details of the original subject which in turn leads to better accuracy.
- **Semi-forensic sketches:** In this method for drawing sketches, the artist is allowed to see a photo of the subject for a while and then is asked to draw the corresponding sketch of the photo after a short while based on his memory.
- **Forensic sketches:** This type of sketch can be drawn by expert artists. Here, artists can draw the sketch based on the description provided by an eye-witness.

Based on the synthesis tools, the sketch can be classified into two categories: hand-drawn sketch [3], [4] and composite sketch [5], [6]. Regardless how the sketch is being generated, the quality of the sketch will mainly depend on the experience of the artists and some other factors.

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

The interpretation of hand-drawn low-quality sketches often leads to under-constrained problems. Moreover, drawing a hand-drawn sketch is a time-consuming task even for a skilled artist.

Generating composite sketches can be done in short order with much more flexibility; this can be used if any changes are to be incorporated into the generated sketch, without involving much human interaction. In the process of generating sketches, drawing basic details that are subjective and depend on the witness and artist [7]. Since the artists need to prepare the sketches based on the verbal description offered by the witness or victim, the prepared sketch might be wrong or misinterpreted. Unlike face recognition area, in the sketch identification domain only one sketch exists for each identity which it makes unreliable in real-world situations.

II. RELATED WORKS

Due to the challenges and limitations mentioned in the introduction Section, face recognition algorithms are not robust enough to apply in sketch identification domain. Many algorithms [5], [8], [9] relied on handcrafted features before the emergence of deep learning-based approaches. Therefore, traditional sketch recognition methods such as multi-scale local binary pattern (MLBP) [10], histogram of oriented gradient (HOG) [11] and scale-invariant feature transform (SIFT) [12] were used to extract features from sketch and photo. Liu *et al.* [8] utilized HOG and SIFT feature descriptors to extract the facial features in both composites and photos. They also detected 66 facial key points and split the face into five key parts.

Han *et al.* [5] proposed a component-based framework for matching composite sketch to photo. In this approach facial landmarks are detected using an active shape model (ASM) [13]. Features from a facial component are extracted using multiscale local binary pattern (MLBP). Further, to obtain the matching results, corresponding components are matched and then fused. Mittal *et al.* [7] proposed an algorithm based on combining facial features alongside facial attributes such as skin color, gender, and ethnicity to create a powerful classifier. They also used the combination of fused DAISY [14] and HOG [11] features to extract local facial features from the salient regions.

Deep neural networks have shown remarkable progress over shallow learning methods for wide range of problems; the continuous deployment of these methods is prevalent in literature [15]–[18]. Therefore, due to the capability linked to the optimal usage of convolutional neural networks, many researchers apply deep neural networks instead of shallow learning methods in sketch identification tasks. To learn the deep shared latent subspace between sketch and photo, Kazemi *et al.* [19] utilized a coupled deep convolutional neural network model. They proposed an attribute-centered loss to train their network to match facial attributes between sketch and photo.

Galea and Farrugia [20] used a state-of-the-art model that is pretrained for face recognition by applying transfer learning to tackle the problem of forensic sketch recognition. In addition, to prevent over-fitting, a three-dimensional morphable software was used to synthesize new images and artificially expand the training data. Iranmanesh *et al.* [21] proposed a coupled deep neural network architecture which utilizes ethnicity, hair, eye, and skin color. They also introduced a joint loss function which is based on an identification-verification model to identify facial attributes and verify a common embedding subspace between sketch and photo.

In this paper, we propose a novel approach based on deep neural networks to solve the problem of sketch identification. The proposed coupled deep convolutional neural network named Sketch-Photo Net (SP-Net) uses a pretrained VGG-Face neural network as a base model. The SP-Net is inspired by the Siamese neural network [22] and contains two subnetworks, namely S-Net, to extract features from sketches and P-Net, to consider photos. This network is fed a pair consisting of a sketch and its corresponding photo, which we refer to as a positive pair. Then to allow the network to learn irrelevant features, a non-corresponding photo to sketch a named negative pair is also used as an input to the SP-Net. We also utilize the contrastive loss function [23] to make sure that the positive pair achieves a higher prediction accuracy than the negative pair.

The S-Net and the P-Net showed the ability to learn distinguishable features between the photo and the sketch, regardless of the modality gap by introducing the concept of elastic learning. Our aim is to extract important and useful features from the input and also to enable the network to learn the most representative features among them, including either the relevant or the irrelevant ones. This can be done by using layers with a higher dimension after the three 512 layers in the VGG-FaceNet. As the network learns all the features, some of which being the least representative, we decrease the dimensionality of the layers. This way, we are able to produce the most representative features and our network can achieve much better learning.

The contributions of the proposed method are summarized as follows:

1) We propose a novel sketch representation learning framework named SP-Net that is based on deep convolution neural networks to solve sketch identification problems.

2) To discover the shared latent structure between sketch and photo we present constructing image pairs to learn the discriminative feature representation.

3) The proposed network is fully able to boost the benchmark for sketch identification; in addition, in terms of recognition evaluation metric, it outperforms the state-of-the-art performance.

4) To evaluate and guarantee the identification accuracy we have created 2282 composite sketches for different face photo datasets using the FACES [24] software.

The reminder of this paper is organized as follows:

Section III presents the proposed composite sketch identification network. Section IV introduces the datasets, evaluation, and the experimental details. In Section V, we show the

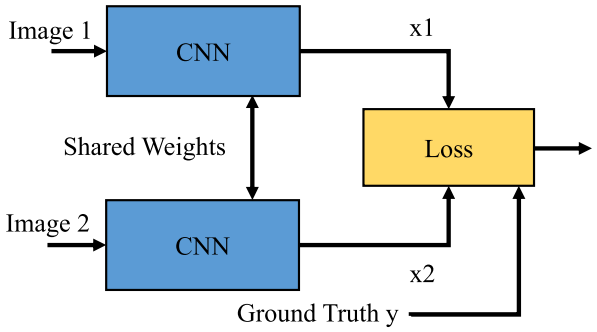

FIGURE 1. Architecture of the Siamese neural network.

superiority of our method compared with the state-of-the-art algorithms. Finally, Section VI concludes this paper.

III. PROPOSED METHOD

A. SIAMESE NEURAL NETWORK

The Siamese neural network is a class of neural network architectures that contains two or more identical subnetworks. Identical here means that they have an identical configuration with identical parameters and weights. Siamese neural networks are popular among tasks that involve finding a similarity or a relationship between two comparable items. A Siamese neural network requires a pair of input, for example, similar or nearly similar images, which is called a positive pair and dissimilar images, or a negative pair, for learning the distance margin. Figure 1 shows a typical Siamese network with two images as input and a ground truth y, to determine the minimum or maximum distance between the two images.

Generally, a Siamese network can be defined as a function of f that can map each input I into an embedding position x, given parameters θ. $X = f(I; \theta)$. Imagine that the parameter vector θ contains all the weights and biases for the convolutional and interior product layers; typically, it will contain 1M to 150M parameters depending on the size of the network. The main purpose of this function is addressing the parameter vector θ such that the embedding produced through f has desirable properties and place similar images nearby. Selecting the right pairs of images in the training part in the Siamese network can be essential to attain the best performance for the network and faster model convergence.

B. PROPOSED S-NET AND P-NET

As there is a significant appearance difference between sketches and images, traditional methods cannot be applied directly to the sketch identification task. To address the problem of appearance gap between a sketch and a photo, the pretrained VGG-Face network is customized to develop the proposed neural network. Therefore, the SP-Net is composed of S-Net and P-Net. To consider digital photos, the final three convolutional layers of VGG-Face are replaced with one convolution layer of depth 4096, two convolution layers of depth 2048 and two convolution layers of depth 1024. To attain relative similarities, the proposed P-Net should extract features to narrow down the appearance difference between sketch and image.

Considering the limited information on the composite sketches, after removing the final three convolution layers of VGG-Face, three convolution layers of depth 4096, 2048, and 1024 were added respectively. From a loss back propagation perspective, the influence of the sketch on losses should be more pronounced than the photo to distinguish joint latent structures.

C. LOSS COMPUTATION

Using deep learning techniques, losses vary from one model to the other. For instance, triplet loss [15], hinge loss [25] and contrastive loss [23] can be used to get the required output of the network. In order to measure the similarity between the pair consisting of a photo and a sketch, we deployed the contrastive loss.

The features extracted from the sketch and the photo are compared using contrastive loss. In the case of a positive pair, the target is defined as zero, as both inputs are identical. For the negative pairs, the distance between the pair of latent is larger than zero and can be assumed to be one in the case of the Cosine distance or the regularized Euclidean distance. As mentioned already, the main goal of the Siamese neural network is to differentiate between input images. Hence, a classification loss such as cross entropy [26], which measures the performance of a classification model whose output is a probability value between zero and one cannot be the correct selection. Instead, this architecture is better suited to a contrastive loss. Intuitively, measuring how well the network can distinguish a given pair of images is the aim of contrastive loss.

The contrastive loss is formularized as a function $L(w, Y, x_p, x_s)$ that is able to evaluate the similarity between x_s and x_p, where x_s and x_p are the outputs of the S-Net and P-Net, respectively. Equation 1 defines the parameterized distance function to be learned, D_w, between x_p and x_s as the Euclidean distance. The function L holds similar images close and keeps dissimilar images further apart. Equation 2 exhibits the definition of the loss on training pairs of images and sketches where p shows the number of training pairs and i indicates the index of pairs. We defined Y as a binary label for pairs.

$$D_w\left(x_p, x_s\right) = \left(\left\|x_p - x_s\right\|_2\right) \tag{1}$$

$$L\left(w, Y, x_p, x_s\right) = \sum_{i=1}^{p} L\left(W, (Y, X_P, X_s)^i\right) \tag{2}$$

In Equation 3, L_P is used for a positive pair and L_N is linked to negative pairs. The label $Y = 1$ (negative case) is used for dissimilar images and the corresponding sketch or the negative pair and the label $Y = 0$ (positive case) for similar images and the corresponding sketch or the positive pair. In a positive pair case, the right-hand additive part is set to zero. Therefore, the output of the loss is calculated through the distance between two similar pairs. If the training image and the sketch are similar, the distance will be decreased as

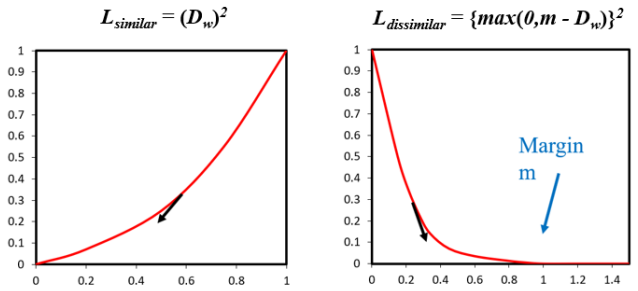

FIGURE 2. Contrastive behavior for similar and dissimilar pairs.

the SP-Net learns. In Equation 4, while the pair consisting of a sketch and a photo are dissimilar, $Y = 1$, and the first term of the equation is set to zero; meanwhile, m is selected as the threshold between positive and negative pairs. A larger m pushes similar and dissimilar pairs further away. In this work, m has been selected as being equal to one. The learning of the network should not be overly impacted by setting m to any value; therefore, we adjusted m to this value based on the best results.

$$L\left(w, Y, x_p, x_s\right) = (1 - Y) \cdot L_P\left(D_w^i\right) + Y \cdot L_N\left(D_w^i\right) \quad (3)$$

Consequently, the contrastive loss function can be defined as follow:

$$L\left(w, Y, x_p, x_s\right) = (1 - Y) \cdot (D_w)^2 + Y \cdot \{\max\left(0, m - D_w\right)\}^2 \quad (4)$$

Equation 4 represents the two states for different inputs in Pair-Network based on Figure 2. The contrastive loss function for the similar pairs is shown in Figure 2. Therefore, per each positive and negative pair of inputs there are two different values for the loss function. I_P and I_S are defined as the image and sketch inputs respectively.

$$L\left(I_p, I_s\right) = \begin{cases} \left(||x_{p+} - x_s||_2\right)^2 & Y = 0 \\ \left(\max\left(0, m - ||x_{p-} - x_s||_2\right)\right)^2 & Y = 1 \end{cases} \quad (5)$$

D. PROPOSED NETWORK ARCHITECTURE

To tackle the sketch identification challenge, we propose a coupled deep convolutional neural network named SP-Net. The main objective of the proposed SP-Net is to find the photos that are most similar to the sketch. Compared to the state-of-the-art networks, the proposed SP-Net contains two subnetworks: S-Net and P-Net. The proposed S-Net is used for extracting the features from composite sketches and the P-Net is applied on positive and negative photos. The novelty of the proposed SP-Net is that the S-Net and P-Net do not share parameters. This means that they are different networks and they work on different types of inputs.

The input of the network is a pair of images and sketches. In this work, a negative pair and a positive pair are used as inputs to the network. A positive pair contains a sketch and its corresponding photo and a negative pair includes a sketch and a photo that does not correspond to the sketch. Figure 3 displays different types of input pairs to the SP-Net.

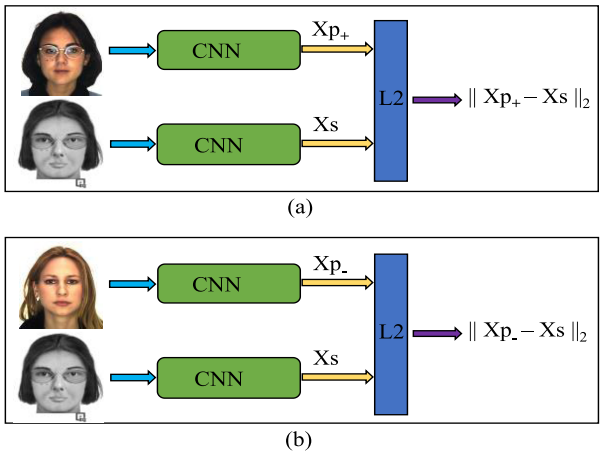

FIGURE 3. Positive and negative input pairs to the proposed SP-Net. (a) A positive input pair includes a sketch and its corresponding photo, (b) a negative input pair consists of a sketch and a photo that does not correspond to the sketch.

As can be seen in Figure 3, X_P is defined as features that are extracted from the photos corresponding, and not corresponding, to the sketch using P-Net. X_S is also defined for features that are extracted from sketches using S-Net. Unlike in the face recognition field, in sketch identification only one sketch exists for each identity. The problem of having just one sketch for each identity is that it creates difficult conditions for the network to learn distinct features. To alleviate this problem and help the network to learn more irrelevant features, negative photos are used as an input to the model.

The distance between X_P and X_S is calculated using the Euclidean metric which in the identification approach determines the closest image to each sketch. In this work, the problem is modeled as one of sketch-photo identification-similarity. Thus, the proposed SP-Net compares the input sketch with all photos in a dataset aiming to find the photo corresponding to the given sketch. It can basically be defined as a $(1 \times N)$ comparison. As mentioned before the proposed network is inspired by the Siamese network architecture which has two CNNs with which that they are trying to minimize the output of a contrastive loss for positive pairs and also to maximize that for negative pairs. The overall architecture of the proposed SP-Net can be visualized in Figure 4. In Figure 4, I_{P+} belongs to the positive images and I_{P-} belongs to the negative images. In addition, I_S can be defined as a sketch input of the SP-Net.

The proposed SP-Net includes a VGG-Face network as its base model followed by two branches, S-Net and P-Net for sketch and photo respectively. The S-Net and P-Net showed the ability to learn distinguishable features between the sketch and the photo, regardless of the appearance gap by introducing the concept of elastic learning. This concept is based

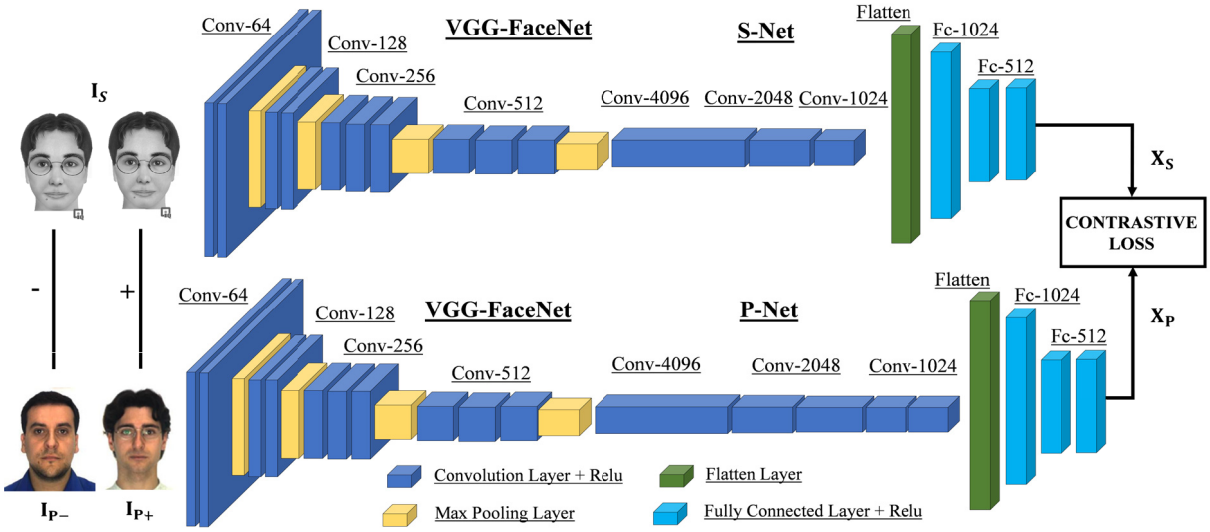

FIGURE 4. Architecture of the proposed coupled deep convolutional neural network. S-Net (upper network) and P-Net (lower network) to learn distinguishable features between the sketch and the photo, regardless of the appearance gap by introducing the concept of elastic learning.

on extracting the most important and representative features from the input.

The Elastic learning allows the network to learn the irrelevant features alongside the relevant ones, by increasing the feature map's dimensionality following the three 512 layers in the VGG-FaceNet. It is then scaled down to a lower dimension, so the most representative features are produced. The rationale for picking out different numbers of convolution layers in the S-Net and P-Net is that photos include much more information than sketches. Therefore, in this work different numbers of convolution layers are selected to limit the split between sketches and photos. In addition, the impact of sketches on loss should be more pronounced than that of photos to determine the common latent structures.

To adopt the sketch identification task, a contrastive loss function is used to discover the discriminative features of different training pairs. In order to calculate the distance, features should be reshaped and flattened. Three fully connected layers are also used to maintain the feature as much as possible. The proposed SP-Net is designed to ensure that the positive image pair is close to zero and the negative image pair is close to one.

IV. EXPERIMENTS

A. DATASETS DESCRIPTION

We evaluated the performance of identification sketches using the proposed network on the E-PRIP [7] dataset. This dataset contains 123 composite sketches and photos, each of them from the AR database [1]. The E-PRIP dataset can be divided into four different collections based on drawing composites by using different artists. One batch was drawn by an American user using the FACES toolkit and two collections were drawn by Asian artists using the FACES and Identi-Kit [27] toolkits. Another set was generated using FACES with an Indian user. To compare the performance of our method in terms of recognition with the stat-of-the-art algorithms, we used the Indian version of the E-PRIP dataset.

For this purpose, we also utilized multiple face datasets to generate their corresponding composite sketches to guarantee the performance of the proposed SP-Net. All sketches had been generated using the FACES toolkit by experts at the Chonbuk National University (CBNU). To draw composite sketches we selected some datasets such as: the Chinese University of Hong Kong (CUHK) [1], FEI [28], CASPEAL [29] (only 561 identities were drawn), MGDB [30], AR, FERET [31] (only 166 identities were drawn), and SCface [32]. The information in some of the datasets used in this study is described below.

The FEI face database contains 200 individuals. The number of male and female subjects is the same and equal to 100. In this dataset, all the images are color and taken against a white homogenous background. The original size of each image is 640×480 pixels. For each face in this dataset, there is a sketch drawn by the FACES software at the CBNU. The CUHK Face Sketch database (CUFS) is aimed at research on face sketch synthesis and face sketch recognition. This dataset includes 188 subjects from the CUHK student database. The CUHK face dataset contains three sub datasets such as AR with 123 subjects, XM2VTS [1] with 295 subjects and CUHK with 188 subjects.

The National Hi-Tech Program and ISVISION by the Face Recognition Group of JDL, ICT, and CAS have created a face dataset called CASPEAL. Currently, the CASPEAL face database [29] contains 99,594 images of 1040 individuals with 595 males and 445 females. The images in this dataset were captured with varying pose, expression, accessory, and lighting (PEAL). Among all images in this dataset, 561 composite sketches have been drawn by utilizing the FACES software at CBNU.

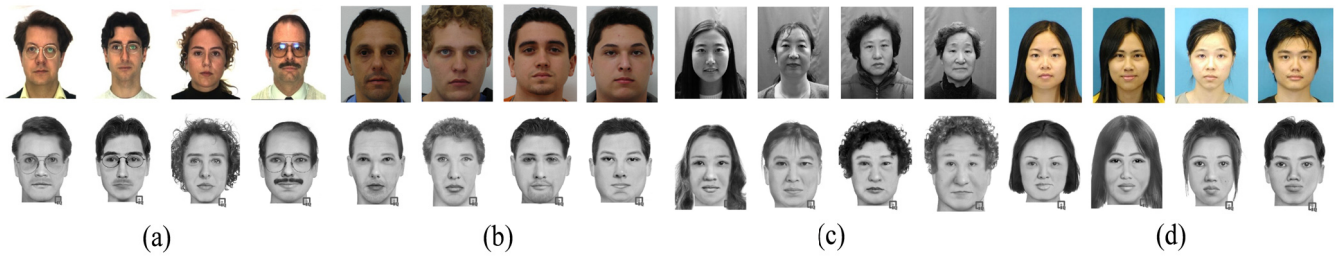

FIGURE 5. Examples of photos and corresponding composite sketches from different datasets drawn by experts at the Chonbuk National University. (a) Photos and composites from an AR dataset, (b) generated composites for images in the FEI dataset, (c) shows the sample of photos and composites from the CASPEAL dataset, and (d) photos and corresponding composites sketches from the CUHK dataset.

TABLE 1. Summary of datasets for sketch identification.

Dataset	Detail		No. of Pairs
	Male	Female	
E-PRIP	70	53	123
CUHK	134	54	188
FEI	100	100	200
CASPEAL	118	443	561
MGDB	65	31	96
AR	70	53	123
FERET	110	56	166
SCface	110	13	123

All possible datasets and their pairs of images and sketches are summarized in Table 1. Figure 5 exhibits several photo samples and their corresponding composite sketches from some of datasets mentioned above. All the composite sketches drawn using the FACES toolkit where only the E-PRIP dataset is publicly available and the others are private.

B. DATA AUGMENTATION

As the number of sketch datasets available is limited, sketch identification became a challenging task in various aspects. To mitigate the issue of limited availability for sketch datasets and to prevent overfitting in this study, we utilized data augmentation techniques such as flipping and scaling. Therefore, composites are scaled by 0.2 percent and flipped horizontally. Figure 6 illustrates the original image with six similar sketches that are generated through data augmentation.

C. DATASET SETUP

The gallery set for all experiments includes the whole dataset, as explained in Section III A of this paper. We used the Indian version of the E-PRIP dataset with 123 identities to compare the performance of our network against state-of-the-art methods.

In the first setup, called D1, 48 subjects were used for training, and 75 subjects selected randomly to test the network in which composites are drawn by FACES software. D2 is the second experiment to define the CUHK dataset with 188 pairs where the composites are drawn using the FACES toolkit at the CBNU. From this dataset 88 identities were selected randomly for the training part and 100 subjects for testing the network.

TABLE 2. Detailed information of the datasets used for training and testing.

Collections Name	Datasets Name	Training Size	Testing Size	Prob Size
D1	E-PRIP	48	75	75
D2	CUHK	88	100	100
D3	FEI	80	120	120
D4	CASPEAL	261	300	300
D5	FERET	76	90	90
D6	AR	48	75	75
D7	MGDB	40	56	56
D8	SCface	48	75	75

In the third experiment, called D3, the FEI dataset with 200 subjects, was divided into training with 80 and testing with 120 subjects. In the fourth order, 561 photos out of 1040 from the CASPEAL dataset were selected randomly to draw corresponding sketches. This setup is called D4 and is then divided into training set with 261 and testing set with 300 sketches.

In the fifth experiment, named D5, only 166 images from the FERET face dataset were used to draw corresponding composite sketches. The training set includes 76 photos and sketches, while the test set contains 90 identities. Using 123 images from the AR dataset another group called D6 is created. The generated composite from the AR dataset is divided into a training set with 48 identities and a test set with 75 identities. MGDB is another face dataset with 96 identities with their corresponding composite that are generated for each subject named D7. The training set includes 40 identities and test set consists of 56 subjects.

In the last group, called D8, the SCface dataset containing 123 characters is divided into two parts; training the network with 48 identities, and testing with 75 subjects. Table 2 provides a layout of the datasets used in our work.

V. RESULTS

A. EXPERIMENT DETAILS

The implementation of the proposed architecture was realized using the Keras [33] deep learning library and four NVIDIA

TABLE 3. Identification accuracies (%) on the E-PRIP sketch database.

Methods	Rank 1	Rank 5	Rank 10
Equal Weighted Sum	10 ± 3.7	21.6 ± 3.7	37.2 ± 3.7
Weighted Sum	4.2 ± 1.4	21.2 ± 1.4	43.1 ± 1.4
AdaBoost	6 ± 0.9	23.1 ± 0.9	45.3 ± 0.9
MCWLD [36]	4 ± 3.4	11.3 ± 3.4	24.0 ± 3.4
Mittal et al. [37]	7 ± 1.4	28.9 ± 1.4	53.3 ± 1.4
Mittal et al. [7]	12.2 ± 1.1	40 ± 1.1	58.4 ± 1.1
Mittal et al. [34]	7 ± 2.9	31.3 ± 2.9	60.2 ± 2.9
Kazemi et al. [19] with facial attributes	-	-	73.2 ± 1.1
Kazemi et al. [19] without facial attributes	-	-	68.6 ± 1.6
Proposed SP-Net using VGG-FaceNet only	13.6 ± 1.1	41.8 ± 1.1	59.18 ± 1.1
The proposed SP-Net	**28.1 ± 1.3**	**53.1 ± 1.3**	**80.0 ±1.3**

TITAN X GPUs. We resized the images and sketches to 224 × 224 during both the training and testing phases. The number of training batch sizes was to 20 while checking the best performance during training. The Euclidian distance metric (L2) is used for any incoming pairs with the learning rate set to 0.001. The best performance was found on the 100^{th} epoch and the stochastic gradient descent (SGD) used as an optimizer in this study. To help the network learning improve, 40 negative pairs and seven positive pairs using augmentation techniques were utilized for each composite sketch and corresponding photo.

B. QUANTITATIVE RESULTS

In the E-PRIP dataset, the reported accuracy [7] in rank-10 using sketches generated by the Indian artist with the FACES software was 58.4%. In addition, the accuracy [27] of the recognition sketches by the Asian user using the Identikit software was 53.1%. Regarding facial attributes they used the attribute feedback algorithm. The proposed algorithm outperforms [34] rank-10 by almost 60.2% while exploiting facial attributes. However, SGR-DA [35] without utilizing facial attributes achieved a 70% accuracy on the E-PRIP dataset using the IdentiKit software.

Compared with the mentioned results, the proposed attribute-centered loss algorithm by Kazemi *et al.* [19] reported a 73.2% accuracy using E-PRIP sketches generated with the FACES toolkit and a 72.6% accuracy for sketches drawn by the IdentiKit software. the attribute-centered loss reported 68.6% without using facial attributes. In addition, Kazemi *et al*. used a contrastive loss during the training of the proposed network and the result was a 65.3% accuracy on FACES and a 64.2% accuracy on IdentiKit.

The proposed SP-Net yielded a result that surpasses the state-of-the-art accuracy. Accordingly, the proposed deep coupled neural network illustrates the significant performance of 80.0% without utilizing any facial attributes on the E-PRIP dataset. Table 3 demonstrates the identification accuracy of the E-PRIP dataset in different ranks. In addition, the recent other methods that used deep neural networks with significant accuracy are mentioned for comparison as well.

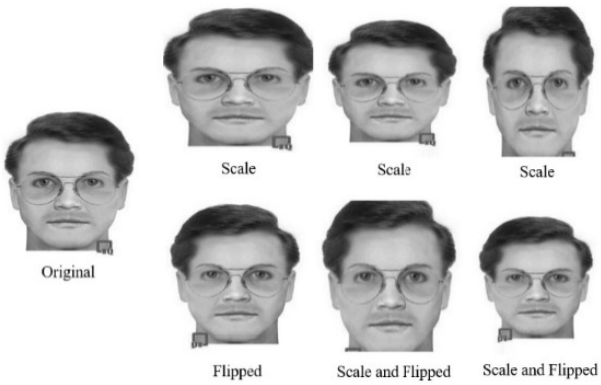

FIGURE 6. Examples of augmented composite sketches using various data augmentation techniques.

To guarantee the result of the proposed SP-Net, both sub-networks (S-Net and P-Net) only used the VGG-FaceNet structure without replacing any of the convolution layers; the performance of this modality being unremarkable compared to state-of-the-art results.

The performance of the proposed method is displayed in Figure 7, comparing it with the state-of-the-art algorithms on the E-PRIP dataset. The performance of the proposed method was compared with the state-of-the-art results through the attribute-centered loss in rank-10.

C. MORE EXPERIMENTS

In this study, we used various private sketch datasets to evaluate the robustness of our proposed deep coupled network. The CUHK dataset includes 188 subjects with 88 identities used for training and 100 subjects selected to test the network. The recognition accuracy was 66.3% in rank-10. Using 80 subjects of the FEI dataset as training samples and 120 composites for testing the network, we achieved a 65.5% accuracy in rank-10. From the 123 photos of the SCface dataset we used 48 subjects for training and 75 identities to test the network. The recognition accuracy for this dataset was 51.7% in rank-10. Among the 123 images from the AR dataset, 48 subjects were used to train the network and 75 subjects for testing. The proposed SP-Net reached a 74.2% accuracy in rank-10.

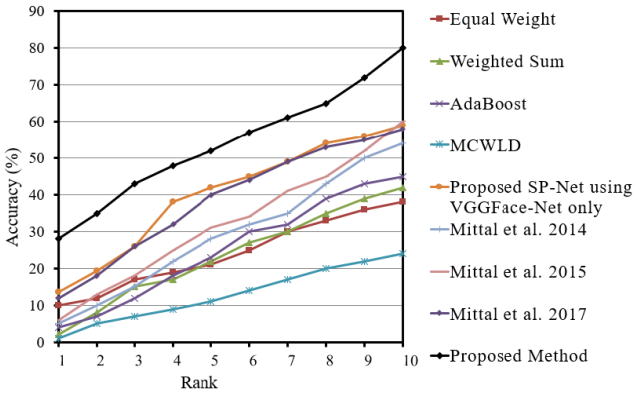

FIGURE 7. Comparison of the proposed method with other recent works in different ranks.

TABLE 4. Identification accuracy (%) on private sketch datasets in different ranks.

Dataset	Rank 1	Rank 5	Rank 10	Rank 20	Rank 30
CUHK	23.5	40.5	71.3	83.4	91.1
FEI	26.8	39.6	63.3	77.0	85.5
CASPEAL	21.8	28.8	41.5	50.8	61.6
FERET	20.1	27.0	33.6	42.2	54.1
MGDB	39.8	73.33	85.1	87.4	89.2
SCface	22.8	38.7	51.7	69.9	82.8
AR	31.6	55.6	74.2	81.2	93.0

One of the largest face datasets that is used in this work is the CASPEAL database. This dataset consists of 1040 identities, and only 561 composites of the images from this dataset have been generated. Since the quality of face images in this dataset are poor, recognition accuracy was only 41.5% in rank-10 by allocating 261 identities for the training of the proposed network and 300 identities for testing.

The FERET face dataset includes 1194 identities, but 166 composites for this dataset have been generated. For training, 76 identities were used, and 90 composites were selected randomly for testing. To evaluate the proposed network under normal conditions, faces were not cropped which may degrade recognition. The reported accuracy of sketch identification in the FERET dataset is 33.81% in rank-10. Another face dataset used in this work is MGDB, which contains 96 identities. 40 composites of this dataset were used for training and 56 composites for testing.

The recognition accuracy was 85.1% accuracy in rank-10. Table 4 provides information on the identification performance of the proposed method on eight different sketch datasets for various ranks. While the proposed SP-Net trained only on a small number of images and sketches and was tested on totally unseen data, the remarkable accuracy achieved indicates the stability and robustness of the proposed network.

D. QUALITATIVE RESULTS

Sample ranking results are presented across datasets to show the performance of the method in composite sketch identification as well as the failure identification results in Figure 8 and Figure 9. It is apparent from the ranking results that the proposed method has an adequate recognition ability due to the rank-1 retrieval. However, the failure identification cases linked to the artist drawing skills as shown in Figure 10.

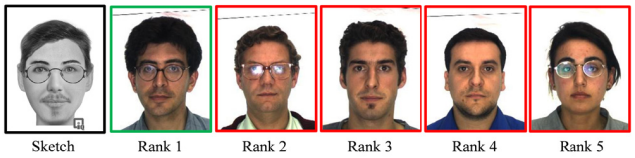

FIGURE 8. The effect of proposed network in sketch-photo identification.

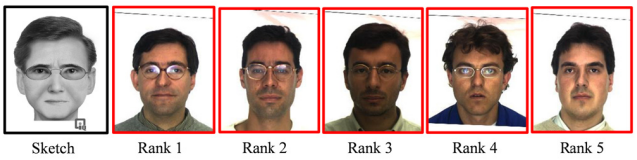

FIGURE 9. Depiction of the failure cases to recognize by proposed network.

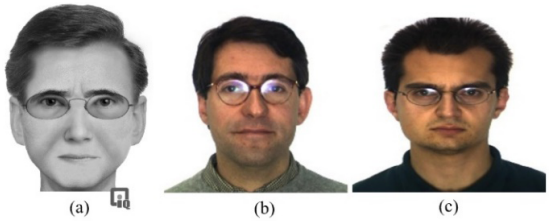

FIGURE 10. Comparison between a sketch and its corresponding photo for the failure cases. (a) Composite sketch from the E-PRIP dataset drawn by the Indian user. (b) Photo not corresponding to the sketch and (c) corresponding photo for the sketch.

As visualized in Figure 10 the sketch drawn is very similar to the non-corresponding photo compared with the original. Likewise, in Figure 9 all photos are almost similar to the sketch provided. The failure cases are mostly attributable to the sketch quality rather than the learning paradigm or the network's identification capability.

VI. CONCLUSION

We proposed a novel algorithm for the composite sketch identification task. First, a pretrained deep neural network was utilized for feature extraction from both composite sketches and photos. The proposed SP-Net consists of two subnetworks, namely S-Net and P-Net, that extract features from sketches and photos, respectively. We used a contrastive loss to learn similarity or dissimilarity between images and composites. We attempted to minimize the distance for the positive pairs and maximize the distance for the negative pairs. Accordingly, the Euclidean distance algorithm was selected as the metric for comparing features. In the test part, every composite was considered with a number of images to find the smallest distances between these based on the rank.

The proposed method advances the state-of-the-art across the E-PRIP dataset in rank-1 (28.3%), rank-5 (53.1%) and rank-10 (80.0%). To prove the robustness of the proposed

network we applied our algorithm on seven different private datasets, and we obtained significantly higher accuracies for various ranks. Further, adapting this approach to hand-drawn sketch recognition and composites generated through the Identikit software is another avenue in the future for more significant research.

REFERENCES

[1] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[2] M. Zhang, J. Li, W. Wang, and X. Gao, "Compositional model-based sketch generator in facial entertainment," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 904–915, Mar. 2018.

[3] N. Wang, X. Gao, L. Sun, and J. Li, "Bayesian face sketch synthesis," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1264–1274, Mar. 2017.

[4] L. Jiao, S. Zhang, L. Li, F. Liu, and W. Ma, "A modified convolutional neural network for face sketch synthesis," *Pattern Recognit.*, vol. 76, pp. 125–136, Apr. 2018.

[5] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.

[6] C. Galea and R. A. Farrugia, "A large-scale software-generated face composite sketch database," in *Proc. Int. Conf. Biometrics Special Interest Group*, 2016, pp. 1–5.

[7] P. Mittal, A. Jain, G. Goswami, M. Vatsa, and R. Singh, "Composite sketch recognition using saliency and attribute feedback," *Inf. Fusion*, vol. 33, pp. 86–99, Jan. 2017.

[8] D. Liu, J. Li, N. Wang, C. Peng, and X. Gao, "Composite components-based face sketch recognition," *Neurocomputing*, vol. 302, pp. 46–54, Aug. 2018.

[9] B. F. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011.

[10] C. Galea and R. A. Farrugia, "Face photo-sketch recognition using local and global texture descriptors," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, 2016, pp. 2240–2244.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[12] B. Klare and A. K. Jain, "Sketch-to-photo matching: A feature-based approach," *Proc. SPIE*, vol. 7667, Apr. 2010, Art. no. 766702.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[14] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[16] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao, "SketchNet: Sketch classification with Web images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1105–1113.

[17] T. Chugh, M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Transfer learning based evolutionary algorithm for composite face sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 117–125.

[18] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 945–953.

[19] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, and N. M. Nasrabadi, "Attribute-centered loss for soft-biometrics guided face sketch-photo recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 499–507.

[20] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1586–1590, Nov. 2017.

[21] S. M. Iranmanesh, H. Kazemi, S. Soleymani, A. Dabouei, and N. M. Nasrabadi, "Deep sketch-photo face recognition assisted by facial attributes," Jul. 2018, *arXiv:1808.00059*. [Online]. Available: https://arxiv.org/abs/1808.00059

[22] S. Appalaraju and V. Chaoji, "Image similarity using deep CNN and curriculum learning," Sep. 2017, *arXiv:1709.08761*. [Online]. Available: https://arxiv.org/abs/1709.08761

[23] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1735–1742.

[24] FACES 4.0. *IQ Biometrix 2011*. Accessed: Mar. 31, 2019. [Online]. Available: http://www.iqbiometrix.com

[25] C. Gentile and M. K. Warmuth, "Linear hinge loss and average margin," in *Proc. Adv. Neural Inf. Process. Syst.*, 1999, pp. 1–7.

[26] P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs. squared error training: A theoretical and experimental comparison," in *Proc. INTERSPEECH*, vol. 13, 2013, pp. 1–5.

[27] Identi-Kit. *Identi-Kit Solutions 2011*. Accessed: Mar. 31, 2019. [Online]. Available: http://www.identikit.net/

[28] *FEI Face Database*. Accessed: Mar. 31, 2019. [Online]. Available: https://bit.ly/2QCpB08

[29] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.

[30] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li, "Forgetmenot: Memory-aware forensic facial sketch matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5571–5579.

[31] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 513–520.

[32] M. Grgic, K. Delac, and S. Grgic, "SCface–surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, 2011.

[33] *Keras Documentation*. Accessed: Mar. 31, 2019. [Online]. Available: https://bit.ly/2HQRKS9

[34] P. Mittal, M. Vatsa, and R. Singh, "Composite sketch recognition via deep network—A transfer learning approach," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 251–256.

[35] C. Peng, X. Gao, N. Wang, and J. Li, "Sparse graphical representation based discriminant analysis for heterogeneous face recognition," Jul. 2016, *arXiv:1607.00137*. [Online]. Available: https://arxiv.org/abs/1607.00137

[36] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1522–1535, Oct. 2012.

[37] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa, "Recognizing composite sketches with digital face images via SSD dictionary," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep./Oct. 2014, pp. 1–6.

HAMID CHERAGHI received the B.S. degree in computer science from Islamic Azad University, Iran, in 2014. He is currently pursuing the M.S. degree with the Division of Computer Science and Engineering, Chonbuk National University, Jeonju, South Korea. His research interests include machine learning, computer vision, and image processing.

HYO JONG LEE received the B.S., M.S., and Ph.D. degrees in computer science from The University of Utah, USA, where he was involved in computer graphics and parallel processing. He is currently a Professor with the Division of Computer Science and Engineering and the Director of the Center for Advanced Image and Information Technology, Chonbuk National University, Jeonju, South Korea. His research interests include image processing, medical imaging, parallel algorithms, deep learning, and brain science.