AN IMPROVED SIAMESE NETWORK FOR FACE SKETCH RECOGNITION

LIANG FAN, HAN LIU, YUXUAN HOU

School of Computer Science and Informatics, Cardiff University,
Queen's Buildings, 5 The Parade, Cardiff, CF24 3AA, United Kingdom
E-MAIL: FanL7@cardiff.ac.uk,liuh48@cardiff.ac.uk,vernonhou@gmail.com

Abstract:

Face sketch recognition identifies the face photo from a large face sketch dataset. Some traditional methods are typically used to reduce the modality gap between face photos and sketches and gain excellent recognition rate based on a pseudo image which is synthesized using the corresponded face photo. However, these methods cannot obtain better high recognition rate for all face sketch datasets, because the use of extracted features cannot lead to the elimination of the effect of different modalities' images. The feature representation of the deep convolutional neural networks as a feasible approach for identification involves wider applications than other methods. It is adapted to extract the features which eliminate the difference between face photos and sketches. The recognition rate is high for neural networks constructed by learning optimal local features, even if the input image shows geometric distortions. However, the case of overfitting leads to the unsatisfactory performance of deep learning methods on face sketch recognition tasks. Also, the sketch images are too simple to be used for extracting effective features. This paper aims to increase the matching rate using the Siamese convolution network architecture. The framework is used to extract useful features from each image pair to reduce the modality gap. Moreover, data augmentation is used to avoid overfitting. We explore the performance of three loss functions and compare the similarity between each image pair. The experimental results show that our framework is adequate for a composite sketch dataset. In addition, it reduces the influence of overfitting by using data augmentation and modifying the network structure.

Keywords:

Siamese network; Modality gap; Data augmentation; Machine learning; Image classification

1. Introduction

Face sketch recognition refers to matching the face photo from a huge dataset based on a given face sketch image. This technology has a wide application in criminals. Especially, criminal suspect photos cannot be captured directly at the crime scene. The police have to make a hand-drawn face sketch image or a composite sketch which is gained by software (such as IdentiKit, FACES 4.0, Mac-a-Mug, Photo-Fit and EvoFIT) by the description from eyewitness and matches it from the dataset. In contrast to human, computers cannot compare the similarity between different modalities' images directly, because of the difference of the visual information which can be perceived from face sketch and photos. For modal interference, the distance between the cross-modal classes of the face feature and the distance between classes cannot be divided. The main challenge is the dissimilarity of the feature representation between photos and sketches because of the modality gap.

Face sketch recognition can be divided into two types: the traditional learning methods and the deep learning methods. One traditional category is to synthesize a pseudo image which is adapted from the other modality image to the real image in order to reduce the distance between different modality images [1]. Another method which projects the photo and the corresponding sketch into a common space bypasses generating a pseudo image. The aim of the method is to maintain intra-class compactness and interclass separability of the entire dataset [2]. The third category is feature-based method that extracts effective features to measure the similarity between the photo and the corresponding sketch using some feature descriptors [3], [4], [5]. Compared with the above-mentioned three methods, deep learning methods extract more effective features and textures. One category is to synthesize images from corresponding modality images, whereas the other one is based on the generative adversarial network that is used to synthesize similar pseudo images with the given real photo [6] [7] [8]. Another is to compare the different modalities' images directly in [2]. Kazemi et al. in [10] designed a new loss function for Siamese network to enhance the recognition accuracy. This loss function consists of two parts, i.e., one is attribute loss to minimize the intra-class distances of photos or sketch-attribute pairs which share the similar combination of facial attributes, whereas another part prevents pushing the centers to keep a minimum distance if the centers converge to a single point in the embedding. Different from traditional methods, the performance of deep learning methods was evaluated by using composite face sketch datasets. However,

some of these methods cannot keep the relationship between modalities' images using a threshold value after training. The accuracy using deep learning methods is all less than 70%.

The case of overfitting and unbefitting loss functions causes low recognition accuracy. At the same time, limited methods of deep learning are used in face sketch recognition. There are three contributions in this paper:

1. A cross-modal loss function is defined to eliminate the interference of modalities in the sample, and it is more effective to measure the distance between features in different modes. The loss function projects the features from two modalities into a common subspace.

2. Based on the loss function, a network is designed to compute the distance between the inter-modal class and the intra-modal class, and there is an interval between different modal samples. The nearest neighbor (NN) classification methods are used to compare the similarity using a threshold value and increase the recognition accuracy.

3. Due to the case that the number of cross-modal homogeneous distance constraints and the number of different types of distance constraints that are usually constructed are severely unbalanced, a constraint is used to reduce the influence.

The rest of this paper is organized as follows: Section 2 shows the pre-processing methods relating to the proposed method. In Section 3, we describe our proposed network which involves combining one feature extracted from each Siamese network's channel and another one (the compress feature) from the spare autoencoder network using HAOG feature. In Section 4, we compare the performance obtained using three loss functions with the one obtained using the state of the art one. In Section 5, this paper is concluded by stating the contributions and further directions.

2. Preprocessing methods

In this section, we show a 3D face model to increase the number of photos, and a basic Siamese network is built to compare the similarity between photos and sketches.

2.1. Data augmentation

The overfitting case which results from the small size of data is the main reason that the recognition accuracy on the test set is lower than the one on the training set in deep learning methods. Especially for face sketch recognition, the overfitting phenomenon is more serious, since the number of instances for each subject is limited. The existing face sketch datasets contain only one photo and one corresponding sketch in each subject. Moreover, the size of each face sketch dataset is small. The data augmentation methods such as rotation, scaling and transforming are useful in the classification problem. However, this method is not suitable for face sketch recognition, since these rotations may produce some unnecessary and negative noise to raise the complexity of the network. One reason is that sketch images, which consist of lines and the shapes, are too simple to be used for extracting available features, given that the information of the sketches is less than the one of the face photo. The second reason is that some of data augmentation methods are infeasible for face recognition, such as vertical and titled rotation. A valid method which can advance the capacity of the dataset is to generate face images of different directions.

We utilized a synthesized 3D face model which was proposed by Bas et al. in [1] to generate different directions face images from a single image. This method uses image edges for face model fitting and syntheses a 3D face model. Then, the 2D face images are obtained after different directions of the 3D model face are obtained from different rotation angles. Despite the edge information and the hair space information loss, it increases the number of instances for each subject. After data augmentation, each subject involves four generated images and the original one.

2.2. Siamese network

It is clear that using a large amount of data can avoid the overfitting problem in training the deep neural network and increases the recognition rate. The first reason is that the trained model cannot display the integral performance since the use of small training data leads to the large model space. Although an effective way of better fitting data is to obtain a huge model space, if the model space is too huge, the chance of selecting a suitable model may be reduced. The risk of selecting the parameters which lead to poor performance on test data may be reduced more effectively.

It is clear that increasing the amount of data is necessary to avoid the overfitting problem in training the deep neural network and to increase the recognition rate. However, the sizes of face sketch datasets are all small and these datasets include one photo and one sketch for each person, so most of the deep neural networks cannot be applied effectively. The Siamese network is utilized by involving two same neural networks as two channels to extract features from two images and compares the similarity after training by using a contrastive loss function. The input shape of the Siamese network is an image pair which consists of two images and the label. Each image needs to be made up of an image pair with other images. The network provides an effective way to compare the similarity and alleviates the overfitting problem.

3. Proposed method

The proposed neural network architecture consists of two identical convolutional networks as channels which accept the distinct images as inputs and shares weights to extract features from face photos and sketches, respectively, in order to reduce the distance between features.

3.1. The network architecture

We adopt to share the weights between the two channel's convolutional networks to ensure the same location of different features which map two input images into a common space, respectively, using the contrastive loss function. After the last layer for each channel of a Siamese network, the outputs of the sub-networks of this network are fused and trained using the contrastive loss function to reduce the distance between positive image pairs and increase the distance between negative image pairs.

The modality-invariant parameters of the Siamese network allow learning the relationship from an image pair which consists of the two modalities' images for each subject and apply the learned relationship for testing. Due to the lack of enough texture information of the sketches, we utilize a large kernel size to capture texture information. Except for the first convolution layer without padding, the kernel size is 7*7 for other convolution layers. All convolution layers are set to involve exclusively rectified linear unit (RELU) to make nonlinear mapping which can keep the learning rate faster than other activation functions and avoids saturation though predigest the process of backpropagation. The board of images does not need any padding of the three convolution layers. It means that all dimensions are valid so that the input image gets covered by the filter and the stride. The filter window stays at a valid position inside the input map, so the output size shrinks 1 size than the filter. In the last forth convolution layer, padding is used to reduce the output dimensions. Even if it may lose some features on the image borderline that has less information on face photos and sketches. One advantage is that padding leads to the reduction of the output dimensionality, and another one is that the number of parameters is smaller.

However, these methods cannot reduce the influence of overfitting in this network, because of the small face sketch dataset. To address the overfitting problem further, the fully connected layer is removed to reduce the number of parameters. Moreover, L2 regulation, which decreases the weight to reduce the complexity of the network as a penalty function, is used on weighting to avoid the risk of overfitting. And the dropout is set as 0.2 before the last convolution layer. It diminishes the number of parameters though throwing units before connecting to the next layer of the neural network during training. In the last layer, a convolution layer is used instead of a fully connected layer. The memory of the convolution layer is smaller than the one of the output, since shaping the output from a matrix into a column vector leads to the decrease of redundant parameters comparing with a fully connected layer and avoiding the case of overfitting. In addition, there is one face photo and one sketch for each person on the face sketch datasets. It causes that the number of positive image pairs is far less than the number of the negative image pairs, i.e., it is not suitable to minimize the distance between different samples for imbalanced data.

Due to the small size of data, the Siamese network cannot extract more effective and similar features from photos and sketches, we proposed to fuse other types of features [4] for increasing the recognition accuracy. The HAOG feature is provided as a suitable feature to match the images for face sketch recognition directly. However, the features are too redundant to help improve the recognition accuracy after fusing the HAOG feature and the features extracted from the two channels. In order to reduce the size of the HAOG feature, we built two spare autoencoder networks to compress the HAOG features separately. Each spare autoencoder network which consists of six convolution layers and several max-pooling layers. Based on this characteristic, our network is built to learn feature vectors by minimizing the discrepancy between the features extracted from the convolutional network and the original features which are generated using this feature descriptor. The architecture of the spare autoencoder network is followed as:

C1: The output filter size: 3, kernel size: 7*7, padding: valid, activate function: Relu.

Downsampling: pooling size: 7*7, padding: same.

C2: The output filter size: 3, kernel size: 5*5, padding: valid, activate function: Relu.

C3: The output filter size: 3, kernel size: 3*3, padding: valid, activate function: Relu.

Encoding layer: pooling size: 7*7, padding: same.

C4: The output filter size: 3, kernel size: 3*3, padding: valid, activate function: Relu.

C5: The output filter size: 3, kernel size: 5*5, padding: valid, activate function: Relu.

Upsampling: pooling size: 7*7, padding: valid.

Decoding Layer: The output filter size: 1, kernel size: 7*7, activate function: Sigmoid, padding: same.

In this spare auto-encoder network, 'RELU' is used as an activate function to reduce the difference between the encoder's output and the decoder's output. Moreover, the padding of each layer is all 'valid' which means the output feature map needs to fill up '0' before sending the output to the next layer to keep the dimensionality unchanged between the encoder stage and the decoder stage. After training the

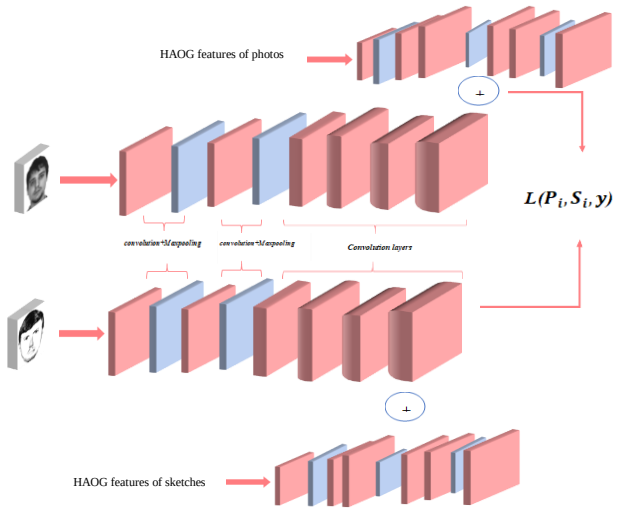

FIGURE 1. *The architecture of the proposed Siamese network*

spare autoencoder network, the features of the encoding layer are extracted as the compress features and these compressed features are fused with the features extracted from each channel of Siamese network. The fused features will be used for training through the contrastive loss function.

3.2. Loss functions

The loss function is to calculate the distance between a face photo and a sketch when learning a mapping which projects different modalities features into a common space. The aim is to separate the inter-modal sample and the intra-modal samples using a constraint condition. Three loss functions are used to train the network, respectively, for comparing the performance of the loss functions.

The contrastive loss function is used to compare the similarity between a pair of images, which is defined as formula (1):

$$L_D(W, X_1, X_2) = \frac{1}{2}(\max\{0, m - D_w\})^2 \quad (1)$$

In this function, X_1 and X_2 are the output images' features from two CNN channels. Y is the binary label which is used to represent the similarity for each image pair. If the label is 0, it shows X_1 and X_2 are the same person. If the label is 1, X_1 and X_2 are considered as two different people. After training using the contrastive loss function, the range of the output distance is too large using chi-square. Before comparing the similarity for an image pair, the output needs to be normalized to a value between 0 and 1. Then the normalized distance is used to compare with a margin value to determine the similarity for each pair. If the distance of an image pair is within the margin value, the image pair is from the same person. The dissimilar image pair is without margin value. This loss function can be used to increase the distance between the different people and to decrease the distance between images originated from the same person.

The Siamese network is trained for a two-class classification task using the Hing loss function to optimize this network. The loss function is described as formula (2):

$$\min \frac{\lambda}{2}||\omega||_2 + \sum_{i=1}^{N} \max(0, 1 - y_i o_i^{net}) \quad (2)$$

λ denotes the weight decay, ω is the weight of network. o_i^{net} is the output feature for the i[th] sample, y_i is the corresponding label for each input image pair. The hinge loss function is used as a part of the loss function which makes the distance close to a probability value. However, the Hinge loss is not differentiable at zero. The Squared l2-morn regularization, which is differentiable at zero, is used to alternate the Hinge loss, in spite of the existing sparsity. The cross-entropy loss function shows good performance for optimized unbalanced samples in two-class classification. The output features from the two CNNs are combined together using a fully connected layer with a single output. However, the output dimension is too high. Dimensionality reduction is used to keep the feature size. The sigmoid is used as an activate function, which maps the output features of the fully connected layer into the common space and measures the probability that two image feature vectors resulting from the last layer are similar. In order to reduce the number of parameters, the output size of the fully connected layer that involves sigmoid as the activate function is set to 4096.

The main idea of the Siamese network [12] is to separate samples of different classes sufficiently based on a threshold on Euclidean distance. The threshold value which is set as 0.5 is defined by a number, which is used to divide the dataset into positive pairs and negative pairs. However, for face sketch datasets, due to some reasons such as high dimensionality of feature vectors, there is no suitable value to ensure the similarity for each of the image pairs. The nearest neighbor (NN) algorithm is used to look for the shortest distance between different modalities' images.

4. Implementation and Experimental Results

Before generating the image pairs, a facial landmark detector needs to be used for face alignment. This facial landmark detector consists of a dlib library which is a model trained using iBUG 300-W dataset to determine the location of face images after extracting the face shape and 68 specific points for each face image. The locations of the eyes are fixed at (100, 50) and (100, 100) after being translated, rotated and scaled. After using the facial landmark detector, all images

are resized to 200*150. Two resized images, which each consists of different modalities' images, are concatenated and input into the network as a single image. Each face photo was paired with all sketches to generate image pairs. If the photo and the sketch show the same subjects, the label is 0 as a positive pair. Otherwise, the label is 1 as negative pairs. Due to the number of positive pairs is far lower than the number of negative pairs. The data will be separated based on the subject of the input photo images, the percentages of the randomly sampled training data and test data are 80% and 20%, respectively. If a pair of images gets separated, the image pair is used as a training instance. In contrast, it is used as a test instance. Also, all instances are normalized into [0, 1] to reduce the sensitivity and increase convergence speed.

The model is trained using the Adam optimizer which adds bias-correction and momentum with the learning rate of 0.000006. Adam optimizer shows the best performance than the stochastic gradient descent and RMSProp optimizers. The weights are initialized randomly and mini-batch is set as 125 for training. Gradient clip is appended in our model to avoid gradient explosion effectively. Several experiments certify the gradient clip is set at 1.0. The other hypermeters keep default values, such as the exponential decay rate and epsilon.

TABLE 1. The structure of the Siamese network

Convolution 1	Filter sizes: 3*3	Kernel size: 3*3	Padding: same
Down sampling	Pooling size: 3*3		Padding: same
Convolution 2	Filter sizes: 7*7	Kernel size: 7*7	Padding: same
Down sampling	Pooling size: 3*3		Padding: same
Convolution 3	Filter sizes: 3*3	Kernel size: 3*3	Padding: same
Convolution 4	Filter sizes: 7*7	Kernel size: 7*7	Padding: same
Convolution 5	Filter sizes: 7*7	Kernel size: 7*7	Padding: same

TABLE 2. The structure of each spare autoencoder

Convolution1	Filter sizes: 3*3	Kernel size: 7*7	Activate function: Relu	Padding: valid
Down sampling	Pooling size: 7*7			Padding: valid
Convolution 2	Filter sizes: 3*3	Kernel size: 5*5	Activate function: Relu	Padding: valid
Convolution 3	Filter sizes: 3*3	Kernel size: 3*3	Activate function: Relu	Padding: valid
Encoding layer	Pooling size: 7*7			Padding: same
Convolution 4	Filter sizes: 3*3	Kernel size: 3*3	Activate function: Relu	Padding: valid
Convolution layer5	Filter sizes: 7*7	Kernel size: 5*5	Activate function: Relu	Padding: valid
Upsampling	Pooling size: 7*7			Padding: same
Decoding Layer	Filter sizes: 7*7	Kernel size: 7*7	Activate function: Sigmode	Padding: same

The performance is evaluated on the composite face sketch datasets, such as e-PRIP, PRIP-VSGC and Uom-SGFS datasets [13]. These datasets include 123 photos for the AR dataset and the corresponding composite sketch using FACES and Indntikit software, respectively. The Uom-SGFS dataset [9] contains 1200 images from the Color FERET datasets and the corresponding viewed software-generated composite sketches. There are two parts in the Uom-SGFS dataset, one is created using the EFIT-V software, and the other one adopts image editing software to make the sketch image closer to the corresponding face photo.

We trained models using the three loss functions, respectively, and compare the results shown in Table 3 and Fig. 2 on different composite face sketch datasets. In particular, while the hinge, cross-entropy loss, and contrastive loss functions are used in our model, Table 3 indicates that the accuracy of contrastive loss with NN classification in Rank-10 is higher than 70% for most of the datasets and that using the improved contrastive loss function obtains the best performance, comparing with using the other loss functions.

TABLE 3. Recognition accuracy by different loss functions in Rank-10

dataset	Hing loss with NN	Cross-entropy loss	Improved contrastive loss
e-PRIP(FACES)	50.75%	78.46%	85.33%
PRIP-VSGC (Indntikit)	62.67%	52.00%	78.67%
Uom-SGFS(A)	46.39%	44.2%	64.15%
Uom-SGFS(B)	58.04%	50.25%	81.74%

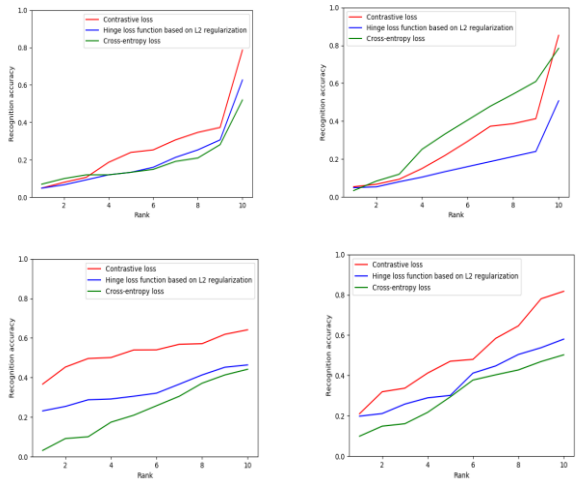

FIGURE 2. Recognition *accuracy of proposed method using three loss function (A) e-PRIP(FACES), (B) PRIP-VSGC(Indntikit), (C) Uom-SGFS(A) and (D) Uom-SGFS(B) from Rank-1 to Rank-10*

We compared the performance with the ones in [10], [14], [9] for each dataset. For Uom-SGFS datasets, since the attribute for each face sketch is very different from the corresponding face photo, the recognition accuracy is lower

than the state of art one.

TABLE 4. Recognition accuracy for e-PRIP datasets in Rank-10

Dataset	Proposed method	[10]	[14]	[9]
e-PRIP(Indntikit)	85.33%	72.6%	52.0%	54.9%

TABLE 5. Recognition accuracy for Uom-SGFS datasets in Rank-10

Dataset	Proposed method	[9]
Uom-SGFS(A)	64.15%	66.13%
Uom-SGFS(B)	81.74%	82.67%

TABLE 6. Performance with [9] and [14] for PRIP-VSGC in Rank-10

Dataset	Proposed method	[9]	[14]
PRIP-VSGC(Indntikit)	78.67%	80.8%	60.2%

5. Conclusion

In this work, we have designed a cross-mode Siamese network to match different modality images. Different from the single modal network and the basic Siamese network, the input of the cross-mode attitude metric is from the two modalities' sample, i.e., one is a photo and the other one is a sketch. The improved loss function eliminates the modal interference in the sample and maps the distance metric for features in different modes to increase the accuracy rate. Moreover, the data augmentation and the regulation methods are used to increase the size of the dataset and reduce both the risk of overfitting and the complexity of the model. The proposed method shows the best performance on different component face sketch datasets using the contrastive loss function from the experiments. The experimental results show that the accuracy rate on most datasets is higher than 70% in Rank-10 using the proposed method. Although the recognition accuracy is higher, the use of the contrastive loss function does not lead to improved classification performance for some component face sketch datasets. In the next step, we will focus on extracting spatial information from images to improve recognition accuracy using a neural network.

References

[1] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, 2004.

[2] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 593–600.

[3] B. 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, 2011.

[4] H. K. Galoogahi and T. Sim, "Inter-modality face sketch recognition," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, 2012, pp. 224–229.

[5] B. Klare and A. K. Jain, "Sketch-to-photo matching: a feature-based approach," in *SPIE Defense, Security, and Sensing*, 2010, pp. 766702–766702.

[6] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. van Gerven, "Convolutional sketch inversion," in *European Conference on Computer Vision*, 2016, pp. 810–824.

[7] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling Deep Image Synthesis with Sketch and Color," *ArXiv Prepr. ArXiv161200835*, 2016.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] C. Galea and R. A. Farrugia, "Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 6, pp. 1421–1431, 2018.

[10] H. Kazemi, S. Soleymani, A. Dabouei, M. Iranmanesh, and N. M. Nasrabadi, "Attribute-Centered Loss for Soft-Biometrics Guided Face Sketch-Photo Recognition," *ArXiv Prepr. ArXiv180403082*, 2018.

[11] A. Bas, W. A. Smith, T. Bolkart, and S. Wuhrer, "Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences," in *Asian Conference on Computer Vision*, 2016, pp. 377–391.

[12] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, 2006, vol. 2, pp. 1735–1742.

[13] 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 Secur.*, vol. 8, no. 1, pp. 191–204, 2013.

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