Composite Sketch Recognition via Deep Network - A Transfer Learning Approach

Paritosh Mittal, Mayank Vatsa, and Richa Singh
IIIT-Delhi
paritosh10059@iitd.ac.in, mayank@iiitd.ac.in, rsingh@iiiitd.ac.in

Abstract

Sketch recognition is one of the integral components used by law enforcement agencies in solving crime. In recent past, software generated composite sketches are being preferred as they are more consistent and faster to construct than hand drawn sketches. Matching these composite sketches to face photographs is a complex task because the composite sketches are drawn based on the witness description and lack minute details which are present in photographs. This paper presents a novel algorithm for matching composite sketches with photographs using transfer learning with deep learning representation. In the proposed algorithm, first the deep learning architecture based facial representation is learned using large face database of photos and then the representation is updated using small problem-specific training database. Experiments are performed on the extended PRIP database and it is observed that the proposed algorithm outperforms recently proposed approach and a commercial face recognition system.

1. Introduction

With the advent in technology, face recognition algorithms [7], [12] are utilized in several applications in e-governance such as nation-wide identification programs and welfare programs as well as law enforcement applications such as border security and forensics. Among several interesting forensic applications, *sketch recognition* helps in crime scene investigation where, facial sketch of the suspect is used as an evidence available for solving the case. In this problem domain, a query (or probe) sketch image is compared against a database of face photographs. While photographs are rich in texture and facial features, sketch images lack the texture details and only provide outline of major facial regions and some remarkable/notable features such as scar and mole. In literature [11], sketches are classified into three categories: viewed, semi viewed, and forensics categories.

- **Viewed sketches** are those which are drawn by the artist while looking at the corresponding photograph for the entire duration.
- **Semi viewed sketches** are those which are created based on the memory of the artist and not by any witness. While this is also used in academic research, compared to viewed sketches, this includes the *memory* component and considered closer to real world challenges.
- **Forensic sketches** are drawn by the artist based on the description provided by an eyewitness. These are real world cases and mostly available through law enforcement agencies (and therefore, very few image sets are available for research purposes). Recognizing forensic sketches is most challenging problem because the eyewitness may have seen a face for a very small duration, generally under stress situations, and sketch formation is dependent on the description given by the eyewitness and expertise of the sketch artist. Few examples of real world sketches are shown in Figure 1.

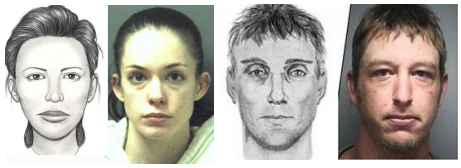

Figure 1. Real world example of sketches and corresponding photographs [1], [2].

Depending on how the sketches are formed, they can be either hand-drawn (artist based) or software generated composites. Hand drawn sketches have been employed in crime scene investigation; however, as technology has advanced, law enforcement agencies have shifted towards software generated sketches. Composite sketches have several recent success stories and therefore, law enforcement agencies have started preferring these tools over hand-drawn sketches. Few examples of composite sketches and associated photographs are shown in Figure 2.

Figure 2. Example of photographs and corresponding composite sketch images.

Once the sketches are generated, they have to be matched against a database of photographs for retrieving possible identities. As mentioned previously, photographs have rich facial texture whereas, (composite) sketches provide an approximate representation. Due to these heterogeneous variations, standard approaches of face recognition cannot be directly used and specific algorithms for matching sketch (both hand drawn and composite) images with photographs are required. In literature, recognizing hand-drawn sketches is relatively explored research area compared to composite sketch recognition. Some notable research directions are: SIFT and MLBP based local feature-based discriminant analysis (LFDA) [13] and genetic optimization based Multiscale Circular Weber Local Descriptor (MCWLD) [8]. Zhang et al. [25] have analyzed the performance of both humans and an automatic approach on forensic sketch database in a recognition experiment. The analysis suggests that humans are better in encoding minute notable features whereas algorithm is better with the sketches that contain less distinctive features. On the other hand, only recently (in 2013), research on automatic composite sketch recognition is initiated by Han et al. [11]. In this model 65 key points of the face are detected, split into 5 key regions, and each part is independently encoded using multi-scale local binary pattern. The corresponding components are matched and then fused to obtain the matching results. Chugh et al. [9] proposed image moments based algorithm to match composites with photographs with large age variations. Mittal et al. [17] proposed an algorithm which used Daisy descriptor and Gentleboost classifier. Circular patches are extracted at key facial regions which is followed by computing Daisy descriptor [22] on each patch. The extracted features are then combined using boosting approach. Mittal et al. [16] recently extended their approach by utilizing a local multiresolution self similarity descriptor [20] based bag of word model learned on CMU Multi-PIE dataset [10]. Klum et al. [14] extended their research and proposed a scalable-operational system using holistic and component based approach for forensic composites.

Existing research in composite sketch recognition do not leverage the abundant knowledge available from matching photo to photo scenario which may help improving the performance of sketch to photo matching. In this research, we propose to utilize *transfer learning* approach on the *feature representation* learned using large photo to photo face matching database for composite sketch recognition problem. As shown in Figure 3, first the deep learning based feature representation [6] is obtained from a large photo face database, followed by updating the representation via transfer learning [19] for addressing domain/problem specific challenges. The experiments are performed on the Extended PRIP (e-PRIP) composite database [11], [16] and results show that the proposed approach is able to improve the performance of composite sketch to photo face recognition.

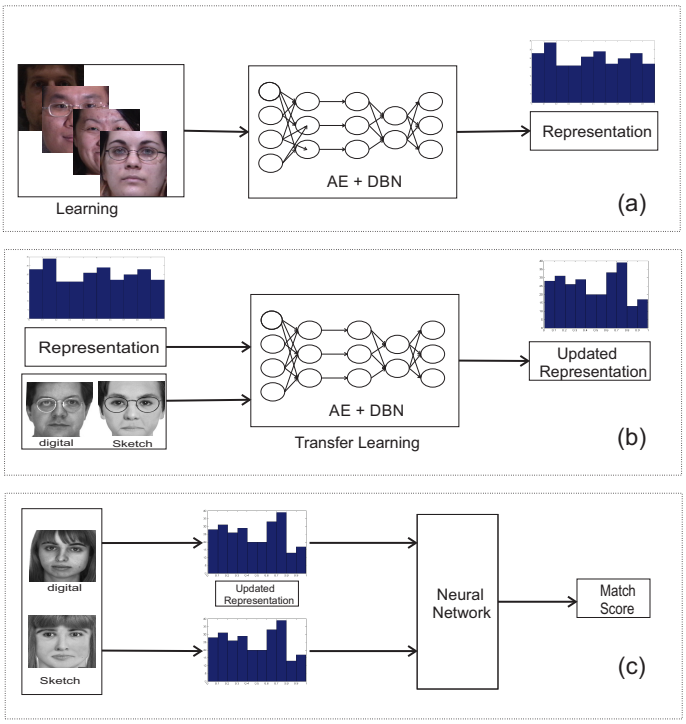

Figure 3. Steps involved in the proposed composite sketch matching recognition algorithm: (a) training deep learning based representation on face photographs, (b) updating the face representation using transfer learning approach to accommodate composite to photo matching variations, and (c) testing on composite to photographs matching using updated representation.

2. Proposed Algorithm

The proposed algorithm is divided into three stages: preprocessing, feature extraction, and feature matching. Detailed explanation of each stage is provided in the following subsections.

2.1. Preprocessing

Both, composites and photos vary in properties, e.g. varying resolution, color variations in photographs (primarily due to presence of skin textures), and composite

sketches, generally contain only outline of key facial regions. Colored images are converted into gray scale and geometric normalization (including size normalization) is performed using eye and mouth coordinates extracted by Viola-Jones face detector [24]. Some samples of preprocessed images are shown in Figure 2.

2.2. Feature Extraction and Matching

A deep learning based approach is employed to learn the feature representation. Stacked autoencoder and deep belief networks both individually learn the important characteristics from the data. Stacked autoencoders are robust to noise and perform dimensionality reduction while the Deep Belief Network (DBN) learns the representation. In this research, we use stacked autoencoder and DBN jointly to learn the representation. Face images (photographs) from the CMU multi-PIE database are used to learn the general feature representation of faces. This is followed by fine tuning via transfer learning for composite sketch and photograph face representation.

2.2.1 Autoencoder and Deep Belief Network

Autoencoder is a simple network which learns the mapping between input layer and output layer [23]. In particular, a function f learns the mapping of the input data x to hidden representation z,

$$z = f(x) = s(Wx + b) \tag{1}$$

where, s is the sigmoid function while (W, b) are the weight parameters. The hidden representation can map the learned representation back to the original data using a function g.

$$\hat{x} = g(z) = s(\hat{W}z + \hat{b}) \tag{2}$$

where, $\hat{W}$ and $\hat{b}$ are the weight parameters and $\hat{x}$ is an approximate reconstruction of the input data x such that the reconstruction error between them is minimized.

$$\underset{W\,,\,\hat{W}}{\operatorname{argmin}} ||x - \hat{x}||_F^2 \tag{3}$$

Multiple layers of autoencoder are stacked together with a sparsity promoting term to create a stacked sparse autoencoder. Different backpropagation variants such as conjugate gradient method [18] and steepest descent [5] can be employed to improve the learned representation.

A deep belief network is formed by training and stacking multiple Restricted Boltzmann Machines (RBM) [21]. DBN is formed by training each layer of RBM and stacking them together in a directed graphical manner. Since the DBN consists of network layers, a fast, greedy, and iterative layer-by-layer unsupervised training method is employed for learning feature representation.

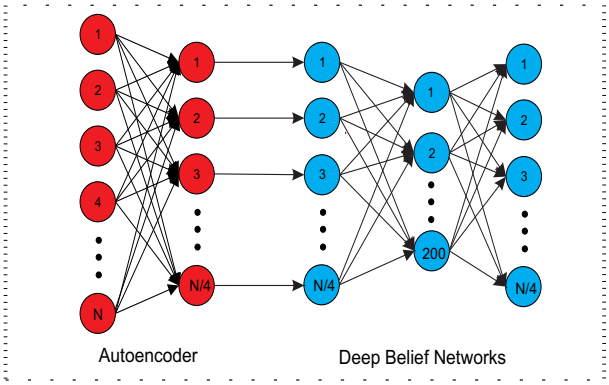

Figure 4. Steps involved in the feature extraction algorithm.

2.2.2 Autoencoders and DBN for Composite Recognition

In the proposed algorithm, autoencoder and DBN are used to obtained dimensionality reduced representation that can be used for recognition. First, this network is trained on a database of face photographs. Let the database used to learn the face representation (features) be represented as T. An image I of size $M \times N$ is converted into vector form $1 \times MN$ and provided as input to the network. After random weight initialization, a layer by layer greedy approach is used to train the autoencoder and DBN, and a the output is a 256 length feature vector. Once the network is trained, a small *transfer set* of composite-photograph pairs, S, is used to fine tune the parameters of the network. This fine tuning step acts as *inductive transfer* where the original representation is learnt on photographs and using set S, the representation is updated for composite-to-photo face matching application.

Once the feature representation, F, is learned, a neural network classifier is used for matching. Let F_s and F_d be the feature representation pertaining to composite and photograph respectively. This feature pair is concatenated as a single joint feature $[F_s F_d]$. Using labeled training data, this joint feature is used to learn a neural network classifier. Since there is a scarcity of training data (composite-photo pairs), we first learn the classifier in verification mode and the network is then undecimated to obtain the match scores which are used to generate the rank list for identification (i.e. we perform identification in verification mode).

2.3. Implementation Details

All the images are converted into grayscale and the size of normalized faces is set to 32×32. Vector form of face data (1×1024) is given as input to the autoencoder. The size of both input and output layers of stacked autoencoder is 1024 while the dimension of learned representation (reduced size) is 256. Using 30,000 frontal face images from the CMU Multi-PIE database, autoencoder learns a compact

(a) (b) (c) (d) (e)

Figure 5. Sample images from the e-PRIP database [11] and [16]. (a) Face photographs, and composite sketches generated by (b) Indian user (FACES), (c) Caucasian user (FACES), (d) Asian user (FACES), and (e) Asian user (Identi-Kit).

representation of photo faces. This compact representation of size 256 is then used as input to the DBN which further learns the feature representation. The size of DBN is [256 200 256], where the size of input visible layer is 256 and hidden layers are of size 200 and 256. Greedy layer wise training is performed to train this network. Once the training on face photographs are complete, a small set of composites and corresponding photographs are used to retrain the network (in inductive transfer learning fashion). After feature representation learning and update, a 3-layer neural network classifier is trained by concatenating the features of training composites and photographs. The output of the network is *undecimated* (or unthresholded) so that the match score is obtained that helps in generating the rank list in identification scenario.

3. Experiment Evaluation

The performance of the proposed algorithm is evaluated on the e-PRIP dataset which combines the original PRIP dataset created by Han et al. [11] and extended by Mittal et al. [16]. This database is the only available database for composite sketches. It has 123 composites and photographs from the AR dataset [15]. The dataset has four sets of composites created by different users. One set is created by an American artist using FACES software [3], two sets of databases are created by Asian artist using both FACES [3] and Identi-Kit [4] tools, and one set is created by an Indian artist using FACES software (images created by the Indian artist contributes in the extended part of the database [16]). Figure 5 shows example images from the dataset. Further, 30,000 images from the CMU Multi-PIE dataset [10] are used for training the deep network. Experimental protocol for this paper is same as given by Mittal et al. [16]. The dataset is divided into training (48 subjects) and testing (75 subjects) and same partitions are used for training-testing (along with five fold random cross validation). The results of the proposed algorithm is compared with state-of-the-art algorithm and commercial-off-the-shelf (COTS) software, FaceVACS. The results are reported in terms of average identification accuracies along with the Cumulative Match Characteristics (CMC) curves. Two set of experiments are conducted to test the proposed algorithm:

Baseline experiment: In this experiment, gallery size is equal to probe size of 75. The results of the experiment are shown in Figure 6 and Table 1. Some of the key observations are:

- As shown in Figure 6, at rank-10, the proposed algorithm outperforms the existing algorithm [16] and commercial system by at least 4% for all four subsets. Highest rank-10 accuracy of 60.2% is observed for the subset created by the Indian user followed by the Caucasian and Asian users.

- *Effect of Transfer Learning:* The proposed algorithm is evaluated without using transfer learning to compare the effect of inductive transfer. For this task, we remove the sketch samples and train the entire algorithm using only photo samples. Table 1 shows that without using sketch samples, we obtain around 20% lower rank-10 identification accuracy compared to when training involves sketch samples. Transferring knowledge from photo domain to sketch domain helps to yield improved results.

- *Effect of DBN and Autoencoder:* Figure 6 shows that using a single deep learning technique i.e. DBN or autoencoder yields less accuracies compared to using them together. Between DBN and autoencoder, in general, DBN performs better than autoencoder (except in the case of identikit).

- *Effect of Artist:* Difference in identification accuracies on all four datasets shows that, unlike the original hypothesis that composite tools mitigate the effect of artist variations, there is *artist effect* in composite sketches as well. However, it is a small database to make any concrete inferences and more research and a larger database is required.

- *Effect of Classifier:* Across all the experiments, as shown in Table 1, neural network classifier outperforms support vector machine classifier (with best performing combinations of kernel and SVM parameters).

Extended Experiment: In this experiment, gallery size is extended up to 2400 subjects while the probe size is 75. The training data is same as in the baseline experiment. The gallery is extended by combining frontal images from multiple face databases. As shown in Figure 7, at rank-40, the best performance of 58.8% is provided by the Caucasian dataset and is closely followed by the Indian dataset. Figure

Table 1. Rank-10 identification accuracy (%) on the e-PRIP composite sketch database.

Algorithm	Faces (Am)	Faces (In)	Faces (As)	IdentiKit (As)
COTS	11.3 ± 2.1	9.1 ± 1.9	7.2 ± 2.2	8.1 ± 2.1
Autoencoder + SVM	23.1 ± 1.8	24.8 ± 1.6	16.0 ± 2.2	18.5 ± 2.0
Autoencoder + NN	46.6 ± 1.7	47.7 ± 1.3	41.8 ± 1.9	46.5 ± 1.3
DBN+ SVM	25.3 ± 2.1	26.6 ± 1.6	24.0 ± 1.3	21.7 ± 0.9
DBN + NN	49.3 ± 2.6	50.7 ± 2.9	43.3 ± 2.1	45.3 ± 2.1
Autoencoder+DBN+SVM	38.7 ± 1.6	40.1 ± 1.7	32.0 ± 2.2	31.6 ± 2.4
Mittal et al. [16]	51.9 ± 1.2	53.3 ± 1.4	42.6 ± 1.2	45.3 ± 1.5
Proposed without transfer learning	36.0 ± 2.9	37.3 ± 3.6	26.6 ± 2.5	32.7 ± 2.7
Proposed	$\mathbf{56.0 \pm 2.1}$	$\mathbf{60.2 \pm 2.9}$	$\mathbf{48.1 \pm 1.7}$	$\mathbf{52.0 \pm 2.4}$

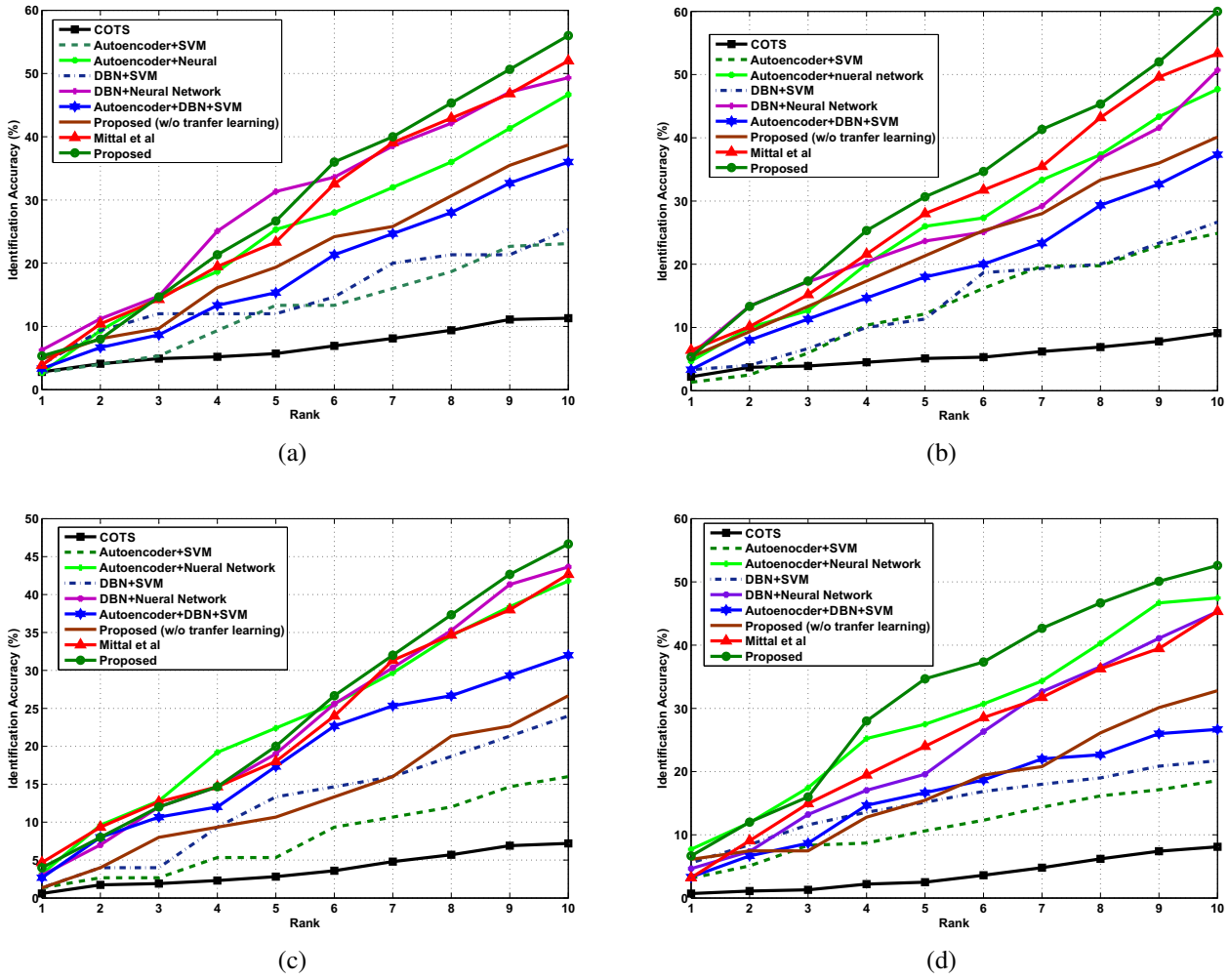

Figure 6. CMC curves of the existing and proposed approaches on the composite sketch database (a) Faces (Caucasian user), (b) Faces (Indian user), (c) Faces (Asian user), and (d) IdentiKit (Asian user).

8 shows an example where the composite sketches created by different artists yield the matching at different ranks by the proposed algorithm. The difference in accuracies obtained using different users (artists) shows that composite sketch are also dependent on the artist; however, due to limited amount of data we cannot compare artist dependency on composite sketch matching.

4. Conclusion

Composite sketch recognition is an important law enforcement application in which a sketch of a suspect is matched against a gallery of known subjects. This paper

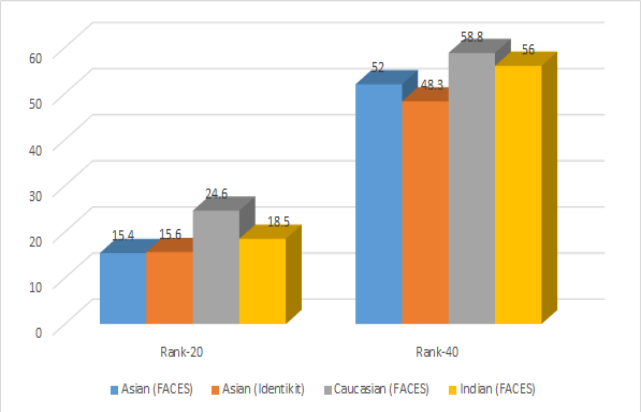

Figure 7. Identification accuracies (%) on the extended gallery experiment of size 2400.

Figure 8. Results with variations in sketch artists. The first column represents the composite probe generated by different artists and other columns show the top matches obtained from the proposed algorithm.

presents a novel algorithm for matching composite sketches with face photographs. The proposed algorithm performs inductive transfer on the features learned using a deep learning architecture to effectively match the heterogenous information. The results of the proposed algorithm on the e-PRIP dataset show improved results compared to existing algorithms. Experiments on extended gallery of 2400 subjects also show that the proposed algorithm is scalable and achieves rank 40 accuracy of 58%.

References

[1] http://www.forensicmag.com/articles/2013/07/click-capture-making-case-digital-composite-images.

[2] http://apps.washingtonpost.com/g/page/local/before-and-after-police-composite-sketches/686/.

[3] Faces 4.0, iq biometrix. http://www.iqbiometrix.com.

[4] Identi-kit, identi-kit solutions. http://www.identikit.net/.

[5] R. Battiti. First-and second-order methods for learning: between steepest descent and newton's method. *Neural computation*, 4(2):141–166, 1992.

[6] Y. Bengio. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[7] H. Bhatt, R. Singh, and M. Vatsa. Covariates of face recognition. Technical report, IIIT Delhi, 2015.

[8] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Memetically Optimized MCWLD for Matching Sketches With Digital Face Images. *IEEE TIFS*, 7(5):1522–1535, 2012.

[9] T. Chugh, H. S. Bhatt, R. Singh, and M. Vatsa. Matching age separated composite sketches and digital face images. In *IEEE BTAS*, pages 1–6, 2013.

[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010.

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

[12] A. K. Jain and S. Z. Li. *Handbook of face recognition*. Springer, 2005.

[13] B. F. Klare, L. Zhifeng, and A. K. Jain. Matching forensic sketches to mug shot photos. *IEEE TPAMI*, 33(3):639–646, 2011.

[14] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain. The FaceSketchID system: Matching facial composites to mugshots. In *IEEE TIFS*, volume 9, pages 2248–2263, 2014.

[15] A. Martinez and R. Benavente. The AR face database. Technical report, 1998. Computer Vision Center.

[16] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa. Recognizing composite sketches with digital face images via ssd dictionary. In *IEEE/IAPR IJCB*, 2014.

[17] P. Mittal, A. Jain, R. Singh, and M. Vatsa. Boosting local descriptors for matching composite and digital face images. In *IEEE ICIP*, pages 2797–2801, 2013.

[18] J. Nocedal and S. J. Wright. *Conjugate gradient methods*. Springer, 2006.

[19] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on KDE*, 22(10):1345–1359, 2010.

[20] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE CVPR*, pages 1–8, 2007.

[21] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *ACM ICML*, pages 1064–1071, 2008.

[22] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE TPAMI*, 32(5):815– 830, 2010.

[23] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ACM ICML*, pages 1096–1103, 2008.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, volume 1, pages I–511, 2001.

[25] Y. Zhang, C. McCullough, J. R. Sullins, and C. R. Ross. Hand-drawn face sketch recognition by humans and a PCA-based algorithm for forensic applications. *IEEE TSMC - A*, 40(3):475–485, 2010.