Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning

Christian Galea, *Student Member, IEEE*, and Reuben A. Farrugia, *Member, IEEE*

***Abstract*— Sketches obtained from eyewitness descriptions of criminals have proven to be useful in apprehending criminals, particularly when there is a lack of evidence. Automated methods to identify subjects depicted in sketches have been proposed in the literature, but their performance is still unsatisfactory when using software-generated sketches and when tested using extensive galleries with a large amount of subjects. Despite the success of deep learning in several applications including face recognition, little work has been done in applying it for face photograph-sketch recognition. This is mainly a consequence of the need to ensure robust training of deep networks by using a large number of images, yet limited quantities are publicly available. Moreover, most algorithms have not been designed to operate on software-generated face composite sketches which are used by numerous law enforcement agencies worldwide. This paper aims to tackle these issues with the following contributions: 1) a very deep convolutional neural network is utilised to determine the identity of a subject in a composite sketch by comparing it to face photographs and is trained by applying transfer learning to a state-of-the-art model pretrained for face photograph recognition; 2) a 3-D morphable model is used to synthesise both photographs and sketches to augment the available training data, an approach that is shown to significantly aid performance; and 3) the UoM-SGFS database is extended to contain twice the number of subjects, now having 1200 sketches of 600 subjects. An extensive evaluation of popular and state-of-the-art algorithms is also performed due to the lack of such information in the literature, where it is demonstrated that the proposed approach comprehensively outperforms state-of-the-art methods on all publicly available composite sketch datasets.**

***Index Terms*— Deep learning, convolutional neural network, software-generated composite sketches, face photos, morphological model, augmentation, database.**

I. Introduction

HETEROGENEOUS Face Recognition (HFR) concerns the matching between two face images belonging in different modalities, one of which is typically a traditional visible band (VIS) face photo image. One of the most difficult HFR scenarios involves the matching of VIS images to sketches obtained from eyewitness descriptions of criminals, since they contain a large modality gap[1] and typically also exhibit several deformations and distortions owing to factors such as eyewitness memory loss and difficulty in describing the face [1], [2]. In fact, even leading Commercial Off-the-Shelf (COTS) Face Recognition Systems (FRSs) have been shown to perform poorly when matching sketches with photos [1]–[5].

There exist two types of sketches, namely *hand-drawn* sketches which are drawn by forensic artists, and *software-generated* sketches that are created with the aid of computer software programs such as IdentiKit [6] and EFIT-V [7]. Most law enforcement agencies are now using software-generated sketches, primarily due to their lower cost [4], [8]. Examples of sketches and the corresponding photos are shown in Figure 1.

Algorithms designed for face photo-sketch recognition can be broadly categorised into two groups [9], [10]. The first is *intra-modality* algorithms, which attempt to reduce the modality gap by transforming a photo (sketch) to a sketch (photo) and then comparing the resultant images with the original probe sketches (gallery photos) using a face recogniser designed to operate in the target modality. However, such methods have only proven to be effective when the sketches are very similar in appearance to the original photographs, and are essentially learning a texture mapping. Moreover, they tend to be complex and computationally expensive [4], [9] and their performance is not as good as *inter-modality* algorithms on more realistic sketches [5], [11], [12]. The performance of the chosen face recogniser also depends on the quality of reconstructed images, which often contain undesirable artefacts. As a result, most recent efforts have focused on the design of inter-modality methods, which learn and/or extract features or classifiers that maximise inter-class separability while minimising intra-class differences [4], [5], [11]–[13].

The majority of inter-modality algorithms use hand-crafted features such as the Scale-Invariant Feature Transform (SIFT) and Multiscale Local Binary Pattern (MLBP), and have shown promising performance [2], [5], [12]. However, it is unlikely that such features are optimal since they were not designed for inter-modality face recognition [13], and it would therefore be desirable to design and use potentially superior feature

Manuscript received May 10, 2017; revised September 18, 2017; accepted December 11, 2017. Date of publication December 29, 2017; date of current version February 7, 2018. This work was supported in part by the Malta Government Scholarship Scheme and in part by NVIDIA Corporation through the donation of a Titan X Pascal GPU. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Karthik Nandakumar. *(Corresponding author: Christian Galea.)*

The authors are with the Department of Communications and Computer Engineering, University of Malta, MSD2080 Msida, Malta (e-mail: christian.galea.09@um.edu.mt; reuben.farrugia@um.edu.mt).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2017.2788002

[1] Modality gap is due to difference in appearance, since photos depict the real-life environment whereas sketches are hand-drawn or computer-generated.

Fig. 1. Photos of three subjects in the Color FERET database [18] and the corresponding sketches from the two sets of the UoM-SGFS database [5].

descriptors that are better adapted for the task of face photo-sketch recognition. This can be performed with the aid of *deep learning*, which has become a hot research topic owing to its great success in several application domains including traditional VIS-VIS face recognition and image super-resolution [14]–[17]. However, there has been limited work in using deep learning for face photo-sketch recognition. This paper aims to rectify this situation with the following primary contributions:

- It is shown that a face recognition system based on a *very deep* convolutional neural network that achieved state-of-the-art performance on popular face recognition databases still yields poor recognition rates when tasked with matching sketches with photographs.
- While transfer learning is beneficial, it is shown that performance still lags behind leading methods since traditional data augmentation techniques are insufficient to avoid the problem of having only one sketch per subject.
- A 3D morphable model is used to create a set of synthetic photos *and* synthetic sketches to artificially expand the number of images per subject. Applying transfer learning to a deep network using these images yields a framework that outperforms state-of-the-art algorithms.
- The UoM-SGFS database [5] has been expanded with twice as many subjects and is used to perform an extensive evaluation of leading algorithms, which is largely unavailable in current literature.

The rest of this paper is organised as follows: an overview of related work is provided in Section II, followed by a description of the proposed methods in Section III. The methodology used for evaluation is outlined Section IV and results reported in Section V. Directions for future work and concluding remarks are finally given in Section VI.

II. Related Work

A summary of salient intra- and inter-modality face photo-sketch recognition algorithms proposed in the literature will first be provided, followed by an overview of deep learning-based face recognition models that have been implemented.

A. Face Photo-Sketch Synthesis and Recognition Algorithms

Several approaches proposed in literature focus on intra-modality algorithms, also known as Face Hallucination (FH) techniques [10]. Some of the best-performing and most popular methods include Eigen-transformation (ET) that synthesises whole faces using a linear combination of photos (or sketches) under the assumption that face photos and the corresponding sketches are reasonably similar in appearance, the Eigen-patches (EP) extension in [11] to perform synthesis at a local level, the use of the Locally Linear Embedding (LLE) manifold learning technique in [19] to create a patch using a linear combination of neighbouring patches, the Multiscale Markov Random Fields (MRF) approach [20] which models the relationships among patches, its extension in [21] to cater specifically for lighting and pose variations, the Markov Weighted Fields (MWF) model in [22] that uses a weighted MRF to model the relation between photo and sketch patches, and the recent Bayesian framework proposed in [23] that considers relationships among neighbouring patches for neighbour selection and weight computation models. A more thorough review of FH algorithms may be found in [10], [23], and [24].

State-of-the-art inter-modality methods include the Direct Random Subspace (D-RS) approach proposed in [25] and modified in [2] which convolves images with three filters, followed by extraction of SIFT and MLBP descriptors from overlapping patches which are compared using the cosine similarity measure. In [26], D-RS was also fused with the Component-based Representation (CBR) method [4] designed to operate on software-generated sketches by comparing MLBP features extracted from the individual facial components. The Histogram of Averaged Orientation Gradients (HAOG) method [9] assigns higher importance to areas having strong orientations and was reported to achieve very high performance on a simple dataset. Recently, the log-Gabor-MLBP-SROCC (LGMS) method was presented in [12] which uses both local and global texture descriptors in the form of MLBP and log-Gabor filters, respectively, along with the Spearman Rank-Order Correlation Coefficient (SROCC) for comparison of the subspace-projected features. LGMS was shown to generally outperform popular and state-of-the-art algorithms including HAOG and D-RS, in the case of both hand-drawn sketches [12] and software-generated sketches [5]. Further information on leading face photo-sketch recognition algorithms can be found in [4], [10]–[13], and [27] [4], [10]–[13], [27].

It should be noted that several algorithms reportedly achieve very high recognition rates, alluding that the problem of face photo-sketch recognition has been solved. However, most of these methods are tested using sketches that bear an unrealistically great resemblance to the original photo, and/or do not use an extended photo gallery to mimic the extensive

mug-shot galleries maintained by law enforcement agencies. For example, it was shown in [11] and [12] that even the HAOG algorithm which was reported to achieve a Rank-1 accuracy of 100% demonstrated significantly degraded performance when more challenging (but more realistic) sketches were used along with an extended gallery. Most algorithms have also been designed and evaluated only for hand-drawn sketches, which are now less popular than software-generated sketches [4]. Consequently, it was recently demonstrated in [5] that more work still needs to be done for this type of sketches. The work in this paper therefore focuses on the design of a system that operates on software-generated sketches. To the best of the authors' knowledge, this work is also the first to develop a system that operates on sketches generated with the EFIT-V software that is used in several countries around the world [7].

B. Deep Learning-Based Methods

One of the earliest and most popular deep-learning approaches is the AlexNet DCNN architecture [28] that was trained for the task of object classification. Several superior approaches based on deep-learning have since been introduced, along with new methods to improve the performance of a network. Of particular interest in this paper are face recognition methods such as Facebook's DeepFace [14], DeepID series [29]–[32], Google's FaceNet [15] and VGG-Face [16], which have provided important observations such as the superior performance that is generally obtained by using more layers [32], the benefit of a high amount of training data (especially for 'deeper' networks having more trainable parameters) [15], [16], the use of multiple DCNNs [29], and a "triplet-based" objective function which aims at decreasing the distance between features of the same subject and increasing the distance between features of different subjects [15], [16]. More information regarding deep-learning-based FRSs may be found in [15], [16], and [29]–[33].

To the best of the authors' knowledge, the only system designed for face photo-sketch recognition that utilises deep learning concepts was proposed in [34], where autoencoders and a deep belief network were bootstrapped to learn a feature representation of VIS face photos and were then fine-tuned for face photo-sketch recognition. However, the system is shallow and does not exploit the spatial relationships inherently present in images, which are important in facial recognition [4].

III. Proposed Method

As shown in Figure 2, the proposed framework consists of a deep CNN and a triplet embedding that optimises the features for verification, and a data augmentation approach to circumvent the lack of multiple images per subject. The framework is demonstrated to outperform leading methods when applied to one of the hardest HFR tasks, namely face photo-sketch recognition, with the resultant method thus denoted the DEEP (face) Photo-Sketch System (DEEPS). A brief overview of the UoM-SGFS database that has been extended with twice the number of subjects and sketches will first be given, followed by a description of the proposed DEEPS framework.

A. Extended UoM-SGFS Database

The UoM-SGFS database[2] presented in [5] is the largest software-generated sketch database and the only one having all of the sketches represented in full colour. To enable better training and testing of algorithms, this database has been extended to contain double the number of subjects and thus contains 1200 sketches of 600 subjects. Similar to the original database, two sets are present: *Set A* having sketches created using EFIT-V [7], and *Set B* containing the sketches in Set A that are lightly altered using an image editing program to make the sketches more realistic (emulating the process performed by law enforcement agencies in real-life).

The new sketches were compiled by the same EFIT-V operator who created the sketches in the original dataset, and was trained by a qualified forensic scientist from a local police force to ensure the adoption of real-life practices in the creation of the database. In addition, despite being created while viewing a face photo, the sketches intentionally contain several distortions and exaggerations to mimic real-world forensic composite sketches which are not publicly available to researchers. More information may be found in [5].

B. Data Augmentation

A drawback of deep learning methods is the requirement of a large amount of data for robust learning, to reduce effects such as over-fitting and to learn more effective functions [16], [28]. Extensive datasets containing not only a high number of unique classes but also numerous examples for each class have been created for tasks such as object and face recognition, and thus allow researchers to train and test their algorithms well. For example, the ImageNet database [35] contains a training set having 1.2 million images of 1000 categories. Such high numbers are possible due to the sheer availability of images on the Internet, where search engines can be used to automatically retrieve images of interest. In the case of face recognition, databases such as the one used to train the VGG-Face network [16] are typically created by assigning celebrities as subjects, many photos of whom are often captured and thus allow a database to be quite easily populated with multiple images per subject. This approach cannot be undertaken in the case of face sketches due to their limited availability as a result of privacy protection issues (in the case of real-world forensic sketches), and the time consuming nature of sketch creation (in the case of publicly available viewed sketch datasets). This means that the number of subjects represented with a sketch image is quite limited, even when combining all available databases. Moreover, sketch databases typically contain only one sketch per subject, and the face photo datasets used to construct sketch databases often contain a limited number of photos per subject as well. However, object and face databases typically contain hundreds of examples for each unique entity which exhibit several variations. In the case of face recognition, these variations span factors such as expression and pose, and allow a network to be robust to intra-class differences. Consequently, a deep network trained using just two images per subject (a sketch and a photo) would find it hard to reliably

[2]Available at: http://goo.gl/KYeQxt, http://wp.me/P6CDe8-4q

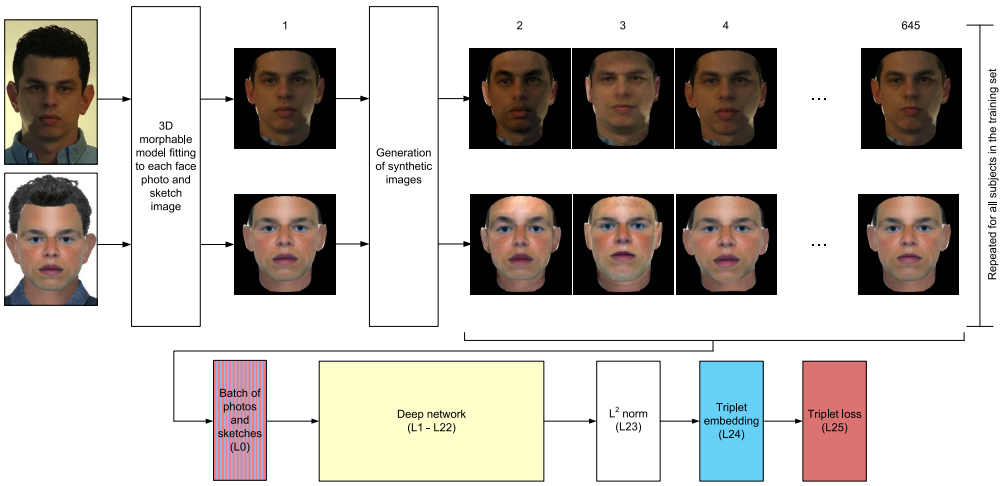

Fig. 2. The architecture of the proposed DEEPS framework, where synthetic images are created and used to train a DCNN. Shown are examples of face images represented with a 3D morphable model and the corresponding synthesised images of one subject, where the first row contains original and synthesised photos of a subject in the Color FERET database [13], and the second row contains the software-generated sketches of the subject in the UoM-SGFS Set A database [5]: '1': Original image represented with the morphable model, '2': More female and taller, '3': Older and fatter, '4': Younger, fatter and shorter, '5': Face shape random adjustment, '645': Mouth random adjustment. Detailed architecture of the deep network shown in Table I.

distinguish them from different identities and at the same time learn intra-class similarities [36]. Even methods designed for face and object recognition tasks (where large datasets are available) have found data augmentation techniques beneficial for system performance [16], [28], [36]. However, it will be shown in Section V that even traditional augmentation methods yield limited performance improvement when using one photo and sketch (the original images). To circumvent this problem, the use of a 3D face morphable model[3] [37] (along with the approach in [38] to fit the model to face images[4]) is proposed to enable the generation of synthetic face photos and sketches, with the additional benefit of normalising off-pose faces to be frontally aligned with no rotation (which is particularly useful in the case of photos). Changes to a face image include: (i) the individual facial features,[5] and more global changes in terms of (ii) age (older or younger), (iii) gender (more female or more male), (iv) height (taller or shorter) and (v) weight (fatter or thinner). Of course, there is a virtually infinite number of ways in which a face image can be altered. In this work, 644 images are created for each face image. These include five random adjustments to the four facial components individually (yielding 20 images), and 624 adjustments to the age, gender, height and weight, both individually (i.e. changing one attribute at a time) and also when multiple attributes are changed simultaneously. The original image is also used, for a total of 645 photos and 645 sketches per subject.[6] Sketches and photos are modified with identical parameters. Some examples of face photos and face sketches created with this approach are shown in Figure 2.

The proposed system thus allows any face database to be expanded with an arbitrarily large number of images. This is particularly important in the case of sketches, since there is typically only one sketch image per subject in both publicly available datasets and in real-life. The increased number of samples will be shown in Section V to be greatly beneficial in training a deep network.

C. Deep Convolutional Neural Network

Since performance of neural networks tends to increase with more layers and filters [16], it would be ideal to design a deep and wide network for face photo-sketch recognition. However, a significant amount of training data is required to counteract effects such as over-fitting as a consequence of the numerous free parameters, which also leads to long training times (on the order of weeks). To mitigate this problem, researchers often apply *transfer learning*, where a pre-trained network's parameters are fine-tuned with a training set that contains samples from the target database. The use of a pre-trained network enables faster convergence, decreases the probability of finding poor local minima, and leverages the regularisation effect that enables better generalisation [39], [40].

The work in this paper also benefits from transfer learning by using the original and synthetic images to fine-tune the VGG-Face FRS model described in [16] and shown in Table I, that was trained with 2.6M face images of 2,622 subjects [16][7] acquired from the Internet. This network was chosen for several reasons: (i) it was designed for recognition of faces as done in this work, (ii) the face photos used for training represent one of the modalities used in the task of face photo-sketch recognition, (iii) the VGG-Face network was shown to be among the leading FRSs for unconstrained face recognition. Hence, the network provides a better basis on which fine-tuning can be performed than a network trained for other tasks.

A similar implementation methodology to that used for the VGG-Face network is employed, whereby the DCNN is first

[3]Available at: http://faces.cs.unibas.ch/bfm/main.php
[4]Available at: https://github.com/waps101/3DMM_edges
[5]Facial features encompass the eyes, nose, mouth and the rest of the face
[6]The exact parameters may be found at: http://wp.me/P6CDe8-7D
[7]Available at: http://www.robots.ox.ac.uk/~vgg/data/vgg_face/

TABLE I

NETWORK ARCHITECTURE WHEN TRAINING FOR VERIFICATION WITH THE TRIPLET LOSS FUNCTION, BASED ON THE VGG-FACE NETWORK [16]. THE ARCHITECTURE USED FOR CLASSIFICATION MAY BE FOUND IN [16]. ALL 'CONV' LAYERS EXCEPT L24 USE THE RELU FUNCTION [28]

layer	L0	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
type	input	conv	conv	mpool	conv	conv	mpool	conv	conv	conv	mpool	conv	conv	conv
support	–	3	3	2	3	3	2	3	3	3	2	3	3	3
filt dim	–	3	64	–	64	128	–	128	256	256	–	256	512	512
num filts	–	64	64	–	128	128	–	256	256	256	–	512	512	512
stride	–	1	1	2	1	1	2	1	1	1	2	1	1	1
pad	–	1	1	0	1	1	0	1	1	1	0	1	1	1
data size	224	224	224	112	112	112	56	56	56	56	28	28	28	28

layer	L14	L15	L16	L17	L18	L19	L20	L21	L22	L23	L24	L25
type	mpool	conv	conv	conv	mpool	conv	dropout	conv	dropout	norm	conv	triplet loss
support	2	3	3	3	2	7	1	1	1	1	1	1
filt dim	–	512	512	512	–	512	–	4096	–	–	$D = 4096$	–
num filts	–	512	512	512	–	4096	–	4096	–	–	$L = 1024$	–
stride	2	1	1	1	2	1	1	1	1	1	1	1
pad	0	1	1	1	0	0	0	0	0	0	0	0
data size	14	14	14	14	7	1	1	1	1	1	1	1

trained for the task of classification using the softmax log-loss objective function (tuning layers L1–L21), after which it is trained for verification using the triplet-loss objective (learning L24). Stochastic gradient descent (SGD) with momentum is used to train the network in each case, using original and synthetic photo-sketch pairs of 450 subjects as elaborated in Section IV. However, the last fully-connected layer (mapping the D-dimensional feature descriptor to classes corresponding to the number of distinct identities in the training set) must be re-initialised since the VGG-Face network was trained using different subjects than the ones considered in this work.

Photos and the corresponding sketches are used for each subject in the training set, allowing the network to learn the relationship between the two modalities. In other words, the aim of the network is to learn modality-invariant parameters such that it may correctly classify both photos *and* sketches.

After the network is trained for classification, the last two layers (the last fully-connected layer and the softmax log-loss layer) are replaced with three layers: (i) a layer that normalises the output feature vector to unit length (L23), followed by (ii) a fully-connected layer (L24) consisting of D inputs and L outputs, $L \ll D$, and (iii) a triplet-loss layer (L25). Layer L24 performs dimensionality reduction and outputs a vector that is suitable for verification, which should yield vectors whose distance with respect to other vectors is small for input face images of the same subject, and large for different subjects. L25 computes the error of the objective function to determine how the parameters must be adjusted using SGD.

D. Triplet-Loss Embedding Scheme

The triplet-loss objective function has been used in several state-of-the-art systems to train the network for the final application of a FRS, namely identification via verification. Given a triplet $\{a, p, n\}$, the aim is to reduce the distance (or increase similarity) between an image a of a subject called the *anchor* and another image p of the same subject known as the *positive* example, while increasing the distance (minimising the similarity) between the anchor and an example n from a different subject called the *negative* example. Two main methods have been proposed in literature: *Triplet Distance Embedding (TDE)* based on Euclidean distance [15], [16], and *Triplet Similarity Embedding (TSE)* based on vector dot product similarities [33]. For both methods, an input face image l_i, $i \in \{a, p, n\}$ yields an output $\phi(l_i) \in \mathbb{R}^D$ that is projected to a $L \ll D$ dimensional space to obtain $\vec{x}_i \in \mathbb{R}^L$:

$$\vec{x}_i = W \frac{\phi(l_i)}{\|\phi(l_i)\|_2} \tag{1}$$

where $W \in \mathbb{R}^{L \times D}$ is the projection matrix that is learned, with L and D set to 1024 and 4096, respectively. The last fully-connected layer now corresponds to the weight matrix W, as shown in Table I. Biases and their learning rate are set to zero along with the parameters of the rest of the network, such that the training process only affects W.

Following recommendations in [15] and [16], all positive pairs (a, p) are selected and n is chosen at random among those images which violate the triplet loss margin. Triplets are chosen while the network is being trained.

IV. IMPLEMENTATION METHODOLOGY

Face photo-sketch algorithms proposed in literature are often compared to few other face photo-sketch algorithms, and COTS FRSs form the sole benchmarks in several cases, e.g. [2], [34], [41]. These are arguably unfair comparisons since COTS FRSs have not been designed for sketch recognition and also makes comparison between algorithms difficult, especially since different databases and methodologies are typically employed. In this work, the proposed DEEPS method is compared not only to a leading FRS but also to several popular and state-of-the-art face photo-sketch synthesis and recognition methods described in literature that are all evaluated under a common protocol. This also enables one of the most comprehensive evaluations of face photo-sketch recognition algorithms performed to date. The methodology will now be provided, with the results discussed in Section V.

A. Training and Testing Details for DEEPS

For the proposed approach, 20% of the images used for training are reserved as validation data for parameter tuning.

The training procedure and parameters are similar to those specified in [16], namely biases are initialised to 0 and weights are randomly sampled from a Gaussian distribution with zero mean and 10^{-2} standard deviation for the re-initialised layers. The learning rate is set to a relatively small value of 10^{-3} to limit the rate of change of the parameters and thus enable better convergence, since the parameters should not require adjustments that are too great given that they are already pre-trained. However, the learning rate of the last layer is increased ten-fold since it is re-initialised without any prior training whatsoever. The triplet loss margin is empirically set to 0.1.

The input to the network is a patch of size 224×224 that is randomly cropped from a face image and flipped with 50% probability, with the mean of the images in the training set subtracted to ascertain stability of the learning algorithm [16]. At test time, a process similar to that employed for the original VGG-Face network as described in [16] is performed, namely the dropout layers are removed and images are scaled to three sizes (256×256, 384×384 and 512×512) to enable multi-scale testing. Feature vectors are then computed for the ten 224×224 patches (the four corners, the centre, and their horizontal flips), extracted at each scale. The final descriptor is the average of the resultant 30 L-D feature vectors that are obtained for each probe (sketch) image and gallery (photo) image. As mentioned in Section III-D, the feature vector of a probe image is then compared with those of all the gallery images using either Euclidean distance [16] or vector dot products [15].

B. Evaluation Methodology

Algorithm performance is reported in terms of the rank-retrieval rates, where Rank-N denotes the number of subjects that were correctly identified in the top N matches. While it is desirable to obtain a Rank-1 rate of 100% (i.e. all subjects correctly identified as the best match), in practice it is difficult to achieve such high performance at this rank due to the significant differences between sketches and photographs. In fact, some sketches may resemble the face photos of other subjects more closely than the true match. In addition, law enforcement agencies would still manually examine the top P best matches to reduce errors, abide by any local laws, and ensure fair legal proceedings. P typically lies between 50 and 150 [4], [11], [12], [42]; therefore, an automatic face photo-sketch recognition system serves to filter the list of potential criminals to examine from the order of thousands or even millions to just a few tens or a few hundreds. The numerical values of the rank retrieval rates are provided until Rank-100, while they are depicted graphically in Cumulative Match Characteristic (CMC) curves until Rank-1000. The CMC curves depict the number of correctly identified individuals below a given rank [43], [44]. For example, a rate of Rank-50 rate of 80% indicates that 80% of subjects have been correctly identified within the top 50 matches. True Accept Rates (TARs) at False Accept Rates (FARs) of 0.1% and 1.0% are also provided, and are shown graphically in Receiver Operating Characteristics (ROC) curves. TARs at given FARs are commonly used to evaluate the performance of traditional face recognition systems, and represent the verification accuracy [43], [44]. Finally, the Equal Error Rate (EER) which corresponds to the point at which the FAR is equal to the False Reject Rate (FRR) is also provided [45]. High values of the rank-retrieval rates and TARs, and low values of the EER are desirable.

All algorithms are evaluated on five train/test set splits with the mean and standard deviation of results reported. It should be noted that all algorithms use the same train/test sets, such that the reported standard deviations are a measure of the consistency in performance of each algorithm. Lastly, only one sketch and one photo for each subject are used during testing (the original non-synthetic images).

C. Databases Used

The sketches in the extended UoM-SGFS database and the corresponding photos in the Color FERET database [18] are used to evaluate the algorithms considered, selecting 450 subjects at random for training and assigning the remaining 150 subjects to the test set. The intra-modality algorithms use the same training set as used by the face recognisers to avoid using too few subjects to train and test the latter. This process is done for each of the two sets in the UoM-SGFS database.

Photos form the gallery set while sketches populate the probe set. The gallery is further extended with the photos of 1521 subjects to simulate the mug-shot galleries maintained by law-enforcement agencies. These include 509 subjects from the MEDS-II database,[8] 476 subjects from the FRGC v2.0 database,[9] 337 subjects from the Multi-PIE database [46], and 199 subjects from the FEI database.[10]

To evaluate the performance of the deep learning-based method in [34], DEEPS is also evaluated on two additional databases: the PRIP-VSGC dataset [4], [26] and the e-PRIP dataset [34], [47]. Although both datasets contain composite sketches of the 123 subjects in the AR database [48], the sketches in the PRIP-VSGC dataset were created using IdentiKit operated by an Asian user while the sketches in the e-PRIP dataset were created using the FACES software operated by an Indian user. The same evaluation protocol described in [34] has been employed to enable direct comparison with the method proposed therein. Specifically, 48 subjects are reserved for training while the remaining 75 subjects are used for testing, while performance figures are computed over 5 train/test set splits. However, DEEPS is not re-trained with these datasets due to the limited number of subjects, and to also determine the robustness of the proposed approach on sketches that are different to those used for training. Only the Rank-10 retrieval rates are given since they are the only results reported with exact values. The extended gallery is excluded when using these two databases, as done in [34].

D. Algorithms Compared

The proposed system is compared to traditional FRSs, and both intra- and inter-modality methods as follows:

[8] Available at: http://www.nist.gov/itl/iad/ig/sd32.cfm
[9] Available at: http://www.nist.gov/itl/iad/ig/frgc.cfm
[10] Available at: http://fei.edu.br/~cet/facedatabase.html

TABLE II

MEANS AND STANDARD DEVIATIONS OVER 5 TRAIN/TEST SET-SPLITS FOR VARIATIONS OF THE PROPOSED APPROACH WHEN EVALUATED ON UOM-SGFS SET A SOFTWARE-GENERATED SKETCHES. *SYNTH* = USING SYNTHETIC IMAGES TOGETHER WITH THE ORIGINALS, *TDE* = USING *TRIPLET DISTANCE EMBEDDING* TO LEARN THE TRIPLET EMBEDDING, AND *TSE* = USING *TRIPLET SIMILARITY EMBEDDING* TO LEARN THE TRIPLET EMBEDDING AS ELABORATED IN SECTION III-D. MORE DETAILED CONFIGURATION DESCRIPTIONS ARE ALSO GIVEN IN TABLE IV

Method/Config.	Matching Rate (%) at Rank-N				TAR@FAR=0.1%	TAR@FAR=1.0%	EER (%)
	N=1	N=10	N=50	N=100			
VGG-Face	9.33±2.45	31.07±3.73	59.73±2.52	73.60±3.58	11.87±2.47	37.33±4.40	14.18±1.05
A (no triplets, synth)	22.80±2.88	49.33±3.65	73.07±2.69	82.53±1.85	29.60±5.53	54.40±3.90	9.99±1.29
B (TDE, no synth)	20.93±2.14	52.93±3.96	76.40±3.25	85.47±3.25	27.47±2.72	61.07±4.68	9.35±1.34
C (TDE, synth)	31.60±1.12	66.13±2.47	86.00±1.25	93.47±1.85	41.87±3.11	73.47±2.08	6.26±0.60
D (TSE, synth)	29.73±3.35	62.67±2.26	86.00±2.00	92.40±2.85	39.60±2.56	70.93±1.86	6.39±0.88

TABLE III

MEANS AND STANDARD DEVIATIONS OVER 5 TRAIN/TEST SET-SPLITS WHEN EVALUATED ON UOM-SGFS SET A SOFTWARE-GENERATED SKETCHES AND OMITTING INDIVIDUAL ATTRIBUTE VARIATIONS (I.E. AGE, GENDER, WEIGHT, AND HEIGHT) FOR PHOTO AND SKETCH GENERATION

Description	Matching Rate (%) at Rank-N				TAR@FAR=0.1%	TAR@FAR=1.0%	EER (%)
	N=1	N=10	N=50	N=100			
All variations (Config. C)	31.60±1.12	66.13±2.47	86.00±1.25	93.47±1.85	41.87±3.11	73.47±2.08	6.26±0.60
No age variations	26.93±2.85	60.53±3.78	85.47±2.28	91.47±1.45	35.73±4.10	69.73±3.08	6.78±0.97
No gender variations	27.07±1.86	59.47±4.31	85.20±3.38	90.67±1.33	35.87±2.18	68.27±5.28	6.93±0.39
No height variations	24.67±3.50	55.87±6.30	80.80±4.51	90.00±2.45	31.47±4.33	65.20±7.92	7.24±0.61
No weight variations	26.27±3.39	58.13±8.74	81.47±3.44	89.73±2.39	35.20±5.91	65.47±6.17	7.62±1.65

TABLE IV

OVERVIEW OF SET-UPS USED TO GENERATE THE RESULTS IN TABLE II. ALL CONFIGURATIONS APPLY TRANSFER LEARNING TO THE BASIC VGG-FACE NETWORK [16]. TDE = *TRIPLET DISTANCE EMBEDDING*, AND TSE = *TRIPLET SIMILARITY EMBEDDING* (REF. SECTION III-D)

Config.	Description
A	No triplet embedding, using all images (original + synthetic)
B	Using TDE for triplet loss and one image per subject (original)
C	Using TDE for triplet loss and all images (original + synthetic)
D	Using TSE for triplet loss and all images (original + synthetic)

1) FRSs: The traditional Eigenfaces (PCA) FRS [49] is used as a baseline for the intra-modality methods, which also use PCA as a face recogniser. The default network[11] of the VGG-Face algorithm [16] is also used and serves two purposes: (i) as a benchmark for the performance of a traditional FRS, since it is considered as one of the leading FRSs for unconstrained face recognition, and (ii) as a baseline performance measure given that the proposed system is initialised with the parameters of this network. The only modification carried out to the VGG-Face network was the removal of the last convolutional layer that outputs the class of the input face image, to instead output a feature vector suited for verification. This vector is normalised to unit length and comparison performed using Euclidean distance.

2) Intra-Modality Methods: The intra-modality methods chosen include the popular Eigen-transformation (ET) approach [50], the Eigen-patches (EP) extension [11], and the Locally-Linear Embedding (LLE) approach [19]. Only performance metrics for photo-to-sketch synthesis are reported since all methods achieve better performance than the sketch-to-photo synthesis case [11]. PCA is used as the face recogniser similar to the implementations described in literature.

[11]Available at: http://www.robots.ox.ac.uk/~vgg/software/vgg_face

3) Inter-Modality Methods: The D-RS [2], [25] and CBR [4] systems are both considered as state-of-the-art methods that were later combined in [26] for further improved performance. The recent LGMS method [12] is also included for comparison due to the state-of-the-art performance achieved for both hand-drawn and software-generated sketches.

Photos and sketches are converted to grayscale, cropped and geometrically normalised to conform with their design specifications, similar to the approach outlined in [12]. However, DEEPS and VGG-Face exploit colour information by retaining the original RGB colour space.

All algorithms were implemented in MATLAB, with the proposed DCNN and the VGG-Face descriptor utilising the MatConvNet toolbox [51]. Graphical Processing Units (GPUs) were used for faster training and testing, with two NVIDIA Tesla K20c GPUs used for initial development of DEEPS and one NVIDIA Titan X Pascal GPU used for final evaluation.

V. RESULTS

The results of the proposed system and the algorithms considered will now be discussed, following an analysis of the effects of each proposal outlined in Section III. Additional results are also provided as part of the Supplementary Material.

A. Ablation Study

The parameters used for the proposed DEEPS system were first evaluated on the UoM-SGFS Set A database, with the best configuration then applied for Set B. The results of several set-ups as described in Table IV are shown in Table II.

1) Transfer Learning: Setup B considers the fine-tuning of the VGG-Face network parameters and the layer employing TDE using only one photo and one sketch per subject (the original images), which yields noticeable improvements compared to the VGG-Face network as a result of being trained with both modalities. Hence, transfer learning is clearly beneficial.

TABLE V

(a) MEANS AND STANDARD DEVIATIONS OVER 5 TRAIN/TEST-SET SPLITS FOR ALGORITHMS EVALUATED ON UOM-SGFS SET A SKETCHES
(b) MEANS AND STANDARD DEVIATIONS OVER 5 TRAIN/TEST-SET SPLITS FOR ALGORITHMS EVALUATED ON UOM-SGFS SET B SKETCHES

(a)

Type	Method	Matching Rate (%) at Rank-N N=1	N=10	N=50	N=100	TAR@FAR=0.1%	TAR@FAR=1.0%	EER (%)
FRSs	VGG-Face [16]	9.33±2.45	31.07±3.73	59.73±2.52	73.60±3.58	11.87±2.47	37.33±4.40	14.18±1.05
	PCA [49]	2.80±1.19	8.40±2.03	17.73±3.22	25.20±3.35	3.33±1.56	9.33±1.94	37.13±0.64
Intra-modality	ET (+PCA) [50]	8.40±2.14	30.00±3.62	54.53±5.82	67.47±2.28	11.73±2.81	34.13±4.33	16.13±2.06
	EP (+PCA) [11]	12.53±2.08	35.60±2.19	62.80±2.88	74.40±3.61	17.20±1.66	40.00±1.70	14.49±1.29
	LLE (+PCA) [19]	6.93±1.92	24.67±2.98	43.60±2.34	57.60±3.79	8.93±1.80	24.13±2.64	23.68±1.00
Inter-modality	HAOG [9]	13.60±2.29	37.33±1.94	52.67±2.62	60.27±1.67	16.93±2.29	39.07±3.73	23.12±1.78
	CBR [4]	5.73±2.09	18.80±1.28	43.33±1.94	52.40±2.14	6.67±2.45	24.27±2.65	21.62±1.29
	D-RS [2], [25]	22.13±1.45	49.33±4.24	69.87±2.18	78.67±2.26	28.53±2.38	53.60±3.35	12.30±0.93
	D-RS+CBR [26]	25.87±4.43	56.00±3.80	76.27±3.90	84.93±1.92	32.27±3.25	62.67±3.89	8.84±0.82
	LGMS [12]	21.87±5.06	51.20±4.01	72.40±3.82	80.80±2.88	28.00±6.53	55.20±4.46	12.34±1.07
Proposed	*DEEPS*	31.60±1.12	66.13±2.47	86.00±1.25	93.47±1.85	41.87±3.11	73.47±2.08	6.26±0.60

(b)

Type	Method	Matching Rate (%) at Rank-N N=1	N=10	N=50	N=100	TAR@FAR=0.1%	TAR@FAR=1.0%	EER (%)
FRSs	VGG-Face [16]	16.13±2.72	48.00±3.30	72.80±2.84	83.73±2.24	23.60±2.73	56.00±4.83	10.24±1.13
	PCA [49]	5.33±2.11	9.87±0.99	18.67±3.43	26.93±3.08	5.87±1.85	11.07±1.98	36.05±1.08
Intra-modality	ET (+PCA) [50]	12.13±1.10	39.07±4.73	63.47±3.18	75.07±3.55	16.80±1.37	43.20±6.04	13.04±1.14
	EP (+PCA) [11]	15.20±1.28	48.27±3.04	70.67±2.49	81.60±2.97	23.33±1.33	52.27±2.14	12.04±1.11
	LLE (+PCA) [19]	10.53±1.59	31.60±1.01	54.53±2.02	67.60±3.42	13.07±0.76	34.40±2.34	19.13±1.42
Inter-modality	HAOG [9]	21.60±3.67	42.27±2.89	57.07±3.39	64.80±2.23	24.40±3.90	45.47±1.37	21.49±1.35
	CBR [4]	7.60±1.98	25.47±2.38	48.27±1.30	57.73±2.03	10.67±3.13	30.53±0.99	20.11±1.21
	D-RS [2], [25]	40.80±1.45	70.80±1.85	86.40±0.89	93.07±1.12	50.93±1.38	77.20±1.97	6.42±0.51
	D-RS+CBR [26]	42.93±1.38	75.87±1.59	90.13±1.45	96.27±1.21	56.80±0.87	83.07±1.80	4.35±1.11
	LGMS [12]	43.47±4.77	73.60±3.96	86.93±1.80	90.40±2.14	53.87±5.32	79.47±2.28	6.62±1.08
Proposed	*DEEPS*	52.17±2.69	82.67±0.94	94.00±0.94	97.50±1.48	66.67±3.22	88.33±1.15	3.82±0.74

However, performance is only on par with the best-performing D-RS+CBR algorithm, results of which are shown in Table Va.

2) Data Augmentation: The use of the synthetic images created with the 3D morphable model for training the network, as performed in configuration C, yields substantial improvements across all performance measures when compared to B (which only used the original images). This enables the retrieval of 6–13% more subjects and a reduction in the error rate by 33%. Since images used to train set-up B were randomly cropped and flipped, which are two of the most common approaches employed to augment data, its inferior performance compared to C indicates that the traditional data augmentation techniques are insufficient for the task of face photo-sketch recognition.

Compared to the VGG-Face network, the proposed data augmentation and transfer learning framework increase rank retrieval rates by 15–35% and reduce the error rate by 55.9%.

3) Facial Adjustments: The proposed data augmentation strategy varies several attributes of a face image in order to generate new synthetic images that are used for model training. An ablative analysis of the effect on performance when omitting individual changes is thus performed, to determine if any set of changes is more important than others. Configuration C is used as the benchmark, since the same network set-up is used and employs all images. As shown in Table III, the omission of any group of changes leads to reduced performance, with the lack of weight and height variations leading to the largest losses. This is likely a result of these variations affecting the roundness of a face, which can easily be represented inaccurately in a sketch image. Hence, the inclusion of images exhibiting these variations allow a network to be more robust to such commonly encountered differences. The least performance loss is observed in the case of age variations. This is likely a result of sketches being generated while viewing the photos, such that the age is represented quite accurately. However, law enforcement agencies may only have old photos of a suspect, yielding significant differences between the age of the subject in the photo and the sketch. Hence, the effect of this attribute would be more significant in real-world applications.

4) Triplet Embedding: Employing a triplet embedding scheme is also beneficial, as observed in comparing the performance of configuration A (which did not employ this scheme) with C and D which both consist of A concatenated with a layer employing TDE and TSE, respectively. More specifically, the additional layer that was tuned for verification (and which also serves as a means of dimensionality reduction) enables the retrieval of approximately 10% more subjects at most ranks, and reduces the error rate by 37%.

5) Triplet Embedding Scheme: Finally, comparing the results of configurations C and D indicates that the two approaches generally yield relatively similar performance. However, TDE tends to outperform TSE. Hence, configuration C (using TDE) is selected as the final proposed architecture.

B. Comparison to Algorithms Proposed in Literature

1) UoM-SGFS Database: The results of algorithms evaluated on Sets A and B of the UoM-SGFS database may be

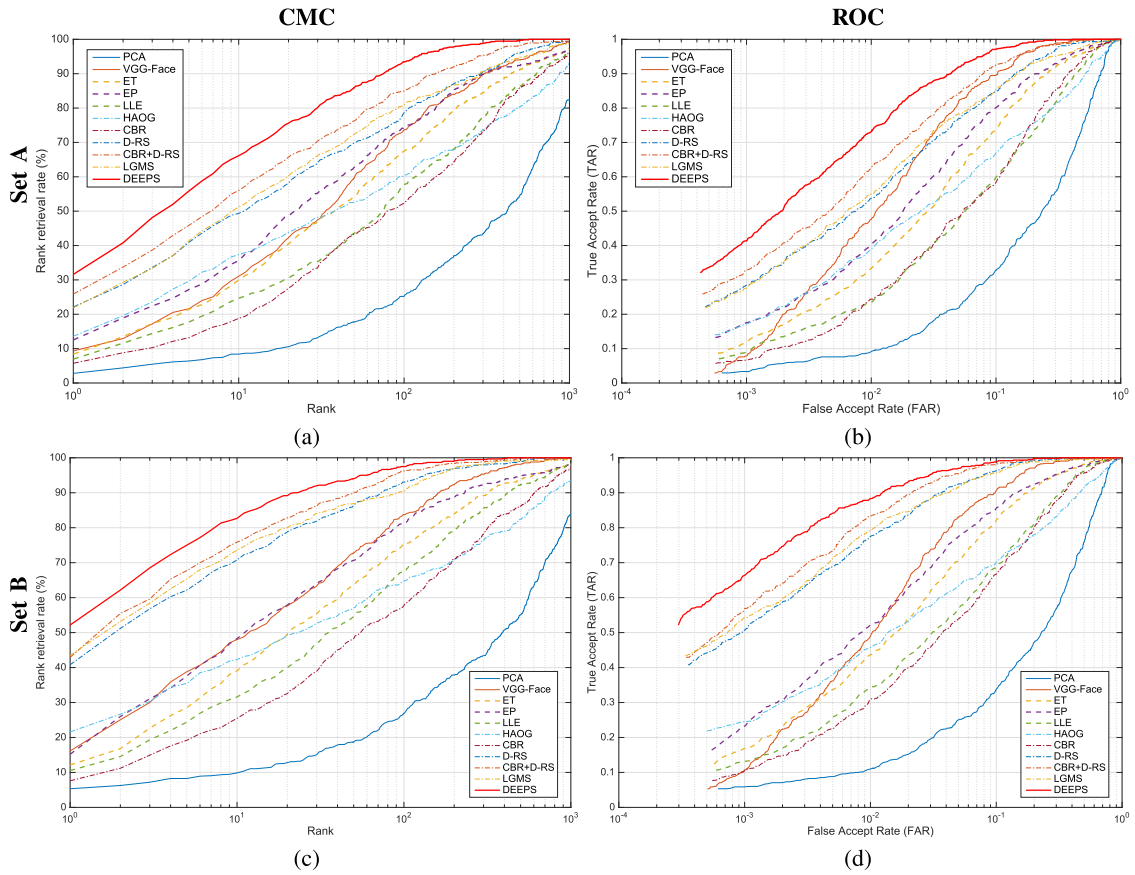

Fig. 3. Cumulative Match Characteristics (CMC) and Receiver Operating Characteristics (ROC) curves for algorithms operating on (a, b) UoM-SGFS Set A and (c, d) UoM-SGFS Set B, respectively.

found in Table Va and Vb, respectively, and in Figure 3. It is evident that the performance of all algorithms is lower on the more challenging Set A sketches, with the inter-modality methods generally outperforming the intra-modality methods at lower ranks. Interestingly, HAOG and CBR are outperformed by the intra-modality ET and EP algorithms at higher ranks. This is despite (i) the fact that CBR was designed to operate on software-generated sketches, and (ii) the perfect recognition rate reported for HAOG on a simple hand-drawn sketch dataset. The relatively low performance of CBR and HAOG indicate that the sketches in the UoM-SGFS database are indeed challenging due to distortions and deformations that mimic real-world forensic sketches.

The best-performing methods proposed in literature that are considered in this work are the D-RS and LGMS inter-modality methods, although fusing D-RS with CBR as performed in [41] yields improved performance over both algorithms despite the poor performance of CBR on this dataset.

It should be noted that the VGG-Face FRS generally yields poor performance even though it achieved state-of-the-art performance (approx. 97–99% accuracy) for unconstrained VIS face recognition, demonstrating the challenging differences between photos and sketches which require the design of specialised face photo-sketch recognition algorithms. However, its performance improves sharply at higher ranks to yield retrieval rates that are only inferior to the top-performing methods. This is largely because (i) the sketches generated with EFIT-V are not dissimilar to normal photos on which VGG-Face operates, so that the network is able to correctly identify global characteristics of the face images compared but is unable to discriminate finer details, and (ii) the VGG-Face method was trained on the same type of object as used in this work (i.e. faces). These observations validate the use of VGG-Face as the basis of the proposed DEEPS system, the latter outperforming all methods considered across all performance measures on both sets of the UoM-SGFS database. The superior performance of DEEPS shows the advantage of using features learned specifically for the task being considered rather than generalised hand-crafted features such as MLBP that are often used in such methods. Indeed, only the proposed DEEPS framework correctly retrieves over 90% of subjects by Rank-100 on the Set A sketches, with a mean match rate of 93.47%. The second best-performing method, i.e. D-RS+CBR, fails to reach this performance even at Rank-150. Together with the mean retrieval rate of 97.5% on Set B, the proposed approach yields the true identity of almost all subjects within the top 100 matches on both sets of the UoM-SGFS database. Compared to D-RS+CBR, DEEPS also reduces the error rate by 29.2% and decreases the number of mis-identified subjects at Rank-100 by 50% on the more challenging Set A sketches. Finally, the standard deviations of

TABLE VI

RANK-10 RETRIEVAL RATES (%) COMPUTED ON THE PRIP-VSGC AND e-PRIP DATABASES, CORRESPONDING TO IDENTIKIT (AS) AND FACES (IN) RESULTS, RESPECTIVELY, IN [34]

Method	Matching Rate at Rank-10 (%)	
	IdentiKit (As)	FACES (In)
Mittal *et al.* [34]	52.0 ± 2.4	60.2 ± 2.9
DEEPS	54.9 ± 3.2	80.8 ± 2.9

results are generally quite small, indicating that the network is not highly affected by the choice of subjects used for training.

2) PRIP-VSGC and e-PRIP Databases: The results of the method in [34] and the proposed DEEPS framework are shown in Table VI. As shown, DEEPS outperforms the method in [34] on both datasets, particularly in the case of the e-PRIP database despite not being re-trained on these new databases. This indicates that the proposed approach yielded a network that did not suffer from over-fitting and that generalises quite well to different types of sketches which were unseen by the system during training, where differences lie not only in terms of the general appearance as a result of the various software programs used to generate the sketches but also in terms of colour (the proposed approach was trained on coloured photos and sketches, whereas the PRIP-VSGC and e-PRIP databases contain sketches in grayscale only). It is also evident that both methods appear to be challenged with the sketches of the PRIP-VSGC database, which were generated using IdentiKit. One of the main reasons may be that the sketches themselves generally bear a subjectively poor resemblance to the original photographs, as a consequence of the software package limitations and the *other-race effect*[12] (since an Asian user generated the sketches of subjects that were mostly Caucasian); indeed, even the creators of the database noted significantly degraded performance when using these sketches and were not used extensively in their experiments [4].

VI. CONCLUSION & FUTURE WORK

This paper presented the first deep convolutional neural network-based system specifically designed for automated face photo-sketch recognition, along with an approach to circumvent the problem of having only a single sketch image per subject. It was shown that a state-of-the-art face recognition system based on a DCNN is inferior to the performance of leading face photo-sketch recognition algorithms, even when the original photo and sketch images are used to fine tune the network with transfer learning and traditional data augmentation methods. This problem stems from the typical availability of just one photo and one sketch per subject in face sketch databases, which were shown to be insufficient to robustly train a large network containing millions of trainable parameters. The proposed artificial expansion of the training set using a 3D morphable model enabled a deep network to successfully learn meaningful representations. An extensive evaluation of numerous popular and state-of-the-art FRSs and face photo-sketch synthesis and recognition algorithms showed that the proposed framework outperforms all methods considered on both sets of the new extended UoM-SGFS database that was also introduced in this paper, and on two other composite sketch databases. Since all databases used in this work are public, the comprehensive algorithm evaluation also serves as a reference to which researchers can compare future face photo-sketch synthesis and recognition algorithms, which is sorely lacking in current literature. Future work includes the use of a more advanced face morphable model to represent both photos and sketches more accurately, and which allows more flexibility in the variation of the facial features, an investigation of different training and testing strategies, and the use of other networks besides VGG-Face. Since the proposed framework was designed for the general recognition of images residing in multiple modalities (i.e. not specifically for face photo-sketch recognition), it can also be applied and evaluated for other HFR tasks such as VIS-Near Infra-Red and VIS-Thermal Infra-Red matching.

[12]The other-race effect concerns the tendency that people find it difficult to recognise persons residing in races or ethnic groups which are different to their own [4], [52], [53]

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Keith Bugeja and Dr. Alessio Magro at the University of Malta for providing two NVIDIA Tesla K20c GPUs. The authors would also like to thank the Malta Police Force for their assistance in this research.

REFERENCES

[1] 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.

[2] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

[3] S. Klum, H. Han, A. K. Jain, and B. Klare, "Sketch based face recognition: Forensic vs. composite sketches," in *Proc. Int. Conf. Biometrics (ICB)*, 2013, pp. 1–8.

[4] 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.

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

[6] *Identi-Kit, Identi-Kit Solutions*. Accessed: Jan. 30, 2017. [Online]. Available: http://www.identikit.net/

[7] *VisionMetric, About EFIT-V*. [Online]. Available: http://www.visionmetric.com/products/about-e-fit/

[8] D. Mcquiston-Surret, L. D. Topp, and R. S. Malpass, "Use of facial composite systems in US law enforcement agencies," *Psychol., Crime Law*, vol. 12, no. 5, pp. 505–517, 2006.

[9] H. Galoogahi and T. Sim, "Inter-modality face sketch recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2012, pp. 224–229.

[10] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014.

[11] C. Galea and R. A. Farrugia, "Fusion of intra- and inter-modality algorithms for face-sketch recognition," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, vol. 9257. 2015, pp. 700–711.

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

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

[14] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[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] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, vol. 1. no. 3.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[18] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.

[19] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun./Jul. 2004, p. I.

[20] 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.

[21] W. Zhang, X. Wang, and X. Tang, "Lighting and pose robust face sketch synthesis," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 6316, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 420–433.

[22] H. Zhou, Z. Kuang, and K.-Y. K. Wong, "Markov weight fields for face sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1091–1097.

[23] 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.

[24] N. Wang, S. Zhang, X. Gao, J. Li, B. Song, and Z. Li, "Unified framework for face sketch synthesis," *Signal Process.*, vol. 130, pp. 1–11, Jan. 2017.

[25] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 1513–1516.

[26] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The FaceSketchID system: Matching facial composites to mugshots," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2248–2263, Dec. 2014.

[27] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C. C. Loy, and X. Wang, "A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution," *Image Vis. Comput.*, vol. 56, pp. 28–48, Dec. 2016.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[29] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.

[30] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[31] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2892–2900.

[32] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015). "DeepID3: Face recognition with very deep neural networks." [Online]. Available: https://arxiv.org/abs/1502.00873

[33] J. C. Chen, R. Ranjan, A. Kumar, C. H. Chen, V. M. Patel, and R. Chellappa, "An end-to-end system for unconstrained face verification with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 360–368.

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

[35] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[36] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. Int. Conf. Document Anal. Recognit.*, Aug. 2003, pp. 958–963.

[37] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," *Proc. Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2009, pp. 296–301.

[38] A. Bas, W. A. P. Smith, T. Bolkart, and S. Wuhrer, "Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences," in *Proc. Asian Conf. Comput. Vis. Workshops*, 2016, pp. 377–391.

[39] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 5. 2009, pp. 153–160.

[40] G. Özbulak, Y. Aytar, and H. K. Ekenel, "How transferable are CNN-based features for age and gender classification?" in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–6.

[41] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The FaceSketchID system: Matching facial composites to mugshots," Michigan State Univ., Lansing, MI, USA, Tech. Rep. MSU-CSE-14-6, 2014.

[42] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications," *IEEE MultiMedia*, vol. 19, no. 1, p. 20, Jan. 2012.

[43] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1931–1939.

[44] S. Z. Li and A. K. Jain, Eds., *Handbook of Face Recognition*, 2nd ed. London, U.K.: Springer-Verlag, 2011. [Online]. Available: http://www.springer.com/us/book/9780857299314

[45] A. K. Jain, P. Flynn, and A. A. Ross, Eds., *Handbook of Biometrics*. Secaucus, NJ, USA: Springer-Verlag, 2007.

[46] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–8.

[47] 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 (IJCB)*, Sep./Oct. 2014, pp. 1–6.

[48] A. M. Martinez and R. Benavente, "The AR face database," CVC, Barcelona, Spain, Tech. Rep. #24, Jun. 1998.

[49] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

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

[51] A. Vedaldi and K. Lenc, "MatConvNet—Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

[52] C. A. Meissner and J. C. Brigham, "Thirty years of investigating the ownrace bias in memory for faces: A meta-analytic review," *Psychol., Public Policy, Law*, vol. 7, no. 1, pp. 3–35, Jan. 2001.

[53] R. K. Bothwell, J. C. Brigham, and R. S. Malpass, "Cross-racial identification," *Personality Soc. Psychol. Bull.*, vol. 15, no. 1, pp. 19–25, 1989.

Christian Galea (S'14) received the bachelor's degree in communications and computer engineering and the master's degree in telecommunications from the University of Malta in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree with the Faculty of Information and Communication Technology. His research interests are in computer vision and image and video processing, particularly applications related to biometrics, automobiles, and objective image/video quality assessment.

Reuben A. Farrugia (S'04–M'09) received the bachelor's degree in electrical engineering and the Ph.D. degree from the University of Malta, Malta, in 2004 and 2009, respectively. In 2008, he joined the University of Malta as an Assistant Lecturer, where he is currently a Senior Lecturer. He has been in technical and organizational committees of several national and international conferences. In particular, he served as the General Chair for the IEEE International Workshop on Biometrics and Forensics and a Technical Programme Co-Chair for the IEEE Visual Communications and Image Processing in 2014. He has been contributing as a reviewer for several journals and conferences, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO AND TECHNOLOGY, and the IEEE TRANSACTIONS ON MULTIMEDIA. In 2013, he was a National Contact Point for the European Association of Biometrics.