MATCHING SOFTWARE-GENERATED SKETCHES TO FACE PHOTOS WITH A VERY DEEP CNN, MORPHED FACES, AND TRANSFER LEARNING

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

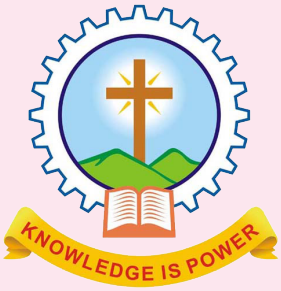
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

JITHESH RAJ J P



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

MATCHING SOFTWARE-GENERATED SKETCHES TO FACE PHOTOS WITH A VERY DEEP CNN, MORPHED FACES, AND TRANSFER LEARNING

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

of

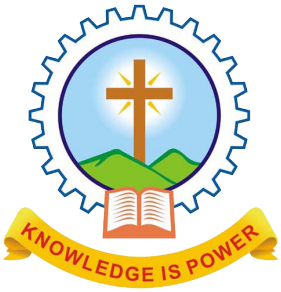
APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

JITHESH RAJ J P

Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM



CERTIFICATE

This is to certify that the report entitled **Matching Software-Generated Sketches to Face Photos With a Very Deep CNN, Morphed Faces and Transfer Learning** *submitted by* **Mr. JITHESH RAJ J P**, *Reg. No.* **MAC15CS033** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by him under our supervision and guidance.*

.....	.....	.....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	Faculty Guide	***Head of the Department***

Date:

Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the God Almighty for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof. Joby George, Prof. Neethu Subash and Prof Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

Sketches obtained from eyewitness descriptions of criminals have proven to be useful in apprehending criminals. Automated methods to identify subjects depicted in sketches have been proposed in literature. But their performance is still unsatisfactory. It is due to the unavailability of training data. The main aim is to create a face photo sketch recognition system based on very deep convolutional neural network. It is trained by applying transfer learning to a state-of-the-art model pre-trained for face photo recognition. Here we use UoMSGFS extended database which consist of 600 subject and 1200 sketches. A 3D morphable model is used to synthesise both photos and sketches to augment the available training data. This is the first deep convolutional neural network-based system. It is designed for automated face photo-sketch recognition along with an approach to circumvent the problem of unavailability of training data.

Contents

List of Figures

LIST OF ABBREVIATION

HFR	Heterogeneous Face Recognition
FRS	Face recognition System
ET	Eigen Transformation
SIFT	Scale-Invariant Feature Transform
MLBP	Multiscale Local Binary Pattern
D-RS	Direct Random Subspace
HAOG	Histogram of Averaged Orientation Gradients
MRF	Markov Random Fields
MWF	Markov Weighted Fields
VGG	Visual Geometric Group
CNN	Convolutional Neural Network
SROCC	Spearman Rank Order Correlation Coefficient

Introduction

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 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 photo-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.

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 and typically also exhibit several deformations and distortions owing to factors such as eyewitness memory loss and difficulty in describing the face. In fact, even leading Commercial Off-the-Shelf (COTS) Face Recognition Systems (FRSs) have been shown to perform poorly when matching sketches with photos.

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 and EFIT-V. Most law enforcement agencies are now using software-generated sketches, primarily due to their lower cost.

Algorithms designed for face photo-sketch recognition can be broadly categorised into two groups. 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 and their performance is not as good as inter-modality algorithms on more realistic sketches . The performance of the chosen face recogniser also depends on the quality of reconstructed images, which often contain undesirable artifacts. As a result, most recent efforts have focused on

the design of inter-modality methods, which learn and/or extract features or classifiers that maximize inter-class separability while minimizing intra-class differences .

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. However, it is unlikely that such features are optimal since they were not designed for inter-modality face recognition , and it would therefore be desirable to design and use potentially superior feature 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 tradi-tional VIS-VIS face recognition and image super-resolution. However, there has been limited work in using deeplearning for face photo-sketch recognition.

Here we are trying to solve the current issues using following method

- 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
- 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
- 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.

Related works

In recent years, there are many techniques proposed for Heterogeneous Face Recognition. Algorithms designed for face photo-sketch recognition can be broadly categorised into two groups . 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 and their performance is not as good as inter-modality[1] algorithms on more realistic sketches . The performance of the chosen face recogniser also depends on the quality of reconstructed images, which often contain undesirable artifacts. As a result, most recent efforts have focused on the design of inter-modality methods, which learn and/or extract features or classifiers that maximize inter-class separability while minimizing intra-class differences .

Several Approaches proposed in literature focus on intra-modality algorithms, also known as Face Hallucination (FH)techniques[2] . Some of the best performing and most popular methods include Eigen-transformation (ET) , Markov Random Fields (MRF) and its extension in to cater specifically for lighting and pose variations, the Markov Weighted Fields (MWF). Inter-modality methods include the Direct Random Subspace (D-RS) approach, (HAOG) method and recently, the log-Gabor- MLBP-SROCC (LGMS) method was presented.

2.1 Intra-modality algorithms

These 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 recognizer 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, these methods tend to be complex and computationally expensive. The performance of the chosen face recogniser also depends on the quality of reconstructed images, which often contain undesirable artifacts.Some of the best-performing and most popular methods include Eigen-transformation (ET) , Markov Random Fields (MRF) and its extension in to cater specifically for lighting and pose variations, the Markov Weighted Fields (MWF).

2.1.1 Eigen transformation

Eigen-transformation (ET) that syn- thesises 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 and face images and sketches are normalized in size, lighting ,poses, expression neutral and no occlusions on test image (eye glasses, beard etc.).

Here we will convert face image into Eigen face representation.From Eigen face Eigen sketches are generated. It is shown in figure 2.1. Then we will project probe sketches(test sketches) on to the sub space spanned by eigen sketches. At last we will use feature vectors of Eigen face representation to match the weight vectors.

Fig. 2.1: Eigen Transformation
First row represents original photo, Second row represents Reconstructed Photo, Third row represents Reconstructed sketch, 4th row orginal Sketch

2.1.2 Markov random Fields

We assume that faces to be studied are in a frontal pose, with normal lighting and neutral expression, and have no occlusions. Instead of directly learning the global face structure, which might be too complicated to estimate, we target at local patches, which are much simpler in structure[2]. The face region is divided into overlapping patches as shown in figure2.2. During sketch synthesis, for a photo patch from the face to be synthesized, we find a similar photo patch from the training set and use its corresponding sketch patch in the training set to estimate the sketch patch to be synthesized. The underlying assumption is that, if two photo patches are similar, their sketch patches should also be similar. In addition, we have a

smoothness requirement that neighboring patches on a synthesized sketch should match well.

During the face sketch recognition stage, there are two options to reduce the modality difference between photos and sketches:

1. All of the face photos in the gallery are first transformed to sketches using the sketch synthesis algorithm and a query sketch is matched with the synthesized sketches,

2. A query sketch is transformed to a photo and the synthesized photo is matched with real photos in the gallery.

3. After the photos and sketches are transformed into the same modality, in principle, most of the proposed face photo recognition approaches can be applied to face sketch recognition in a straightforward way

The comparison between MRF and MDF is shown in the figure 2.3

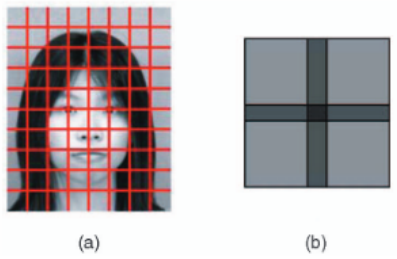


Fig. 2.2: Markov Random Field patches
(a) The face region is divided into patches. (b) The neighboring patches overlap

Here, we use the euclidean distance between intensities or colors of two photo patches as the matching measure.

2.1.3 Markov weighted fields

We firstpropose a novel MRF model that is capable of synthesizing new sketch patches. Unlike the commonly used MRF model [16, 19] in which each node in the sketch layer corresponds to a single variable (i.e., a single candidate sketch patch), each node in the sketch layer in our model corresponds to a list of variables (i.e., a list of weights for the candidate sketch patches), and a target sketch patch is represented by a linear combination of some candidate ketch patches (see Figure 1). We hence call our model the Markov Weight Fields (MWF) model. MWF model is su- perior to the commonly used MRF model in that it can syn- thesize new sketch patches and can be formulated into a convex Quadratic Programming (QP) problem to which the optimal solution is guaranteed. Note that, being a large scale QP

problem, MWF model still cannot be solved easi- ly by off-the-shelf optimization algorithms. By exploiting the Markov property of our model, we propose a cascade decomposition method (CDM) to decompose the original large scale QP problem into a number of small conditionally independent QP problems, each of which can be solved by some common optimization algorithms, resulting in a highly parallelizable computing framework. The contributions of this paper are: (1) Proposing a MWF model which is capable of synthesizing new sketch patches, and can be formulated into a convex QP problem to which the optimal solution is guaranteed; (2) Proposing a cascade decomposition method to solve the large scale QP problem based on the Markov property of MWF.

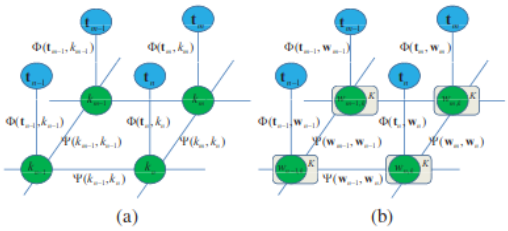


Fig. 2.3: Markov Weighted Field patches
Graphical representations of the MRF and MWF models. (a) MRF model each node in the lower sketch layer corresponds to a single variable (b) Our proposed MWF model: each node in the sketch layer corresponds to a list of variables

2.2 Inter-modality algorithms

However, intra-modality 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 and their performance is not as good as inter-modality algorithms[3] on more realistic sketches. 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. 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. However, it is unlikely that such features are optimal since they were not designed for inter-modality face recognition, and it would therefore be desirable to design and use potentially superior feature descriptors that are better adapted for the task of face photo-sketch recognition.

State-of-the-art inter-modality methods include the Direct Random Subspace (D-RS),The Histogram of Averaged Orientation Gradients (HAOG) method and log-GaborMLBP-SROCC (LGMS)

2.2.1 Direct-random subspace

State-of-the-art inter-modality methods include the Direct Random Subspace (D-RS) approach 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. D-RS was also fused with the Component-based Representation (CBR) method designed to operate on software-generated sketches by comparing MLBP features extracted from the individual facial components. The different steps in D-RS is shown in figure 2.4

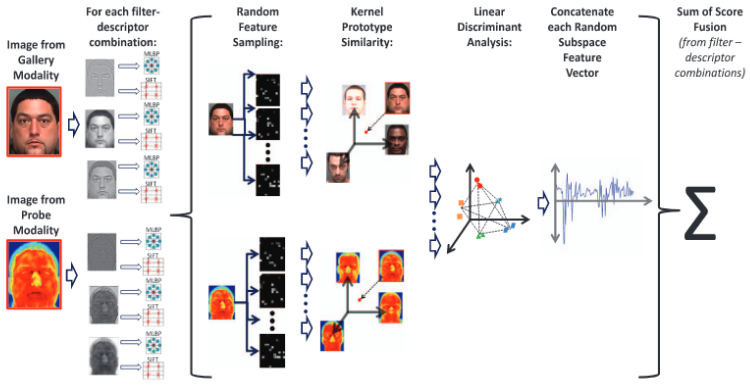


Fig. 2.4: Direct Random Subspace

The scale-invariant feature transform (SIFT) is a feature detection algorithm in computer vision to detect and describe local features in images. Image feature generation transforms an image into a large collection of feature vectors, each of which is invariant to image translation, scaling, and rotation, partially invariant to illumination changes and robust to local geometric distortion. These features share similar properties with neurons in primary visual cortex that are encoding basic forms, color and movement for object detection in primate vision.Key locations are defined as maxima and minima of the result of difference of Gaussians function applied in scale space to a series of smoothed and resampled images. Low-contrast candidate points and edge response points along an edge are discarded. Dominant orientations are assigned to localized keypoints. These steps ensure that the keypoints are more stable for matching and recognition. SIFT descriptors robust to local affine distortion are then obtained by considering pixels around a radius of the key location, blurring and resampling of local image orientation planes.

Local binary patterns (LBP) is a type of visual descriptor used for classification in computer vision. LBP is the particular case of the Texture Spectrum model proposed in 1990. LBP was first described in 1994.It has since been found to be a powerful feature for texture classification; it has further been determined that when LBP is combined with the Histogram of oriented gradients (HOG) descriptor, it improves the detection performance considerably on some datasets. A comparison of several improvements of the original LBP in the field of background subtraction was made in 2015 by Silva et al. A full survey of the different versions of LBP can be found in Bouwmans et al.

2.2.2 Histogram of averaged orientation gradients

Image gradients can be used to extract information from images. Gradient images are created from the original image (generally by convolving with a filter, one of the simplest being the Sobel filter) for this purpose. Each pixel of a gradient image measures the change in intensity of that same point in the original image, in a given direction. To get the full range of direction, gradient images in the x and y directions are computed.

Local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In practice this is implemented by dividing the image window into small spatial regions (cells), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation. For better invariance to illumination, shadowing, etc., it is also useful to contrast-normalize the local responses before using them. This can be done by accumulating a measure of local histogram energy over somewhat larger spatial regions (blocks) and using the results to normalize all of the cells in the block. We will refer to the normalized descriptor blocks as Histogram of Oriented Gradient (HOG) descriptors.

While convolution with the derivative of a Gaussian yields more accurate derivatives in the presence of noise, the smoothing that this convolution entails would remove useful detail. In addition, some degree of noise averaging will occur when histograms are computed in later steps of the HOG computation, so Gaussian smoothing is both less beneficial and unnecessarily expensive.

2.2.3 Log-Gabor multi scale local binary pattern- Spearman Rank-order correlation Coefficient

method was presented in 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 and software-generated sketches.

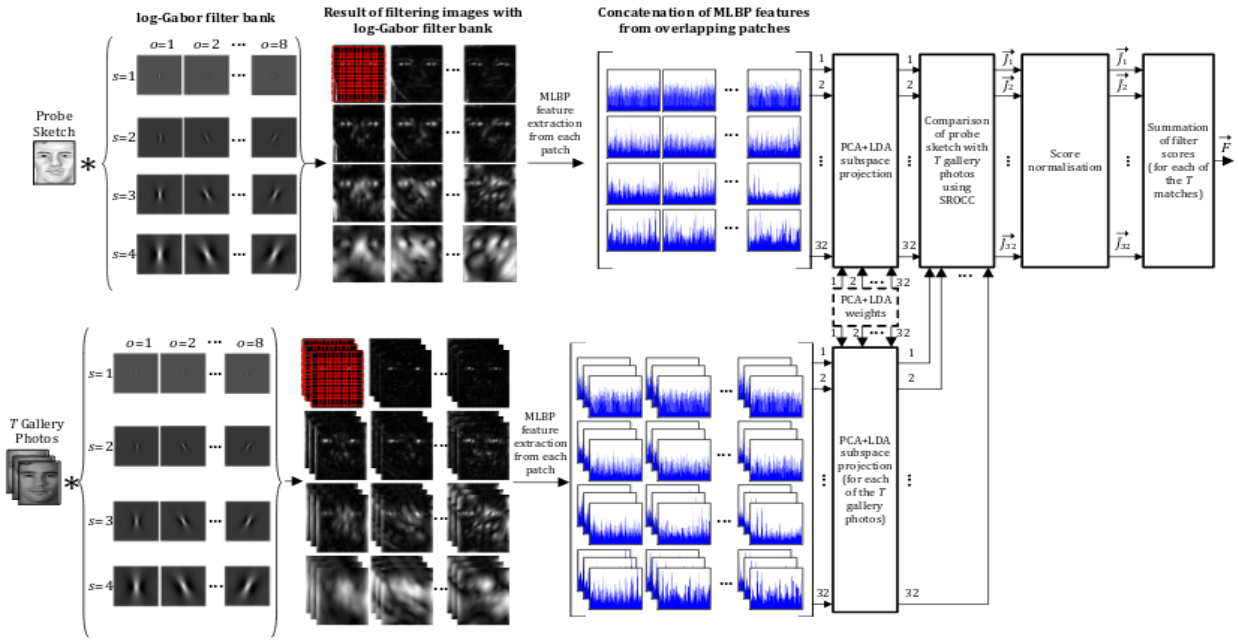


Fig. 2.5: Log-Gabor MLBP-SROCC

The system flow diagram of the proposed method is shown in Figure 2.5 First, all photos and sketches are aligned such that the eyes and mouth are in the same position for all images, which are then filtered with 32 log-Gabor filters to yield 32 images for each sketch and each photo. Gabor filters are able to represent signals localised in both time/space and frequency and have been used in a vast number of applications. Their use is motivated by the observation that these filters can model the Human Visual System (HVS) and have yielded good performance within their application domains. However, log-Gabor filters were proposed in to better model natural images, to remove the DC component, and to reduce the number of filter banks required. They are less commonly used in literature than Gabor filters and to the best of the authors knowledge have thus far not been used for face photo sketch recognition. MLBP descriptors from overlapping patches of the images derived in the filtering stage are

then extracted. While both MLBP and log-Gabor filters extract texture information, MLBP characterises the type of texture present within local areas. Hence, log-Gabor filtering extracts texture information at global level, while MLBP extracts local texture information. Following discriminant analysis, the Spearman Rank Order Correlation Coefficient (SROCC) between the resultant descriptors of the sketches and photos to be compared is found and used as a similarity measure. Whilst not often used for FR, it will be shown that SROCC outperforms popular comparison metrics. Scores are finally normalised and summed to yield the final similarity score. The proposed method is thereby named log-Gabor-MLBP-SROCC (LGMS). Further details will now be given here under.

2.3 Deep learning based method

Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.Deep learning architectures such as deep neural networks, deep belief networks and recurrent neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases superior to human experts. Deep learning models are vaguely inspired by information processing and communication patterns in biological nervous systems yet have various differences from the structural and functional properties of biological brains (especially human brains), which make them incompatible with neuro science evidences.

In this application one of the earliest and most popular deep-learning approaches is the AlexNet Deep Convolutional Neural Network (DCNN) architecture that was trained for the task of object classification[6]. 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 Facebooks DeepFace, DeepID series, Googles FaceNet and VGG-Face , which have provided important observations such as the superior performance that is generally obtained by using more layers, the benefit of a high amount of training data (especially for deeper networks having more trainable parameters), the use of multiple DCNNs 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.The only system designed for face photo-sketch recognition that utilises

deep learning concepts is implemented using 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.

Proposed method

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. The framework of proposed method is shown in figure 3.1.

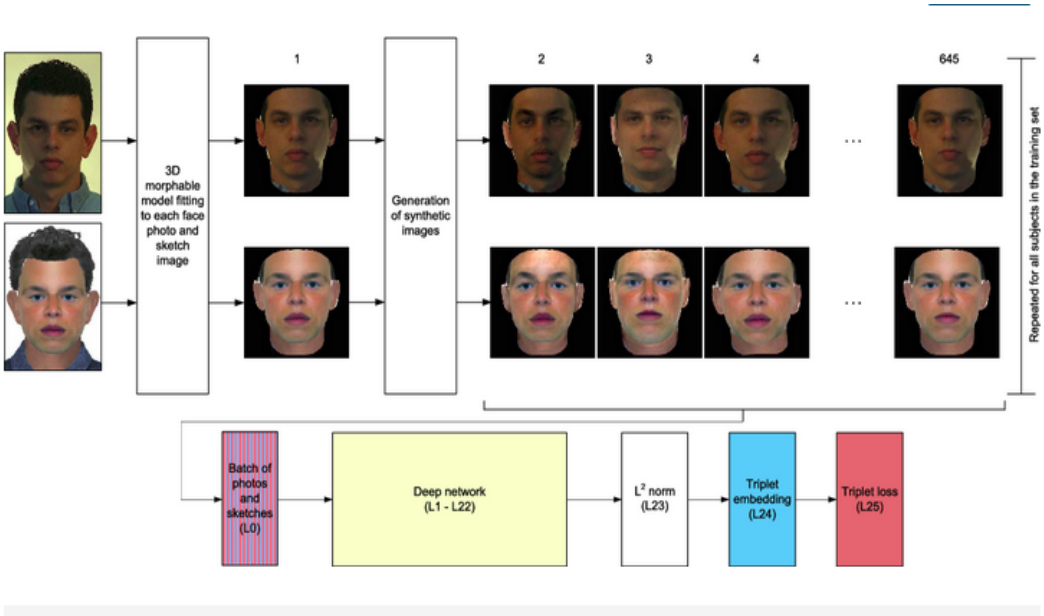


Fig. 3.1: Proposed method framework

3.1 Extended uom-sgfs database

The UoM-SGFS database is the largest software-generated sketch database that is made publicly available. All the composite sketches in the UoM-SGFS database are represented in full color and thus enable the use of color information for face photo-sketch recognition. This database contains software generated sketches of 300 subjects in the Color-FERET database(figure 3.2), created using the EFIT-V software which is commonly used by law enforcement agencies.

The EFIT-V operator was trained by a qualified forensic scientist from the Malta Police Force so as to ensure that practices adopted in real-life were also used in the creation of the UoM-SGFS database.

This database contains two viewed sketches for each of the 300 subjects considered, and is thus partitioned into two sets, where each contains the sketch of one subject. Set A contains those sketches created using EFIT-V where the number of steps performed in the program was minimised so as to lower the risk of producing composites that are overly similar to the original photo. In fact, the average time taken to create sketches varied between approximately 30 to 45 minutes. The sketches in Set A were then edited using the Corel PaintShop Pro X7 Image editing software to fine-tune details which cannot be easily modified with EFIT-V, yielding Set B. Consequently, sketches in Set B are generally closer in appearance to the original face-photos. On average, editing spanned approximately 15 to 30 minutes only, to retain inaccuracies as found in real-life forensic sketches. The Corel software was also used for sketches in Set A, but only to modify the hair component. The EFIT-V software also allows the depiction of shoulders in the sketch, which can indicate the type of clothes that the perpetrator was wearing and the physique (e.g. fat, muscular, etc.). While the type of clothing is important, more emphasis was given to correctly representing the physique of the subject since it provides more salient information. In addition, any accessories such as jewellery and hats are generally slightly different to those shown in the original photograph and sometimes omitted in the UoM-SGFS sketches to mimic memory loss effect of eyewitnesses.

3.2 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.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 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 are typically created by assigning celebrities as subjects,many photos of whom are often captured and thus allow adatabase 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

Fig. 3.2: UoM SGFS
Photos of eight subjects from the Color FERET database and the corresponding sketches (left Set A, right Set)

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 caseof 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 persubject 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 distinguish them from different identities and at the same time learn intra-class similarities . Even methods designed for face and object recognition tasks (where large datasets are available) have found data augmentation techniques beneficial for system performance .

To circumvent this problem, the use of a 3D face morphable model (along with theapproach in to fit the model to face images 4) is proposed[4]. Here each of our face models is created from a set of 3D face scans. The model has two components.

1. A mesh consisting of the mean face
2. Two matrices, for shape and texture

The number of modes of variation depends on the size of the mesh, and also is different for shape and texture. Hence the appearance of a given face can be summarised by a set of coefficients that describe how much there is of each mode of variation. 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:

1. The individual facial features 5 , and more global changes
2. Age (older or younger)
3. Gender (more female or more male)
4. Height (taller or shorter)
5. 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 as shown in figure 3.3. Sketches and photos are modified with identical parameters. Some examples of face photos and face sketches created with this approach are shown in Figure . 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.

3.3 Deep convolutional neural network

CNNs are feed forward networks in that information flow takes place in one direction only, from their inputs to their outputs. Just as artificial neural networks (ANN) are biologically inspired, so are CNNs[4]. The visual cortex in the brain, which consists of alternating layers of simple and complex cells (Hubel Wiesel, 1959, 1962), motivates their architecture. CNN architectures come in several variations; however, in general, they consist of convolutional and pooling (or subsampling) layers, which are grouped into modules. Either one or more fully connected layers, as in a standard feedforward neural network, follow these modules. Modules

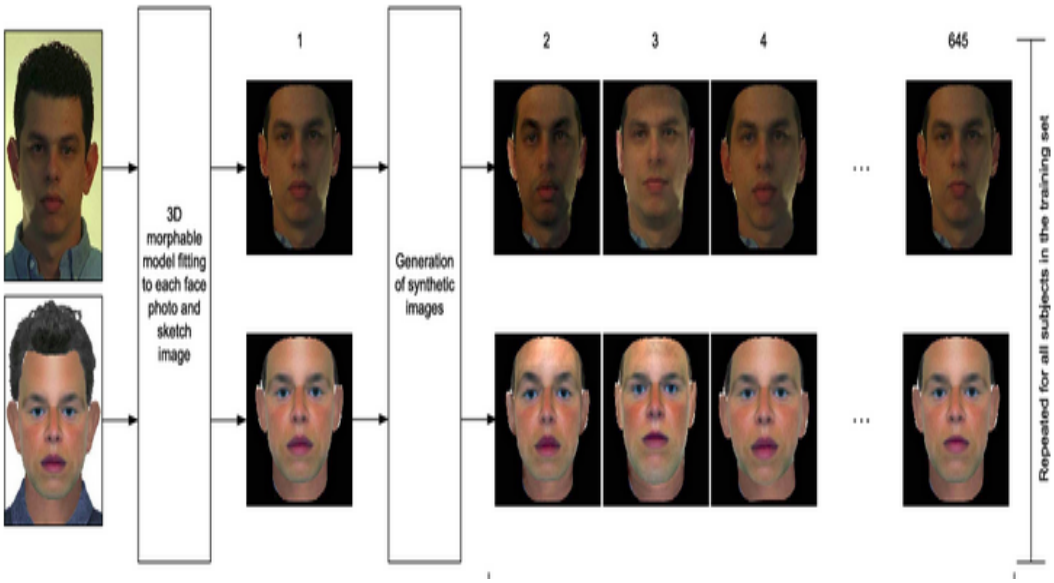


Fig. 3.3: Data Augmentation

are often stacked on top of each other to form a deep model. Figure illustrates typical CNN architecture for a toy image classification task is shown in figure 3.4. An image is input directly to the network, and this is followed by several stages of convolution and pooling. Thereafter, representations from these operations feed one or more fully connected layers. Finally, the last fully connected layer outputs the class label. Despite this being the most popular base architecture found in the literature, several architecture changes have been proposed in recent years with the objective of improving image classification accuracy or reducing computation costs.

Since performance of neural networks tends to increase with more layers and filters[6] , 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 networks 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 . 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 that was trained with 2.6M face images of 2,622 subjects 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

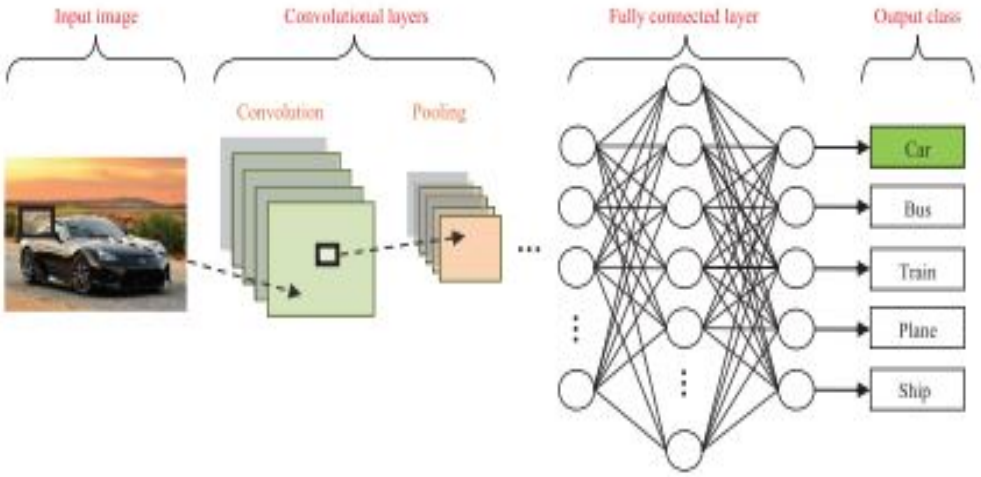


Fig. 3.4: CNN

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 trained for the task of classification using the softmax log-loss objective function (tuning layers L1L21), after which it is trained for verification using the triplet-loss objective (learningL24). Stochastic gradient descent (SGD) with momentum is used to train the network in each case. 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:

1. A layer that normalises the output feature vector to unit length (L23)
2. a fully-connected layer (L24) consisting of D inputs and L outputs
3. 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.

3.3.1 VGG face network

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 .This network was chosen for several reasons:

- It was designed for recognition of faces
- The face photos used for training represent one of the modalities used in the task of face photo-sketch recognition,
- 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. So we use VGG-Face FRS model that was trained with 2.6M face images of 2,622 subjects acquired from the Internet. The VGG-Face architecture consists of 11 blocks, each block contains a linear transformation followed by nonlinear operation such as rectification layer (ReLU) and max pooling. The first eight blocks are convolutional layers, in which the linear transformation is a bank of linear filters (or convolutional filters). The last three blocks are fully connected layers, which are the same as convolutional layers, but the size of the filters is the same as the size of the input data, such that each filter encodes information from the 229 entire image. All the convolutional layers are followed by a ReLU operation. The first two fully connected layers output are 4096 dimensions and the last fully connected layer has either N = 2622 or L = 1024 dimensions, depending on the loss functions used in the optimization. The last fully-connected layer and the softmax log-loss layer are replaced with three layers in proposed system.

The softmax activation function is often placed at the output layer of a neural network. Its commonly used in multi-class learning problems where a set of features can be related to one-of- K classes. For example, in the CIFAR-10 image classification problem, given a set of pixels as input, we need to classify if a particular sample belongs to one-of-ten available classes: i.e., cat, dog, airplane, etc.As the name suggests, softmax function is a soft version of max function. Instead of selecting one maximum value, it breaks the whole with maximal element getting the largest portion of the distribution, but other smaller elements getting some

of it as well. This property of softmax function that it outputs a probability distribution makes it suitable for probabilistic interpretation in classification tasks.

3.3.2 Transfer learning

Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task.

It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems.

We can give the new dataset to fine tune the pre-trained CNN. Consider that the new dataset is almost similar to the orginal dataset used for pre-training. Since the new dataset is similar, the same weights can be used for extracting the features from the new dataset. If the new dataset is very small, its better to train only the final layers of the network to avoid overfitting, keeping all other layers fixed. So remove the final layers of the pre-trained network. Add new layers . Retrain only the new layers. If the new dataset is very much large, retrain the whole network with initial weights from the pretrained model.

Fine tuning if the new dataset is very different from the orginal dataset the earlier features of a ConvNet contain more generic features (e.g. edge detectors or color blob detectors), but later layers of the ConvNet becomes progressively more specific to the details of the classes contained in the original dataset. The earlier layers can help to extract the features of the new data. So it will be good if you fix the earlier layers and retrain the rest of the layers, if you got only small amount of data. If you have large amount of data, you can retrain the whole network with weights initialized from the pre-trained network.

3.4 Layers in deep convolutional neural network

3.4.1 Convolutional layers

The convolutional layers serve as feature extractors, and thus they learn the feature representations of their input images. The neurons in the convolutional layers are arranged into feature maps. Each neuron in a feature map has a receptive field, which is connected to a neighborhood of neurons in the previous layer via a set of trainable weights, sometimes referred to as a filter bank (LeCun et al., 2015). Inputs are convolved with the learned weights in order to compute a new feature map, and the convolved results are sent through a nonlinear

activation function. All neurons within a feature map have weights that are constrained to be equal; however, different feature maps within the same convolutional layer have different weights so that several features can be extracted at each location .

3.4.2 Pooling layers

. The purpose of the pooling layers is to reduce the spatial resolution of the feature maps and thus achieve spatial invariance to input distortions and translations. Initially, it was common practice to use average pooling aggregation layers to propagate the average of all the input values, of a small neighborhood of an image to the next layer , max pooling aggregation layers propagate the maximum value within a receptive field to the next layer . Formally, max pooling selects the largest element within each receptive field. Figure 3.5 llustrates the difference between max pooling and average pooling. Given an input image of size 4 X 4, if a 2 X 2 filter and stride of two is applied, max pooling outputs the maximum value of each 2 X 2 region. There are also several concerns with max pooling, which have led to the development of other pooling schemes

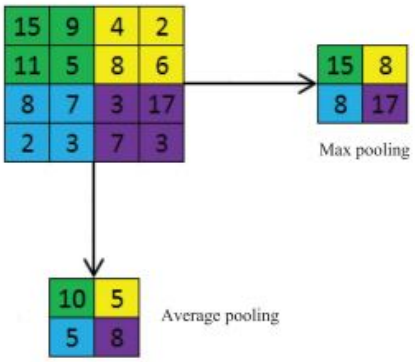


Fig. 3.5: Pooling

3.4.3 Fully connected layers

Several convolutional and pooling layers are usually stacked on top of each other to extract more abstract feature representations in moving through the network. The fully connected layers that follow these layers interpret these feature representations and perform the function of high-level reasoning. For classification problems, it is standard to use the softmax operator success was enjoyed by using radial basis functions (RBFs), as the classifier on top of the convolutional towers found that replacing the softmax operator with a support vector machine (SVM)leads to improved classification accuracy . Moreover, given that computation in

the fully connected layers is often challenged by their compute-to-data ratio, a global average-pooling layer, which feeds into a simple linear classifier, can be used as an alternative . Not withstanding these attempts, comparing the performance of different classifiers on top of DCNNs still requires further investigation and thus makes for an interesting research direction

3.4.4 Dimensionality reduction

Features from pre-trained Convolutional Neural Newtorks (CNN) have proved to be effective for many tasks such as object, scene and face recognition.

Compared with traditional, hand-designed image descriptors, CNN-based features produce higher-dimensional feature vectors. In specific applications where the number of samples may be limited as in the case of histopatological images high dimensionality could potentially cause overfitting and redundancy in the information to be processed and stored. To overcome these potential problems feature reduction methods can be applied, at the cost of a moderate reduction in the discrimination accuracy. In this paper we investigate dimensionality reduction schemes for CNN-based features applied to computer-assisted classification of histopathological images. The purpose of this study is to find the best trade-off between accuracy and dimensionality. Specifically, we test two well-known techniques (i.e.: Principal Component Analysis and Gaussian Random Projection) and propose a novel reduction strategy based on the cross-correlation between the components of the feature vector. The results show that it is possible to reduce CNN-based features by a high ratio with a moderate decrease in accuracy with respect to the original values.

3.4.5 Triplet-loss embedding scheme

Our method uses a deep convolutional network trained to directly optimize the embedding itself, rather than an intermediate bottleneck layer as in previous deep learning approaches (see Figure 3.6). To train, we use triplets of roughly aligned matching / non-matching face patches generated using a novel online triplet mining method. We learn an embedding f(x), from an image x into a feature space R^d , such that the squared distance between all faces of the same identity is small, whereas the squared distance between a pair of face images from different identities is large. The loss that is being minimized is then showed in equation 3.1.

$$Loss = \sum_{i=1}^{N} \left[||f_i^a - f_i^p||_2^2 - ||f_i^a - f_i^n||_2^2 + \alpha \right] \qquad \text{(Equ:3.1)}$$

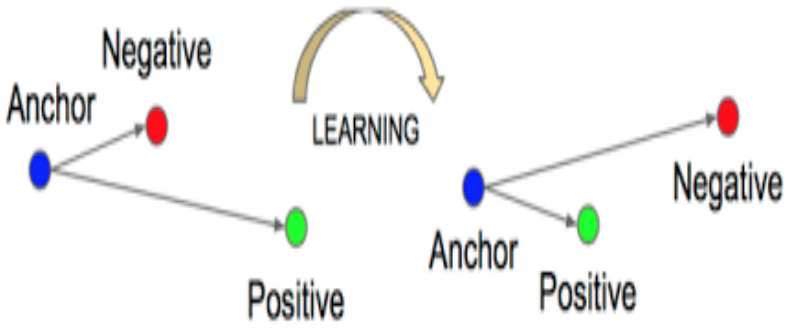


Fig. 3.6: Triplet loss embedding scheme

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 , and Triplet Similarity Embedding (TSE) based on vector dot product similarities.

3.4.6 Training and testing details

For the proposed approach, 20 percentage of the images used for training are reserved as validation data for parameter tuning. 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 50percentage probability, with the mean of the images in the training set subtracted to ascertain stability of the learning algorithm. At test time, a process similar to that employed for the original namely the dropout layers are removed and images are scaled to three sizes (256256, 384384 and 512512) to enable multi-scale testing. Feature vectors are then computed for the ten 224224 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.

3.4.7 Performance and evaluation

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 percent (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, 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. For example, a rate of Rank-50 rate of 80indicates that 80 percent of subjects have been correctly identified within the top 50 matches. True Accept Rates (TARs) at False Accept Rates (FARs) of 0.1percent and 1.0 percent are also provided, and are shown graphically in Receiver Operating Characteristics

The sketches in the extended UoM-SGFS database and the corresponding photos in the Color FERET database 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 , and 199 subjects from the FEI database 10 . To evaluate the performance of the deep learning-based method in , DEEPS is also evaluated on two additional databases: the PRIP-VSGC dataset and the e-PRIP dataset . Although both datasets contain composite sketches of the 123 subjects in the AR database, thesketches in the PRIP-VSGC dataset were created using IdentiKit operated by an Asian user while the sketches in the ePRIP dataset were

created using the FACES software operated by an Indian user. The same evaluation protocol described in 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.

Conclusion

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 photosketch synthesis and recognition algorithms showed that the proposed framework outperforms all methods considered on both sets of the new extended UoM-SGFS database.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.

REFERENCES

[1] H. Galoogahi and T. Sim, Inter-modality face sketch recognition, in IEEE Int. Conf. Multimedia and Expo (ICME), July 2012, pp. 224229.

[2] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, A comprehensive survey to face hallucination, Int. J. Comput. Vision, vol. 106, no. 1, pp. 930, 2014.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman, Deep face recognition, in British Machine Vision Conference, 2015.

[4] 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. 295307, Feb 2016.

[5] B. Klare, Z. Li, and A. K. Jain, Matching forensic sketches to mugshot photos, IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 3, pp. 639646, 2011.

[6] X. Wang and X. Tang, Face photo-sketch synthesis and recognition,IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 19551967, 2009.