Face sketch recognition using deep learning

August 2020

Liang Fan

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Supervisors: Dr Xianfang Sun, Professor Paul L. Rosin

School of Computer Science and Informatics

Cardiff University, United Kingdom

Abstract

Face sketch recognition refers to automatically identifying a person from a set of facial photos using a face sketch. This thesis focuses on matching facial images between front face photos and front face hand-drawn sketches, and between front face photos and front face composite sketches by software. Because different visual domains, different image forms, and different collection methods exist between the matching image pairs, face sketch recognition is more difficult than traditional facial recognition.

In this thesis, three novel deep learning models are presented to increase recognition accuracy on face photo-sketch datasets. An improved Siamese network combined with features extracted from an encoder-decoder network is proposed to extract more correlated features from facial photos and the corresponding face sketches. After that, attention modules are proposed to extract features from the same location in the photos and the sketches. In the third method, in order to reduce the difference between different visual domains, the images are transferred into a graph to increase the relationship for different face attributes and facial landmarks. Meanwhile, the graph neural network is utilized to learn the weights of neighbors adaptively.

The first is to fuse more image features from the Siamese network and encoder-decoder network for increased the recognition results. Moreover, the attention modules can fix the similarity positions from different domain images to extract the correlated features. The visualized feature maps exhibit the correlated features which are extracted from the photo and the corresponding face sketch. In addition, a stable deep learning model based on graph structure is introduced to capture the topology of the graph and the relationship after images have been mapped into the graph structure for reducing the gap between face photos and face sketches.

The experimental results show that the recognition accuracy of our proposed deep learning models can achieve the state-of-the-art on composite face sketch datasets. Meanwhile, the recognition results on hand-drawn face sketch datasets exceed other deep learning methods.

Acknowledgements

I would like to thank my supervisors Dr. Xianfang Sun and Professor Paul L. Rosin for their consistent support and guidance. Without their guidance and encouragement, I would have no content for my thesis. And sorry for all the extra work.

And I cannot forget to thank all staff of the School of Computer Science and Informatics at Cardiff University and the Cardiff university for all the unconditional support in academic years.

My biggest thanks to my parents for all the support you have shown me through this research.

Liang Fan

List of Acronyms

BBox	Bounding Box
CNN	Convolution neural network
CUFS	CUHK Face Sketch database
CUFSF	CUHK Face Sketch FERET dataset
DSGNN	Deep Sparse Graph Neural Network
DrLIM	Dimensionality Reduction by Learning an Invariant Mapping
E-HMM	Embedded Hidden Markov Model
EM	Expectation Maximization
FAR	False Acceptance Rate
GANs	Generative Adversarial Networks
GCN	Graph Convolution network
HAOG	Histogram of Averaged Oriented Gradients
HED	Holistically-nested Edge Detection
HOG	Histogram of Oriented Gradients
IOU	Intersection Over Union
JB	Joint Bayesian Method

KNDA	Kernel based Nonlinear Discriminant Analysis
LBP	Local Binary Patterns
LDA	Linear discriminant Analysis
LDoGBP	Local Difference of Gaussian Binary Patterns
MoNet	Mixture Model Networks
MRF	Markov Random Fields
MrFSPS	Multiple Representations on Efficient Markov Network-based Framework
MTCNN	Multi-task Cascaded Convolutional Networks
MvDA	Multi-View Discriminant Analysis
NN	Nearest Neighbor Classification
PCA	Principal Component Analysis
PLS	Partial Least Squares
Relu	Rectified Linear Unit
RBM	Restricted Boltzmann Machine
SLIC	Simple Linear Iterative Clustering
SLFNs	Single-hidden Layer Feedforward Neural Networks
SPPLayer	Spatial Pyramid Pooling Layer

Table of Contents

Publication list

Liang Fan.et al. 2019. An improved Siamese network for face sketch recognition. Kobe, Japan, 7-10 July 2019 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE

Liang Fan.et al. Siamese Graph Convolution Network for Face Sketch Recognition, 25th International Conference on Pattern Recognition (ICPR 2020).

Liang Fan.et al. Attention-Modulated Triplet Network for Face Sketch Recognition, IEEE Access.

Chapter1:

Introduction

1.1. Introduction

Face recognition systems seek to authenticate a subject from video or photo evidence using facial information. It has five advantages that other biometric identification technologies do not possess: it is non-intrusive, convenient, friendly, non-contact, and scalable. Nowadays, face recognition is applied for payments, access to premises, security, and criminal identification. In the context of criminal identification, a clear front-face image that uses a traditional face recognition cannot be captured, because the criminal suspects may deliberately avoid the range of the monitor. For example, crime suspects tend to cover any unique facial features to reduce the chance of being recognized by an electronic system. In addition, a photo of a suspect is taken much later than any photo in the dataset and thus allows mistakes of identity to be made. If the photo of a suspect is unavailable, a sketch can be taken as an important alternative way of reinforcing recognition by witnesses. The sketch is drawn according to the description by eyewitnesses and can be generated with software. It can record the location and relationships in the eyewitnesses' evidence from their descriptions and can eliminate confusing and unnecessary details. Moreover, it may contain a clue to assist detectives to build up their view of a crime scene and to question suspects

or witnesses. The automatic face sketch recognition system is a valuable aid to law enforcement.

1.1.1. Face photo-sketch recognition

Face sketch recognition refers to the matching of a sketched facial image from a set of face image. Face sketches are either complete hand-drawn sketches or made up from descriptions of features. Hand-drawn sketches are drawn by an artist and composite sketches are generated by software.

Due to the different generating mechanisms, a large modality gap caused, for example by, a different form of image representation, separates facial sketches and photos. Photos are achieved by the object's projection using shadow, space perspective method. Sketch utilizes the sparsity of the lines to embody the 3D effect. At the same time, some details may be lost or be exaggerated, because of the differences between people's memories. The result of these is that textures and shapes in sketches and in the corresponding photos are not closely alike. The main challenge is the difference in the feature representation between photos and sketches, because of the modality gap and the exaggerated description for the sketch. The early strategy for recognition was to transfer the facial photo and

the sketch to the same modality to reduce the difference. In this method, the input is the existing facial sketch and the corresponding facial photo. Then the relationship which can be used to synthesise a pseudo image between the two modalities can be learned with the machine learning method. The synthesised Pseudo image is generated based on the images' characters from the dataset. The generated image called pseudo image, because it does not exist in the dataset. Due to some face features are obtained in the process of generating pseudo image, the identification between photo and sketch is better than between photo/sketch and pseudo photo/pseudo sketch. The second method has been is to project cross-modal facial images into a common space (Kukharev et al., 2016). The third method uses feature descriptors, such as local binary patterns (LBP) (Klare and Jain, 2010), to extract similar features from the photo and the sketch. Then the features are used to measure the similarities between the two (Oh et al., 2017). If a hand-drawn sketch is studied, all these three methods confer great accuracy of identification, because the details and features are all acquired from descriptions and drawn in the sketch and therefore the facial photo and the facial sketch are closely similar. On the CUHK Face Sketch database (CUFS) (Tang and Wang, 2002) which is drawn by an artist on the basis of a front-face photo in

this dataset, the accuracy of feature-based method reaches 100% (Galoogahi and Sim, 2012a) (Yi et al., 2015) (Cao et al., 2020). However, in many crime investigations the hand-drawn sketch that is shown to witnesses is harder to capture than a composite face sketch which painted based on existing photos or an 'invisible' composite facial sketch which painted following the witnesses' description without any photos as reference would be. Although unique facial features are represented in a composite facial sketch, the loss of details makes the performance unsatisfactory. Facial recognition using deep learning has more than 99% recognition accuracy (Taigman et al., 2014) (Schroff et al., 2015) (Sun et al., 2014b) in its identification. The key advantage of deep learning is that it can learn a hidden representation from the training dataset using computational technologies. The feature representation of the deep convolutional neural networks, as a feasible approach to identification, has wider application than other methods allow. It is able to extract features, which eliminates the difference between facial photos and sketches. The recognition rate is high for neural networks constructed by learning optimal local features, even if the input image shows geometric distortions. However, the case of overfitting ultimately makes deep learning methods unsatisfactory. Moreover, the sketched images are too

simple to be used for extracting effective features. This thesis focuses on increasing recognition accuracy from hand-drawn sketch dataset and composite facial photo-sketch datasets using deep learning.

To sum up, we built three deep learning models to minimize the difference between the face images of the same class after projecting the extracted features into the common space. On the other hand, the shared-weight architectures keep the invariant features of different modalities' images.

1.2. Research aims and hypotheses

There are three challenges in facial sketch recognition. The first challenge is the gap between the modalities of a facial photo and a sketch, as seen in the difference between the features that are represented. Sketches of all kinds imply the subjective input and painting style from the artist who makes them. The structure of the facial sketch is more complex than the structure of others' images' sketches. Some facial attributes, such as the front and sides of the human face, deep sunken eye sockets, protruding nose, and two cheekbones, are difficult to represent the three-dimensional effect on 2D sketches. Thus, artists adopt structural sketch and chiaroscuro to represent three-dimensional effects. The

structural sketch is used to definite geometric showing stereo. Meanwhile, chiaroscuro refers to utilize light and dark lines which are composed of black, white, and gray color are an import method to create the depth of each facial attributes. Otherwise, in order to represent the structure of the shape, the artist usually adopts perspective drawing. For example, the artist discards a whole nostril to represent the stereo of different types of nose for the front facial sketch image. Thus, subtle elements will be ignored to highlight the three-dimensional of the facial sketch. Even if for the component sketches which are generated by software, the templates for each face attributes leave subtle elements, in order to represent stereo. Otherwise, the illumination and location are not completely the same for two images of the same person.

The second challenge is that training in the deep learning method requires datasets that are larger than any published facial photo-sketch dataset. The deep learning method uses a complex nonlinear system to extract abstract features at a higher level and thus increase the accuracy of recognition. Because the number of parameters is too great for training purposes, overfitting may impair the network performance of small datasets. If the training data are sparse, directly updating the model for all weights of neuron usually leads to overfitting and diminishes the

network performance. Moreover, some features are hidden in each node. Because deep learning systems behave like black boxes, prior information is difficult to acquire for training in the network.

The third challenge is that facial sketch datasets normally provide only one photo and one sketch for each person. Deep learning models often fail to convey reliable information from the distribution of a class. Normal deep learning methods are difficult to generate a strong generalizable model.

Hypothesis 1 (Extract effective features)**:** Traditional facial photo-sketch recognition cannot produce a high recognition rate for all facial sketch datasets, because the use of extracted features cannot help to eliminate the effect of images with different modalities. The Siamese neural network obtains a good performance on one-shot recognition (Vinyals et al., 2016), a classification task where one example of each class is given; however, it is more complex than other images in representing the features of facial images and sketches. We hypothesise the autoencoder network (Gao et al., 2015) as a channel structure used to extract features that more efficiently compare the distance between facial photos and facial sketches using contrastive loss function. Instead of Euclidean distance in the contrastive loss function, Chi-square distance is used to separate

the features and compare the distance between facial photos and facial sketches. The best result is obtained using Chi-square distance combined with NN as classification.

Hypothesis 2 (Pays more attention to the same regions)**:** The Siamese neural network ensures that two similar samples will not map to different parts of the embedded space using the same functionality for each branch. It supports the recognition of facial sketch datasets using the network of shared parameters. However, the representation of features in facial photos and facial sketches is not the same. The Siamese neural network cannot learn all features that are similar in both a facial photo and facial sketch. The mind's mechanism (Vaswani et al., 2017) pays more attention to facial attributes of images with different modalities in order to capture the represented features from the same regions. We hypothesise that the mechanism of attention tries to build and improve recognition accuracy by searching similar regions of the image, which include abundant information in order to distinguish different persons in photos and sketches. Meanwhile, after cropping the image, a spatial pyramid pooling layer (He et al., 2015) is used to reduce the information loss from using cropping as a pre-processing method.

Hypothesis 3 (Learn the relationship from image's graph topology)**:** Because the variations in the thickness and lightness of the lines increase the noise with the sketch, the features extracted from the Siamese convolution network model increases the similarity of the extracted features between different persons beyond that for the same person. We try to keep all the features of the facial photo and facial sketch using graph topology. However, it is difficult to treat the built graph topology as an irregular data structure using CNN to extract the features. We hypothesise that a graph convolution network (Defferrard et al., 2016) should be used to extract the information from the node according to its neighbours in the graph. One consideration is that graph topology keeps its fine grain to increase the recognition accuracy. Another is that topology reduce the modality gap between the photo and the sketch after transferring the facial features to a node. This method avoids inconsistency in the representation of features from the photo and the sketch. Otherwise, we adopt the Siamese network structure as the overall framework in which the graph convolution network shares the parameters of the networks', in order to learn which features from each branch are similar.

Contribution1 (Extract abundant features using encoder-decoder network**):**

We proposed an improved Siamese network to increase the matching rate using the Siamese convolution network structure. The framework used more facial features from each images' pair to reduce the modality gap. We explore the performance of three loss functions and examine the similarities between each pair. The experimental results show that our framework is adequate for a composite sketch dataset. In addition, it reduces the influence of overfitting by using data augmentation and modifying the network structure. Then, we proposed a new Siamese network combined with training classifiers (Support Vector Machine, RandomForest, and XGBoost). This model involves parameter sharing and is combined with VGG-19 as a pre-trained model to extract similar features using the contrastive loss function to reduce the data imbalance. The results show that the performance is better than the one obtained using the original Siamese network and other methods of deep learning.

Contribution 2 (Extract the similar features from the same attribute**):**

We proposed a novel triplet network with an attention module for face sketch recognition. This method uses the attention module which is adapted to the sketch image to extract information about the same attributes to deal with the similarity between a facial photo and its corresponding facial sketch. In order to avoid the

influence of the noise and distortion after cropping the images, instead of a fully connected layer in the convolution neural network, we use a spatial pyramid pooling layer to deal with the input images of random sizes without pre-processing methods, such as cropping and scaling. The network structure consists of an attention model, spatial pyramid pooling layer. Experiments show that our method achieves better performance than the state-of-the-art result on composite face photo-sketch datasets. Although the sketches in Set B are more different than the sketches in Set A of UoM-SGFS dataset, the accuracy of our attention module for Set B is higher than 81%.

Contribution 3 (Extract the face attributes' relationship**):**

We proposed a Siamese graph convolution network (GCN) for face sketch recognition. This method uses graph structure to deal with cross-modalities' gap problem. It is designed to transfer a graph structure which is generated by a facial photo and a facial sketch using a CNN and two types of superpixel method, into an embedding space with the intrinsic structural properties of graphs. The network structure consists of two graph convolutional layers on graph-structure data. It extracts more similar cross-modal graph features than were extracted by the original weight-sharing Siamese network. Experiments show the Top-1

recognition accuracy for the UoM-SGFSA dataset is better than the state-of-the-art methods.

Structure of the Thesis

Chapter 1: Introduction of the facial photo-sketch recognition project. It includes the research motivation, research aims and hypotheses.

Chapter 2: Literature review, which describes the history of this topic, its development and its relationship with this technology.

Chapter 3: This chapter describes the structure and detail of the proposed Siamese network. The performance is used to explore and compare the each pair of images. The experimental results show the capacity of our framework to present composite facial photo-sketch datasets and hand drawn photo-sketch datasets of facials.

Chapter 4: In this chapter a novel triplet network is proposed for facial sketch recognition. A spatial pyramid pooling layer is introduced into the network to deal with images of different sizes, and an attention model on the image space is proposed to extract features from the same location in the photo and the sketch,

so that the cross-modal differences between photo and sketch images are reduced when they are mapped into a common feature space.

Chapter 5: A novel Siamese graph convolution network (GCN) for facial sketch recognition is proposed, to share messages between the cross-modal graphs in this model. The graphs from both a facial sketch and a facial photo are input into the Siamese GCN for recognition.

Chapter 6: This chapter summarizes the contributions made by our project and its achievements in performance.

Chapter2:

Literature review

2.1. Facial recognition technology

Facial recognition technology uses a camera to capture an image and a video stream which contains human facial features, and calculates the nearest distance between a probe facial image and a gallery of facial images. This technology includes facial image capture, feature localization and identification. After facial image is captured, similar features can be extracted from the probe facial image and the gallery facial images, such as eyebrows, mouth and so on, to match and identify the information about the face in question without the need for personal human judgements. In ideal conditions, the recognition accuracy is more than 99% using deep learning methods on MNIST dataset and LFW dataset (Hoffer and Ailon, 2015) (C. Wang et al., 2017) (F. Wang et al., 2017) (Liu et al., 2016). MNIST dataset (LeCun et al., 1998) is handwritten digit database which consists of 60000gray-scale images for training and 10000 images for testing. The features of images in LFW dataset (Learned-Miller et al., 2016) are complex for training. However, there are five factors for reducing the recognition performance, namely illumination, facial pose, facial expression, occlusion, and aging.

1. Illumination. The lighting problem leads to a less than satisfactory recognition effect. A face is a 3D structure; hence, the shadow cast by a facial feature may

highlight or weaken the feature itself. Especially at night, insufficient light makes facial features too indistinct to recognise. The reason is that different lights shining on the same person make a greater difference than the difference between two individuals in the same light.

2. Facial position. Facial recognition is mainly based on a person's facial features. However, if the head is turned the facial image that is captured has lost some of its facial information and this reduces the accuracy of the recognition.

3. Facial expression. Different facial expressions, such as crying, smiling, and so on, also affect the recognition accuracy.

4. Masked features. In non-cooperative situations, such as a surveillance context, the captured facial image may be impossible to extract features from for recognition, because it is incomplete, being partially covered by spectacles, hats and other accessories. Thus, the captured characters, such as eyes, eyebrows, are most important for recognition.

5. Aging. The facial features may change significantly over time, thus reducing recognition accuracy. The recognition rate of the facial recognition algorithm is also different for different age groups.

In addition, the different kinds of collection equipment may change the quality of the facial image obtained. Low-resolution, noisy, poor-quality facial images reach a recognition accuracy of just 40% on TinyFace dataset using CSRI model (Cheng et al., 2018).

In summary, these factors reduce recognition accuracy for a real-time facial recognition system. However, in the context of crime, when a facial image is difficult to capture from the monitor, unique clues can come from the descriptions by witnesses. In this case, the direct solution is to match the features between the facial sketch which is drawn to reflect the features described by the witnesses and the facial photo.

2.2. Facial photo-sketch recognition

People have been drawing one another's faces since the dawn of history, trying to capture a fleeting impression of a scene or person before it changes. The recorded image looks for the essence of a person or object, instead of an accurate likeness. No two faces are exactly alike, but facial features and head shapes lend themselves to a classification system and recognize one person from thousands of portraits. We can mentally encode images of faces using these identifiable

features and store them for later retrieval. Law enforcement agencies use a sketch to aid their investigations where evidence is scant and the name of the perpetrator is unknown. In the 1880s, Alphonse Bertillon, sometimes called the father of scientific detection, developed an identification system referred to as the "Portrait Parle" or "speaking likeness." This system was a compilation of facial features taken from photographs with descriptive detail provided. Originally, Bertillon meant the catalogue to provide a method of identification that would help to identify local prisoners, but it was later found to be useful for obtaining descriptions of unknown suspects. Bertillon's classification (Laws, 2020) provided a basis for modern recall systems that would help artists to produce sketches and develop composite kits and computer systems. An early example of the composite sketch was made in 1920 after a bomb that exploded in an office on Wall Street. The investigation found a witness from the forge of a nearby blacksmith, who had shod the horse of a stranger observed carrying a covered object in the back of his wagon. In an interview the blacksmith stated that he felt capable of providing enough facial detail to let an artist prepare a drawing of the stranger. A commercial artist was hired to make a sketch which so closely resembled the stranger that he was later identified and arrested.

2.3. Face photo-sketch Datasets

hand-drawn facial sketch datasets and composited face sketch datasets were used in our research. One was the hand-drawn sketch dataset, in which a sketch is drawn on the lines of a full-face image. It shows a close similarity in its proportions and features to a photo of the same face. The other kind is the composite facial photo-sketch dataset. In this kind, the composite sketch is generated by some software program, such as IdentiKit, FACE 4.0, Mac-a-Mug, Photo-Fit and EvoFIT. The composite sketches display the important facial features, but the similarity with the corresponding facial photo is much lower than that of the hand-drawn sketch.

Figure 2-1 The examples of CUFS dataset

1. CUHK facial sketch dataset (CUFS dataset): This dataset contains 188 photo-sketch pairs and is often used in facial sketch synthesis and recognition (Wang and Tang, 2009). 123 pairs of facial images came from the AR database (Martinez

and Kak, 2001), and 65 pairs came from the XM2VTS dataset (Messer et al., 1999).

2. CUHK facial sketch FERET dataset (CUFSF dataset): This dataset is usually taken as the benchmark in methods of photo-sketch recognition (Zhang et al., 2011a).1194 subjects are translated, rotated, scaled and collected from the FERET database. Every photo has a corresponding sketch for each subject. Unlike the CUFS dataset, which uses frontal light, photos with CUFSF use variations in lighting. At the same time, the sketches contain elements of shape exaggeration. Therefore, this dataset entails more challenges for the facial sketch and photo algorithm and is closer to a practical scenario. In the experiment, the training set randomly selects 500 persons from the dataset, and the test dataset selects 694 persons.

Figure 2-2 The examples of CUFSF dataset

3. IIIT-D sketch Database: This is another popular facial sketch dataset (Bhatt et al., 2012). It can work not only with viewed sketches but also with semi-forensic and forensic sketches. The IIIT-D facial sketch database consists of three types of facial sketch datasets, namely the IIIT-D viewed, IIIT-D semi-forensic and IIIT-D forensic sketch datasets. The IIIT-D viewed sketch dataset contains 238 photo/sketch pairs collected from different sources. A professional sketch artist draws these sketches on the basis of photos. 67 sketch-image pairs are derived from the FG-NET aging dataset, 99 sketch-digital images come out of the Wild dataset, and other image pairs come from the IIIT-D student & staff dataset. For the IIIT-D semi-forensic dataset, sketches are drawn from an artist's memory or an eye-witness description, rather than being copied directly from photos. The semi-forensic dataset consists of 140 digital images from the viewed sketch dataset. The forensic dataset in IIIT-D contains 92 and 37 forensic sketch-photo pairs respectively, with the remaining pictures taken from the internet.

Figure 2-3 The examples of IIIT-D dataset

4. PRIP-VSGC dataset : This dataset is composed of three kinds of composite sketch and facial photo pairs (Peng et al., 2016b). For each image that is drawn from the AR database in the PRIP-VSGC database, three composites are created. Two composite images are generated by FACIALS, while the other one is created by IdentiKit.

Figure 2-4 The examples of PRIP-VSGC dataset

5. PRIP-HDC dataset: This contains images of 265 persons (Klum et al., 2014). As with other facial sketch datasets, it allocates only one photo and one sketch for each person. Some of them are drawn by an artist; others are generated by by the Pinellas County Sheriff's Office, the Michigan State Police and the Internet. However, not all of the 265 composites are released; at least 47 hand-drawn composites are publicly available from the Internet.

Figure 2-5 The examples of PRIP-HDC dataset

6. e-PRIP dataset: The sketch in this dataset is generated by FACE software and Identi-Kit tool (Mittal et al., 2014). It contains 123 pairs of images, including 123 images from AR face photo dataset and 123 composited images using software.

Figure 2-6 The examples of e-PRIP dataset

7. UoM-SGFS database: This dataset contains two groups of viewed sketches, which are drawn according to the colour FERET dataset using EFIT-V (Galea and Farrugia, 2016). There are two set of sketches in UoM-SGFS database. The sketches in set B are closer than in set A and have authentic photos of faces. 1200 colour sketch images of 600 subjects are in this database. This dataset is greater in size than other datasets.

(A)

(B)

Figure 2-7 The examples of UoM-SGFS dataset

8. CASIA HFB: There are 5 categories of images, including color face images, gray face images under visible light, near-infrared images, thermal infrared images, and 3d face images (Li et al., 2009). Although there are many kinds of face images, the number just has 202 subjects (persons).

9. CASIA NIR-VIS 2.0: This dataset adopts VIS and NIR cameras to capture more details from facial images (Li et al., 2013). It contains 725 pairs of images, including a set of visible light images and a set of corresponded near-infrared images. And this dataset utilizes an eye detector to correct the error of the eyes' coordinates.

We choose composited face sketch datasets (UoM-SGFS dataset and e-PRIP dataset) and hand-drawn sketch datasets (CUFS dataset and CUFSF dataset) in our research. One reason is that our model has a strong generalization of different types of sketches. Another reason is that the number of images in these datasets supports more features to obtain a stable model for training.

2.4. Literature review on facial photo-sketch recognition

Photos and sketches are generated by different generating mechanisms and typify different kinds of representation. This is not because of the artist's drawing skills, but because no artist gets enough valid information from the witnesses or victims. In extreme cases, witnesses or victims may forget all or part of the

experience, because it can cause severe trauma; hence, as a self-protection mechanism, people "seal up" the memory of it.

The main challenge in facial sketch recognition is the gap between a photo of a face and the corresponded sketch. Facial sketch recognition depends either on traditional methods or the deep learning methods. The traditional methods can in turn be sub-divided into synthesis-based methods, common space-based methods and feature-based methods.

2.4.1.Synthesis-based methods

The synthesis-based method is an effective approach that can help to reduce the modality gap between face photos and face sketches. This method obtains a good quality pseudo image which narrows the modality gap effectively.

The earliest automatic retrieval method of facial sketch images was proposed by Tang and Wang (2004). To transfer a sketch and its corresponding photo to the same domain, an eigenface approach is used to generate a pseudo sketch image from a photo image. In this step, the Karhunen–Loeve Transform calculates a set of eigenvectors from the ensemble facial covariance matrix. The advantage of this step is that it reduces both the dimensions of the image and the amount of

data. Then these eigenvectors are used to project the image into eigenface space. In the recognition stage, the reconstruction coefficient vector, sketch eigenspace, and photo eigenspace are used one by one to compute the recognition accuracy. This method has an accuracy rate of 73%. And then, they use rank-10 accuracy for identify. Rank-10 accuracy is that a correct sketch is in the top10 predictions result. The rank-10 accuracy increases to 96% on the CUFS dataset than the recognition result on rank-1 accuracy. However, the performance of recognition may be affected because principal component analysis (PCA) (Wold et al., 1987) cannot be synthesised the whole details of the sketches, especially when the subject's hair is included.

The Karhunen–Loeve Transform (Gerbrands, 1981) is a local linear algorithm; to replace it, Liu et al. (2005) have proposed a nonlinear method to synthesize a pseudo image from sketch. A pseudo image synthesizes many details of a facial image. The idea of a nonlinear method based on preserving local geometry is to compute neighbour-preserving mapping between a high-dimensional space in the original data and a low-dimensional feature space, based on a simple geometric intuition that each datum and its neighbours lie in or close to a local patch of the manifold. The synthesis weight is computed according to the surrounding patches.

In classification, the performance of the kernel based nonlinear discriminant analysis (KNDA) (Roth and Steinhage, 2000) is better than that of linear discriminant analysis (LDA) (Chelali et al., 2009) and PCA. KNDA is a nonlinear kernel trick with linear discriminant analysis. Compared with the best recognition rates of LDA and PCA, which are 85% and 64.33% respectively, the highest recognition rate on the CUFSF dataset for KNDA is 87.67%.

Gao et al. (Gao et al., 2008) propose a different nonlinear mapping method to assess the relationship between images with different modality; they synthesize a pseudo image based on a machine learning method called the embedded hidden Markov model (E-HMM). The E-HMM not only has moderate computational complexity and the ability to extract 2-D facial features, but also synthesizes a real image for different poses and in different contexts. Then ensemble strategy is used to synthesize the final pseudo image. However, blurring and noticeable block edges may be discerned in the synthesized sketches and photos because the derived pseudo-sketch patches and pseudo-photo patches are combined by averaging the overlapping regions. Moreover, much useful information for recognition may be lost if all facial photos are transformed into sketches. E-HMMI and E-HMM methods increase 23.81%

recognition accuracy than the accuracy using the nonlinear method by PCA for the dataset from the Chinese University of Hong Kong (the CUFSF dataset).

Wang and Tang (2009), taking the LLE approach (De Ridder et al., 2003), use the multiscale Markov Random Fields (MRF) model (Luettgen et al., 1993) to synthesize a pseudo image. In the MRF model the local structure of facial images is synthesized on patches of different sizes. It can learn facial structures even when presented on different scales. The sketch patch which is used in synthesizing is estimated after a photo patch and corresponding sketch patches are found in the training set. This method captures the overall facial structure and the shape of features, while the synthesized sketch is sharper and cleaner than with other methods. The result of using the multiscale Markov model and random sampling LDA is the highest recognition rate for the CUHK database. The rate increases to 96.3%, ranked first among all the eigenface methods (Tang and Wang, 2004).

Zhou et al.(2012) proposed introducing MRF into the sketch synthesis step. In pre-processing, each photo and the corresponding sketches are divided into *N* overlapping patches. Then each node in the sketch layer corresponds to a list of sketch patches. Unlike the multiscale Markov random field, a target sketch patch

is represented using a linear combination of *K* candidates of sketch patches. It is an efficient method which can synthesize a new sketch patch without needing a corresponding sketch patch in the training dataset. However, it cannot cope well with certain non-facial factors, such as hairpins and spectacles, when these factors are excluded in the training data. This method is validated on the CUFS dataset. The Rank-1 recognition accuracy reaches 80% after synthesized a pseudo image using MRF. The accuracy exceeds 95% in Rank-10 using PCA as recognition.

A novel method based on sparse coding and dictionary learning, which reduces the dimensions of the raw image patches and keeps their distinguishable characteristics can improve the effects of synthesized images (Zhang et al., 2015). It uses sparse representation to build local patches and compute the similarity scores between them by means of the nearest neighbour algorithm. Then the synthesized pseudo sketch is generated according to the possible similarities between the photo patches and the candidate sketch patches using MRF model.

In order to discover the relationships between pairs of sketch-photo patches, a probabilistic graphic model was designed (Wang et al., 2013). The reconstruction fidelity of the input image and the synthesis fidelity of the target output image are

all entered into the model. Then an alternative optimizing method which converges on a small number of iterations is used to obtain a local solution. The proposed method can handle these different processes because of the symmetry between sketch synthesis and photo synthesis. In the experiment, the proposed method attains the highest rate of recognition 97.7% on the CUFSF dataset when it is combined with RS-LDA (Wang and Tang, 2004). The results show that it is an efficient and effective probabilistic framework for facial sketch-photo synthesis. Meanwhile, the proposed method can achieve a visual image of better quality than several other methods. Furthermore, using photos synthesized by the proposed method can reach a higher rate of facial recognition than other methods. Although the cumulative scores for this proposed method are higher than those achieved by a traditional method, its performance is lower than those achieved by the feature-based method.

Chen et al. (2018) adopts a dual-scale Markov network to fuse more features than a single Markov network can. Because the scale of the patches is an important and effective factor in synthesizing pseudo images, the dual-scale Markov network, to avoid distortion, brings in larger- and smaller- scale Markov Networks in a kind of mosaic. After synthesizing a pseudo image, the effective features are

assessed for recognition. As with other recognition methods, the best result comes from fusing features of different kinds. However, the fusion of multi-information creates dimensions which are too high for recognition. Although the recognition accuracy achieves 100%, the number of images in the CUHK databases and AR databases is always less than 200. In addition, the sketches are all viewed-sketches which are drawn by the artist on the basis of authentic photos.

Peng et al. (2016) proposes to synthesize pseudo images on multiple features. These scholars use multiple filters to collect a number of features, such as DoG, SURF and LBP features, before learning the weights of multiple representations in an efficient Markov network-based framework (MrFSPS). The proposed framework could use an alternating optimization strategy; in the experiments it normally converges on only five outer iterations. The first advantage of this is that multiple representations can increase the amount of information obtained from the input images. Therefore, features which are robust as illustrations and sensitive to edge structures should be used, while irrelevant features may interfere with the synthesized result. The second advantage is that it provides a

synthesis strategy which depends on the database and scores 100% on the rank-10 rate when combined with RS-LDA on the CUFS dataset.

A new facial descriptor which called the Histogram of Averaged Oriented Gradients (HAOG) is inspired by the fact that orientations of stronger gradients are more modality-invariant than orientations of weaker gradients. It achieved 100% recognition accuracy on the CUFS database using Chi-square distance. However, the method has not been tested on large datasets, such as CUFSF and others, and it is not suitable for recognizing facial sketches which employ exaggeration. Another improved HOG feature descriptor achieves 100% on a CUHK dataset (Radman and Suandi, 2018). It is composed of an HOG descriptor and Principal Component Analysis (PCA) and it reduces the difference between sketch and photo modalities. The main advantage is that this method, unlike other existing methods, does not require training samples. The synthesized pseudo sketches simulate real sketches drawn by an artist. However, the recognition accuracy of this method relies on a high-quality pseudo sketch, especially with regard to the shapes of faces and facial attributes. Another disadvantage is that the process of synthesis cannot cope with some details, such as hairpins and spectacles. In

addition, the quality of the synthesized images depends too much on the quality of authentic photos.

2.4.2.Common-space methods

The aim of the common space method is to project facial photos and facial sketches into a common feature space. To be more specific, common space-based methods consider facial photos and sketches as cross-domain image classification problems, in which the facial photo and corresponding sketch are collected from the source and target domains, respectively.

Sharma (2011) proposes using Partial Least Squares (PLS) to gain linear projections for facial photos and sketches. Then these projections from different modalities are mapped into a common space. The nearest neighbour algorithm is used for multi-modal recognition, after maximising the common covariance. The CUHK database is used and a rate of 93.6% accuracy is attained, using a holistic algorithm.

Kan et al. (2012) uses the method of Multi-View Discriminant Analysis (MvDA) (Kan et al., 2012) to build a discriminant common space which is generated by jointly learning multiple view-specific linear transforms from multiple views. The

discriminant common space for multiple views, which is computed using jointly optimized linear view-specific transforms, is more efficient and better able to generalize. The generalized Rayleigh quotient maximizes the distance between the images of different modalities to improve recognition. This approach is evaluated by three different kinds of heterogeneous facial recognition, namely, the Multi-PIE dataset, CUHK Facial Sketch FERET dataset, and Heterogeneous Facial Biometrics dataset. The results of these three datasets secures the best performance. Especially on the CUFSF dataset, the rank-1 recognition reaches 53.4% and 55.5%, respectively, for photo-sketch recognition and sketch-photo recognition. However, the features of sketches are highly similar and the classifiers may not have a strong function.

Lei and Li (2009) propose an efficient coupled spectral regression to reveal a discriminant common space. They use two different projection methods to map a photo and the corresponding sketch into the same space. According to the abundant nonlinear and low dimensional information on the facial image, the kernel trick is used to project the image data into a hidden high dimension space or infinite dimension space. This notion refers to the use of nonlinear embedding to improve the recognition performance. In order to avoid overfitting, the least

squares regularized sense and impose are used. In the test stage, all the test images are projected into a common space. Then the cosine distance is calculated in this space before using the nearest neighbour for classification. In Lei et al. (2012), all the examples are projected from different modalities into sub-space as the discriminant and gain more discriminant power. Therefore, the kernel information is introduced into the sub-space and obtains better generalization. The rate of recognition reaches 81.43% on the VIS-NIR facial dataset.

The coupled spectral regression does not make sufficient use of discriminative information among the images from different classes. Hence, Huang et al. (2013) introduce two new regularization terms in coupled spectral regression to increase the recognition rate. Their function is to minimize the distances between low-dimensional representations of the same class and maximize the distances between the low-dimensional representations of different classes. For the viewed facial sketch dataset, the accuracy is 4% higher than it is with the coupled spectral regression method (Lei and Li, 2009).

In a previous cross-domain approach, they (Lei and Li, 2009) (Kan et al., 2012) (Sharma et al., 2011) use training data and test data to construct the common

space directly. But they do not accept the gender information. It leads to limiting the capacity to solve the practical problems of heterogeneous facial recognition. Huo et al. (2017) propose the cross-modality metric learning method of designing a suitable and efficient metric function. Metric learning supports an effective method which can learn a distance function to satisfy a set of distance constraints from a training set. One advantage of Margin-Based Cross-Modality metric learning is the intrapersonal cross-modality distance constraints which are used for minimizing intrapersonal distances. Another is that the margin which is forced between the intrapersonal cross-modality and interpersonal cross-modality distances. The interpersonal cross-modality distances that are inseparable are useful for learning the metric. The role of interpersonal cross-modality pairs is similar to that of the support vectors in support vector machines. Other methods work pairwise, but this method uses triplet-based constraints to optimize the distances between intrapersonal and interpersonal for images of different modalities and can be widely applied to different kinds of heterogeneous facial recognition, such as viewed facial sketches and VIS-NIR facial datasets.

There are two drawbacks in these methods. One is that the discriminative power of the classifiers will be reduced if the inter-modality between the extracted loss

correlative discriminative information from the images' pairs is widely different. In addition, the projection processing always loses information, which diminishes the rate of recognition. Zhang et al (Zhang et al., 2011) design a coupled information-theoretic encoding which is used at the feature extraction stage in order to reduce the modality gap. It requires the extracted codes to be uniformly distributed across different subjects, which leads to high discriminative power, and the codes of the same subject's photo and sketch to be highly correlated, which leads to a small inter-modality gap. The CUHK facial sketch FERET database is used in this test. Although the recognition rate for the proposed method is not very much higher than for others, the error rate is much lower than that of LFDA for a 0.1% False Acceptance Rate (FAR). Shi et al. (2017) adopt a joint Bayesian (JB) method (Han et al., 2016) to separate the intra- and inter-facial pairs effectively. This method uses the two inputs as samplings from two different Gaussian distributions and optimizes the asymmetric metric with respect to the log-likelihood ratio across modalities. It uses the expectation-maximization (EM) method to optimize HJB for a few iterations. It uses only two datasets for testing: one is the CASIA HFB (Li et al., 2009) and the other is the CASIA NIR-VIS 2.0 (Li et al., 2013). From the result, it seems that the original JB method performs

better than the LDA method. Compared with the original JB, the performance enhances 7-15% in VR with different FARs. With CASIA NIR-VIS 2.0, the accuracy can increase to 91.65% if combined with local Gabor method. Although this method is successfully applied in facial sketch recognition, the recognition rate is not very high, unless it is combined with other feature descriptors.

2.4.3.Feature-based methods

The feature-based method compares the similar features extracted from facial photos and sketches using local and global feature descriptors.

The first feature-based method was proposed by Klare and Jain (2010). This method proposed using SIFT (Lindeberg, 2012) and multi-scale LBP (Ahonen et al., 2004)extract gradient information. In this case, LFDA improves the recognition accuracy. Despite the high accuracy achieved by this method, LFDA does not overcome the modality difference between sketches and photos. This method uses an existing mug-shot dataset to test the validity. The success rate is from 10 to 50. The accuracy is increased between 18.37% and 44.90% if a race/gender filter is used. This result is better than Face-VACS (Huang, 2016) and the proposed method without a race/gender filter. However, the SIFT and MLBP are

not robust against a modality difference in the facial photo-sketch recognition problem. Zhang et al. (Zhang et al., 2011a) proposed to maximize the common information between facial photo and facial sketch in feature space using information theoretic. This descriptor captures more discriminative information to improve recognition accuracy.

Kernel similarities are used in Klare and Jain (2013). These similarities generate a high dimensional, non-linear representation of a facial image through compact feature vectors. It is the first effective approach for matching facial images using feature descriptors. Although both viewed and forensic sketches are drawn by an artist, the difference is that the forensic sketches are drawn not from looking at a person or photograph but following a verbal description from an eyewitness or a victim. When a forensic sketch is being made, the witness can seldom recall exactly the facial appearance of a suspect. Additionally, a disparity is often found between an artist and the eyewitness in the understanding and depiction of facial features. Thus, additional challenges are posed when matching forensic sketches against facial photographs. Klare and Jain's experiment uses five datasets, including from the near-infrared to visible range images, the thermal to visible images, the viewed sketch to visible images, the forensic sketch to visible image

and standard facial images for recogntion. As the data show, different descriptors can be used in the P-RS method to represent the probe images and the gallery images. At the same time, a rank-1 accuracy obtains good performance. In particular, the average rank-1 recognition of the P-RS method achieves a rate of 99.47% without the extended gallery for the viewed face sketch dataset (CUFS dataset). There is a 99% recognition rate for viewed sketches when SIFT and Gauss SIFT are used in the probe feature corresponding to the gallery feature. The recognition rate is 98.5% for near infrared dataset named NIR-VIS dataset when SIFT are used in the probe feature corresponding to the gallery feature. This dataset includes 1580 VIS images and 1884 NIR images in the training dataset. And 515 VIS images and 1118 NIR images are in the test dataset.

Galoogahi et al (2012) propose a Local Radon Binary Pattern framework as a new facial descriptor to directly match facial photos and sketches of different modalities. The Local Binary Pattern method encodes micro-information about facial shapes in new space. This feature descriptor does not incur complex computing and critical parameters. However, the handcrafted features, such as LBP and SIFT, are not designed for inter-modality facial recognition. The extracted features from photos and sketches may exhibit great inter-modality

variation. The result of this method is a 99.51% rate and one of 91.12% respectively for the CUFS and CUFSF datasets. However, most of these approaches involve common features which were not originally designed to solve the recognition problem of images of different modalities. Handcrafted features, such as LBP and SIFT, were not designed for inter-modality facial recognition. The extracted features from photos and sketches may have large inter-modal variations. A modality-invariant feature is urgently needed for facial sketch recognition to deal specifically with the presence of modality differences between facial photos and sketches (Galoogahi and Sim, 2012a).

The sketches and photos are similar, although they have different textures and shape distortions. Aware of this property, Alex et al. (2013) proposed a novel method exploiting the Local Difference of Gaussian Binary Patterns (DoG). It uses Gaussian Difference as an image filter to capture the most relevant features shared by sketches and corresponding photos. DoG is one of the effective extraction features of a method of approximation that uses the Laplacian of Gaussian. The image is close to the mechanism of the human retina. LBP is used to encode the DoG image for every patch to generate the feature vector of the proposed model. In order to test the recognition of this method, it uses the CUFS

and CUFSF datasets. Compared with traditional LBP and other improved methods, the recognition rate of the LDoGBP increases respectively to 96.53% and 91.04% for the two datasets, because of the superiority of LDoGBP to the LBP based descriptors. Yi et al. (Yi et al., 2015) used feature descriptors to make a high nonlinear relationship for heterogeneous facial images. The proposed method extracts local Gabor features around many facial points for the two modalities in turn. Then an unsupervised learning method called the Gaussian Restricted Boltzmann Machine (RBM) is used to learn the representational features of facial photos and sketches. All representations are concatenated before using PCA to reduce the dimensions. Finally, a Cosine metric is used to evaluate the similarity of the photo features to the sketch features. The first advantage is the local Gabor feature has strong discriminative ability in traditional facial recognition. Second, the shared representation which is learned from RBM reduces the data dimension and low dimensional data more easily prevent overfitting. Meanwhile, PCA can effectively remove the redundancy and heterogeneity for different modalities. It uses a CASIA HFB dataset and a CASIA NIR-VIS 2.0 dataset to evaluate. With the CASIA HFB dataset, if only the Gabor feature is used, the recognition rate is 50.47% in Rank 1. After using PCA, the

recognition rate jumps to 94.87%. With VR, too, the accuracy is still low: it reaches only 84.50%. After using RBM, however, the performance increases to 99.38% and 92.25% for Rank1 and VR respectively. With the CASIA NIR-VIS 2.0 dataset, the accuracy goes down to about 20% for each method by tuning the CASIA HFB dataset parameters. With the CUFS and CUFSF datasets, Rank1 achieves rates of 100% and 98%. Oh et al. (Oh et al., 2017) propose the design of a discriminative classification model to reduce the gap between modalities. It uses a novel three-layer Gabor-based extreme learning machine model. This type of model chooses hidden nodes and the output weight of single-hidden layer feedforward neural networks is determined randomly by analysis (SLFNs) (Huang et al., 2011). To begin with, a geometrically localized image blocks the input to each hidden node for a hidden layer. Then every image block is convolved with Gabor kernels followed by a magnitude function. The final decision is obtained by calculating the output of each hidden node. The model uses BERC VIS-TIR and CASIA VIS-NIR datasets to evaluate the accuracy. The rate achieved is 88% and 98% for BERC VIS-TIR and CASIA VIS-NIR when the weights are more than 5000. This method decreases the number of Gabor computations using random sampling and effectively reduces the computational cost of recognition accuracy.

In order to enhance the recognition rate, Siddharth and Kisku (2017) use two different LBPs and fuse the two descriptors for identification photos and sketch images. The two LBPs are the Modified LBP descriptor and Multi-Block LBP descriptor. LBP can increase the recognition rate though reducing the influence of illumination and pose. A modified LBP descriptor calculates local spatial relationships to obtain a simple local contrast measurement taking account of texture features. In Multi-Block LBP, it provides more robust and smooth features than other descriptors. After a single feature vector is obtained from two feature vectors using Modified LBP and Multi-Block LBP, the three feature vectors are concatenated in the horizontal direction as a descriptor. This descriptor composes a unique feature vector for each image. For testing, the distance is calculated using a distance metric called Euclidian distance or City-Block. Then a non-parametric method named K-nearest neighbour is used for classification. LDHF and IIIT Delhi datasets are used for testing. To sum up, the identification for LDHF is better than for IIIT Delhi. For the LDHF dataset, when k=20, the identification accuracy rate can be as much as 98% using City-Block; this is better than the identification that uses Euclidean distance. For IIIT Delhi Forensic and IIIT Delhi Semi-Forensic and viewed sketches, the identification accuracy is always 90% at

k=20. The accuracy is higher than it is with the original LBP method, which combines two different LBP descriptors.

Chugh et al. (Chugh et al., 2017) proposes the use of two descriptors: Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) to restrain image distortion on orientation space and Histogram of Image Moments (HIM) for improved accuracy. Before training the dataset, all images are resized to 192 × 224 pixels and a distance of 100 pixels is kept between the eye-coordinates and inter-eye distance using OpenCV's boosted cascade. In the training stage, HOG and HIM are combined as a new feature vector. The Histogram of Image Moments calculates such details as the orientation information of image moments. It also averages localized weighted pixel intensity and the centroid. However, the variation in the orientation of facial features is less or varies consistently. The Histogram of Oriented Gradients uses the image gradients to encode the intensity variations in local regions. In the test stage, the feature representation, parameters, relational knowledge transfer to the test model after extracting the feature from a test photo and corresponding sketch are obtained. This method uses three kinds of hand-drawn database along with the digital images for training. In this method, combined with transfer learning, the recognition accuracy which

uses HIM and HOG features improves by more than 5% in rank-10 on CMU-PIE dataset and the IIIT-Delhi Sketch Database. If a hand-drawn sketch is in the source domain, the accuracy achieved is 34% in rank-10. The accuracy is similar to that gained by using semi-forensic images. This method uses composite sketches as the source domain to verify the proposed algorithm. The recognition rate increases around 5% in rank-1 for composite face photo-sketch dataset than the accuracy on the hand-drawn sketch dataset. This is lower than the results obtained with semi-forensic images and it seems to us that the lower accuracy may be due to the smaller number of training images in the source domain. Using the IIITD Composite Facial Sketch Databases, this algorithm gives better identification than the existing algorithm does.

2.4.4.Deep learning based methods

Deep learning methods transform the original input into shallow features, middle features, and high-level features to a final task using the state of a hidden layer in a deep neural network. DeepFace (Taigman et al., 2014) proposes a complete facial system from facial detection, facial alignment, and facial expression to classification using a CNN model. This CNN model adopts eight layers to extract features. The convolution layers in the first three layers focus on extracting low-

level features, such as edges and textures and a max-pooling layer in between two convolution layers is used to increase robustness. Other layers comprise three local convolution layers, one full connect layer, and a SoftMax layer. The advantage of this approach is that local convolution layers use an unshared convolution kernel to reduce the number of parameters. Moreover, the DeepFace method decreases the number of max-pooling layers to avoid missing features of texture. The method is trained and evaluated on the LFW dataset. The recognition accuracy reaches 97% on the front-end of a single CNN model. The DeepID (Sun et al., 2014a) (Sun et al., 2014b) (Sun et al., 2015b) (Sun et al., 2015a) uses deep learning to capture advanced feature representation for classification. This model adopts 4 convolution blocks which consist of one convolution layer and a max-pooling layer, a fully connected layer and a SoftMax layer as a classifier. Although the structure of DeepID is similar to the original convolutional neural network, the last convolution layer of the DeepID model is used to obtain complementary and completed feature representations after inputting different facial images into the CNN model. The recognition accuracy reaches a rate of 97.25% on the LFW dataset using the DeepID models, thanks to its strong ability to generalize. The advantage of deep learning is that the multi-layer nonlinear structure in a neural

network brings a marked ability to represent features and serve as a model for complex tasks. Moreover, the deep neural network with multiple hidden layers has excellent ability to learn features. The deep learning methods can be subdivide into two methods for a facial photo-sketch recognition project: one method is class relies on synthesising the pseudo-image (L. Zhang et al., 2015) (Zhu et al., 2017) (Jiao et al., 2018). The pseudo image is synthesised from corresponding modality images using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) (Gauthier, 2014) (Mirza and Osindero, 2014). The other method is to use a Siamese network. This type of network model uses the metric learning method and matches the features which it extracts using a convolution neural network directly between images of different modalities.

Güçlütürk et al. (2016) uses a DNN model to synthesise high-quality images based on 'Perceptual Losses for Real-Time Style Transfer and Super-Resolution' (Johnson et al., n.d.). the loss function for obtaining a synthesized image which is close to a real image consists of three loss functions, that is, a standard Euclidean loss function, a Euclidean loss function and a pixel's loss function. The integrated loss function compares features and calculates the features of the image in pixels, which use a standard Euclidean loss function to measure the distance between a

real image and on that is predicted. Finally, RS-LDA (Wang and Tang, 2004) is used to verify the recognition accuracy between synthesized colour images and real images. The accuracy rate achieved is 99.79%, using line sketches to synthesise colour images. However, one disadvantage is the model generates only one image from one sketch. Another is that the synthesized colour is not precise for all images. In convolution sketch inversion, the edges of pseudo sketches are not clear. Sangkloy et al. (2016) uses a generic feed-forward network to synthesize a realistic image. The generic feed-forward network is built to learn the process of transformation between a sketch and a colour photo. In addition, the network obtains more details from facial sketches after training. The architecture is an encoder-decoder network with three residual blocks. In the encoder stage, three convolutional layers are used to extract more features as in a convolutional sketch inversion. In the decoder stage, two deconvolutional layers replace bilinear upsampling layers by residual blocks. In order to avoid overfitting because of the small numbers for training, it uses 21,848 images of different sketch styles for pre-training, and uses the parameters to train the network. From the result, the shape of the synthesised photo looks clearer and has higher resolution than others which are combined with adversarial loss procedures.

However, a blurred outline is still observed in some images, because uncommon colours cannot be generated using an adversarial loss function in this network. Although the synthesis-based method using deep learning can obtain a realistic facial image and get high accuracy with all kinds of facial sketch datasets, the methods available for deep learning are fewer than for traditional methods. In particular a cycle-consistent adversarial network (Zhu et al., 2017) not only generates sketch images from a photo but also generates a photo image from a sketch. Moreover, this network can ascertain the obtain the relationship between facial photos and sketches without paired training data. The main reason is that deep learning method relies on a huge number of training samples, but a large number of facial sketch datasets is hard to amass. Another method exploits a kind of network structure to gain large numbers of data.

Another deep learning method for facial photo-sketch recognition is based on Siamese network architecture. The earliest Siamese model was proposed by Bromley et al. (1994). They designed a multi-branch network structure which consisted of two identical networks with shared parameters to extract the features of the input images. Although the CNN model gets high scores for facial recognition, it is not suitable for real-time facial recognition systems or real scenes.

One reason is that the monitor captures one facial image without constraints. The captured face photo which is not obtain by perfect environment may be not clear to extract feature for recognition. However, the CNN model does not extract features from small datasets efficiently. In a facial recognition system, the best advice is to recognize the person who corresponds to the image from a single facial image. Thus, Chopra et al. (2005) obtain the feature vector by replacing the original structure with a convolution neural network to extract the description operator, and then use the feature vectors of the two pictures to determine the similarity using the contrastive loss function. The main idea is to map the input pair of images into the target space and use a distance such as Euclidean distance to compare the similarities in the embedded space. The DrLIM method (Hadsell et al., 2006) maps the data from the high-dimensional space into a low dimension using the relationship between samples. The advantage is that it preserves the relationship between the input data using nonlinear transformation after dimensionality has been reduced. Khalil-Hani and Sung (2014) propose a simple Siamese network based on a convoluted neural network. Instead of the max-pooling layer, a subsampling layer is adopted to increase transformation invariance. Therefore, the subsampling pooling layer reduces the number of

parameters using a large receptive field to avoid the overfitting problem. Otherwise, the pooling layer efficiently reduces the dimensions of the features. Next, to reduce the effect of reducing nonlinear dimensionality, stochastic gradient descent is used to minimize the loss function. When evaluating the verification performance for this Siamese network, Equal error rate (EER) is the threshold for the same value of False Rejection (FR) and False Acceptance (FA). This index is used to evaluate the model's performance. EER of the Siamese network reaches 3.33% on the AT&T facial database for testing. Zagoruyko and Komodakis (2015) explode three types of Siamese network to match image patches, namely a two-channel network, a shared weight Siamese network, and a pseudo-Siamese network. The difference between the two types of Siamese network is that the pseudo-Siamese network adopts unshared weight for the training stage. The similarity of the shared weight Siamese network and the pseudo-Siamese network is that each branch is a feature extraction descriptor, and the last layer is used for the function of calculating the similarity of the feature vector. Unlike the two Siamese networks, two-channels network makes its input an image pair which it combines into a single dual-image channel. This method is significantly more flexible and easier to use in training than others. An improved

HybridCNN (Melekhov et al., 2016) is adopted to measure the similarities between the feature vectors of similar image pairs and dissimilar image pairs using Euclidean distance. Each branch of the Siamese network consists of convolutional layers, a rectification layer as a nonlinear convolutional layer, and a fully connected layer to extract the optimal feature vector. However, the contrastive loss function does not work for the feature vectors of large distances. For sketch recognition, getting information on the missing colour and texture is the main challenge. To obtain more information about the spatial structure, one efficient method is to use the large-scale convolution kernel to replace the small convolution kernel in the convolution layer. The Sketch-a-Net model (Yu et al., 2017) uses a multi-scale multi-channel deep neural network framework for hand-drawn sketch recognition. First, a large kernel is used in the DNN model to extract the textural features from a sketch. Second, the LRN layer is removed from the DNN model to reduce the number of parameters. The reason is that the source of illumination is not shown in sketches. In spite of removing the LRN layer, the performance of this model cannot be affected. If it could, a high dropout would reduce the risk of overfitting. Sketch-based image retrieval (Qi et al., 2016) uses the Siamese network architecture to reveal the intra-class variability between one

sketch and another. In one-shot learning, the Siamese network (Koch et al., 2015) solves the target problem by learning the characteristics of a specific field or generating a hypothesis that can distinguish attributes. Except for the first layer using ReLU as a activate function, the sigmoid function is adopted in other layers. For the pair of input images, this method to avoid data imbalance adopts random sampling and generates pairs of images as training data from the training dataset. In addition, the output of the classification layer can serve as an attention mechanism for the extracted features. The process with this layer is that the extracted features for the image pair uses L1 distance (the absolute value between the feature vectors corresponding to the two images) to multiply by a weight which is a set of parameters generated by training. These methods all adopt the contrastive loss function to train the model. The main idea of this loss function is to increase the difference between intra-classes and to reduce the difference between inter-classes. However, this loss function needs to specify a margin for each pair of images. Thus, the margin is fixed for all the training and testing samples. Further, it keeps the embedding space for training. The Center Loss function (Wen et al., 2016) is proposed to determine a flexible margin based on a SoftMax loss function. The Center loss function increases the constraint

distance of the samples between the feature space and the class centre after calculating a centre for the class of each sample. The aim is to pay attention to the distribution of intra-class samples, in order to minimize the distance between the intra-class samples using the calculated centre of the class. Range loss function (X. Zhang et al., 2017) adds an intra-class constraint to close the distance of the same sample. Another Siamese network structure is proposed by Wu et al. (2017a), who use it to recognize facial images of different modalities. This method uses a new trace norm on the fully connected layer. The first advantage it has is to enhance the correlation between different modalities. Another advantage is to constrain the bound of the parameters to prevent overfitting on a small number of datasets. However, most of the heterogeneous facial datasets contain fewer than 200 images. Besides, there are not enough labelled NIR and VIS images for training by the SoftMax loss method. The complexity of a triplet loss may affect the quality of the images, and a simple triplet loss slows down the training, but the cross-modal triplet ranking can compensate for these weaknesses. An improved convolution neural network can be used to synthesize a pseudo sketch (Jiao et al., 2018). The core of CNN is made up of the convolution layer, shared weights for each layer and a pooling layer. The function of the convolution layer is to

produce a set of feature maps. However, there are too many parameters in the convolutional layers for training purposes. In order to reduce the number of parameters that might cause overfitting, a pooling layer which partitions each feature map into non-overlapping rectangles and outputs the maximum for each rectangle is connected to the convolutional layer. In addition, the main sources of recognition accuracy are the photo and sketch in a training pair which are usually not registered in full. The overlapping max-pooling layer which receives the maximum value within the receptive field using a sliding window is used to keep the original resolution of the sketches.

He et al. (2017) propose a Siamese network called W-CNN to minimize the Wasserstein distance between NIR and the feature distribution of VIS modalities. The main idea of optimizing is to project the invariant features of the modality into a low dimension subspace, and combine the output of the two channels to compare the distance. Compared with using contrastive loss in a Siamese network, it combines triple loss and the contrastive loss function to reduce the modality gap. Another problem is overfitting, which is a major problem for a small dataset in deep learning, for two reasons. One is that heterogeneous facial datasets are all small and insufficient. The other is that the number of parameters

is much greater than the number of datasets. Redundant parameters from a fully connected layer are the main reason for overfitting, A correlation prior is introduced into the fully connected layers of deep models to mitigate the overfitting problem on small-scale datasets. Kazemi et al. (2018) propose a new loss function for deep coupled network structures to enhance the recognition accuracy. Unlike the existing methods which are feature-based, this method focuses on relevant facial attributes. After projecting the features into the embedded space, the proposed loss function fuses the facial attributes provided by eyewitnesses and the geometrical properties of forensic sketches to improve the accuracy. Although contrastive loss can reduce the distance between the same samples and increase the distance between different samples, it is still used as a part of the new loss function. Moreover, most images of different modalities cannot be separated as images of the same modality can. Therefore, two loss functions join to distinguish them. One is called attribute loss and minimizes the intra-class distances of photos or sketch-attribute pairs which share combinations of facial attributes. The other loss function can prevent pushing the centres and keep a minimum distance if all the centres converge on a single point in the embedding. Iranmanesh et al. (2018) propose a deep coupled model to match

facial images with different modalities. The recurrent problem of overfitting is remedied in this method by three devices. First, a pre-trained model (VGG-16) is used. An increased parameter occurs after the convolution layers and several fully connected layers. The model saves many parameters if a trained model is used for new data without retraining. especially if the number of parameters in the fully connected layer increases more often than before. In order to reduce the number of parameters, a global pooling layer can replace the max pooling layer. The last device is to build an unshared Siamese network model. Unlike the traditional Siamese network, the parameters for each layer are not shared, despite optimizing with the contrastive loss function. For embedding the data from the network, PCA and T-SNE use dimension to reduce the output and visualize it on two dimensions. It uses CMU Multi-PIE and a Notre Dame LWIR facial dataset for training, unlike other methods, to obtain a set of parameters. The test dataset is the Polarimetric Thermal Facial dataset. The exploiting polarization information extracted from the network increases the Rank-1 identification rate to 94% and 88% for the Polar and Thermal images, respectively.

Siamese networks have two ways of measuring the samples' distance. Triplet networks (Hoffer and Ailon, 2015) (Parkhi et al., 2015) (Hermans et al., 2017)

add an anchor sample as input data. Thus, the input data of this network is composed of three images, either two positive images and one negative image or two negative and one positive. The principle of the triplet loss function is similar to that of the contrastive loss function. The aim is to reduce the features of the anchor sample and of each positive sample. FaceNet (Schroff et al., 2015) uses a triplet network model composed of three convolution neural networks as branches to map facial images into Euclidean space for measuring the features' distance. Instead of a fully connected layer as a classification layer, this model adopts a 1*1 kernel size convolution layer as an embedding layer to extract the features of a facial image. Then the triplet loss function selects a large mini batch to increase the number of samples for each batch. It increases the speed of convergence in the training stage. The recognition accuracy reaches 99.63% and 95.12% with LFW datasets (Huang et al., 2008) and the Youtube Facials DB dataset (Wolf et al., 2011) , respectively. Galea and Farrugia (2018) propose using a CNN network to match software-generated sketches and authentic facial photos. Either a hand drawn sketch or a software-generated sketch always has a small dataset. In order to reduce the overfitting problem, a useful data augmentation is to use a 3D morphable model (Bas et al., 2016a). The VGG-

facial model (Huang et al., 2008) as a pre-training model is used to support fast convergence. The proposed training framework consists of a deep CNN, a triplet embedded to optimize the features for verification, and a data augmentation approach to circumvent the lack of multiple images per subject. An embedded triplet adopts a triplet loss function to reduce the Euclidean distance between the target sample and the same subject, while increasing the distance between the target sample and a sample from a different subject. Two software-generated sketch datasets are used, among such datasets as UoM-SGFS, PRIP-VSGC, and e-PRIP. The best result exceeds 60% in Rank-10 for all the datasets that were used. The results do not show very high accuracy, even with VGG-facial as pre-trained model and triplet loss function. Zhang et al. (2017) propose a cross-model network based on a convolution neural network to extract the correlation facial features from facial images of different modalities . Comparing it with traditional CNN to extract features, the GAN model generates points of close similarity from a real image. In this method, GAN is used not only to generate a pseudo image for increasing the similarity between two images of different modalities, but also to capture some common features. Then the inputs of cross-modal CNN are real colour images and the corresponding pseudo images. This cross-modal CNN is

a kind of Siamese network except for the loss function. The loss function consists of a Softmax loss function and a correlation loss function, which ensures the same categories and minimizes the distance between them in feature space. Because this method involves matching 2D facial images and 3D facial images, three datasets are used: BU3D, Bosphorus and CASIA-3D dataset. The recognition result is related with parameter λ. If $0.4 \leq \lambda \leq 0.8$, and yields good and stable results. The accuracy reaches 96.88%.

Most research achieve high recognition performance on traditional machine learning methods. For the deep learning method, most of the projects adopt synthetic-based methods to reach a high recognition rate. However, the time for generating pseudo images is too long for application. Our research focuses on improving face photo-sketch recognition accuracy and increasing the recognition speed using deep learning methods between face photo and face sketch directly.

Attention mechanism using deep learning

To improve the efficiency and accuracy of facial photo-sketch recognition, we decided to use an attention module to focus on the features of similar locations and a spatial pyramid pooling layer in a triplet network. From a set of states in the

network, the attention mechanism (Ba et al., 2014) (Li et al., 2018) (Li et al., 2018) (Xu et al., 2015) selects a state similar to a given one, and then extracts information from it. The aim is to compare the similarities between the vector collection $v_1, v_2 \ldots v_n$ and to assign big/small weight values respectively to those with high/low similarity. Minh et al. (2014) proposed an attention mechanism based on a reinforcement learning model. It extracts the information from a picture or a video, and selects a range of regions or locations, which are then processed at high resolution. The recurrent model is allowed directly to train for a given task using past information and mission requirements. It not only extracts the features from the whole image, but also extracts the necessary features using the relationship between the image pixels. Vaswani et al. (2017) use a self-attention module to capture the related global information by means of the hidden state from the source input and the target input data. Unlike the traditional attention mechanism, their mechanism uses the relationship between the source and the target data to obtain the dependency of the source and the target data. Yin et al. (2016) apply an attention mechanism to a convolution neural network. The CNN model selects the down-sampling layer to preserve the scale and spatial invariance of the input images. However, the pre-selected fixed size is limited so

as to adapt to the deformation, and for this reason the feature map of the images cannot reflect the deformation of any image or the features of the image as a whole. In the spatial transformer network (Jaderberg et al., 2015), the localisation network uses a sub-network that is a component of the CNN model to generate the spatial transformation parameters which can transfer the input map to the expected output map. The learned spatial transformation network automatically extracts the local data features from the area under attention and eliminates the deformation of the target image by applying it to a reverse spatial transformation. SENet (Hu et al., 2018) builds a correlation between feature channels to intensify the important features for recognition. The core idea of SENet is to learn feature weights based on the loss function through the network and automatically note the importance of each feature channel. The useful features in these tasks are increased and better results are achieved by attending to the relative importance of features. The network extracts the spatial information as a 'global descriptor' before two fully connected layers generate a feature map. Finally, the feature map is multiplied by the original space after global average pooling, to recalibrate the output feature map. However, this does not reflect the significance of attention in the spatial dimension. The similar features are related to each other without

distance. Woo et al. (2018) separately apply the attention mechanism to the channel and to the spatial dimensions to improve the ability of network models to extract features without the need to significantly increase the amount of calculation and the parameters. The channel and the spatial attention modules generate refined features by working on the input feature maps in sequence.

Chapter3:

An improved Siamese network

3.1. Introduction

We propose a novel network based on a shared-weight Siamese structure. This network architecture is composed of identical convolution networks as branches to extract similar features which can reflect the relationship between a photographed face and a corresponding facial sketch. Next, two autoencoder-decoder networks are built to keep as much information as possible without noise from the photo of the face and the sketch of it. Finally, the features from the Siamese network and the features from the autoencoder-decoder network are fused to increase the diversity. One advantage is that the framework is used to extract useful and more detailed features from each pair of images to reduce the modality gap between the photo and the sketch which is produced by the imaging principles. Another advantage is that, the input data format of the Siamese network avoids the problem of overfitting which increases the amount of input data to several times more than the amount of the dataset. Unlike traditional methods which obtain high performance when viewed from datasets of drawn faces in photos and sketches, deep learning methods always keep the accuracy less than 70%. One reason is that some of these methods cannot maintain the relationship between the image modalities using a threshold value after training.

Another is that the case of overfitting and underfitting loss functions causes low recognition accuracy. At the same time, limited methods of deep learning are applied in face photo-sketch recognition. Our model not only achieved high recognition accuracy on viewed datasets of the above kind, but also was evaluated by using composite face sketch datasets. Our Siamese network makes three main contributions:

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

2. Based on the contrastive loss function, our network is designed to compute the distance between the inter-modal class and the intra-modal class, with an interval between the modal samples. The methods of nearest neighbour (NN) classification are used to compare the similarities, using a threshold value and increasing the accuracy of the recognition.

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

4. Three types of feature are extracted and fused after training the proposed type of Siamese networks. These features are fed into the training of classifiers to increase recognition accuracy. The weight of each class in the Siamese network is optimized at the training stage to reduce the negative effect of the data imbalance. Unlike the original Siamese network which obtains a result using a threshold value, we propose to use the classifiers trained on the features extracted by means of the Siamese network that we have designed.

3.2. The proposed Siamese network architecture

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

parameters which lead to poor performance on test data can be reduced more effectively in other ways.

However, the face sketch datasets are all small in size, containing one photo and one sketch for each person, so most of the deep neural networks cannot be applied effectively. The Siamese network is used after training by using a contrastive loss function by involving two identical neural networks as two channels to extract the features from two images and compare the similarities. The input shape of the Siamese network is a pair of images which consists of two images and a label. Each input needs to be a pair of photo and sketch. The network provides an effective way to compare the similarities and alleviates the overfitting problem.

As shown in Figure 3-1, the proposed Siamese network architecture consists of two identical convolutional networks as channels which accept the distinct images as inputs and share weights to extract the features from both the photos and sketches of the face. We adopt the sharing weights model among the two channels' convolutional networks to map the features extracted separately from two input images into a common space, using the contrastive loss function. After the last layer from each channel of the Siamese network, the outputs of its sub-

networks are fused and trained using the contrastive loss function to minimise the distance between the pairs of positive images and to maximise the distance between the negative pairs of images.

The modality-invariant parameters of the Siamese network make it possible to learn the relationship from a pair of images (consisting of the images of each subject in the two modalities) and apply the learned relationship for testing. Because the sketch uses monochromatic lines to reflect the object's structure, a convolution network of small kernel size makes it difficult to extract enough textural features for recognition. For example, in the photo image, according to the grey-level distribution of pixels and the surrounding space, the small round patches can be recognized as human eyes. However, the sketched representation pays attention to the shape, and therefore they may not be recognized as eyes if we use the extracted texture features. Thus, we use a large kernel size to capture more structural features than texture features for the purpose of recognition. Except for the first convolution layer without padding, the kernel size of the other convolution layers is 7*7. All the convolution layers are set to involve an activate function 'rectified linear unit' (RELU) to make nonlinear mapping which can keep the learning rate faster and strengthen the

representational ability more than other activation functions can. In addition, the linear model avoids saturation though predigesting the process of back-propagation. Moreover, all the photo-sketch datasets of faces that we used come from full-face images. Thus, the images do not need any padding to change the number of the photos' pixels and the corresponding sketches' pixels for the three convolution layers in which the characters are all mapped. This means that all dimensions are valid so that the input image gets covered by the filter and the stride, the filter window stays at a valid position inside the input map. In the last four convolution layers, padding is used to increase the number of pixels for the input images. Even if it may lose some features on the border of the image, this carries less information in facial photos and sketches. One advantage is that padding keep sufficient features for deeper layer, and another is that the number of parameters is less.

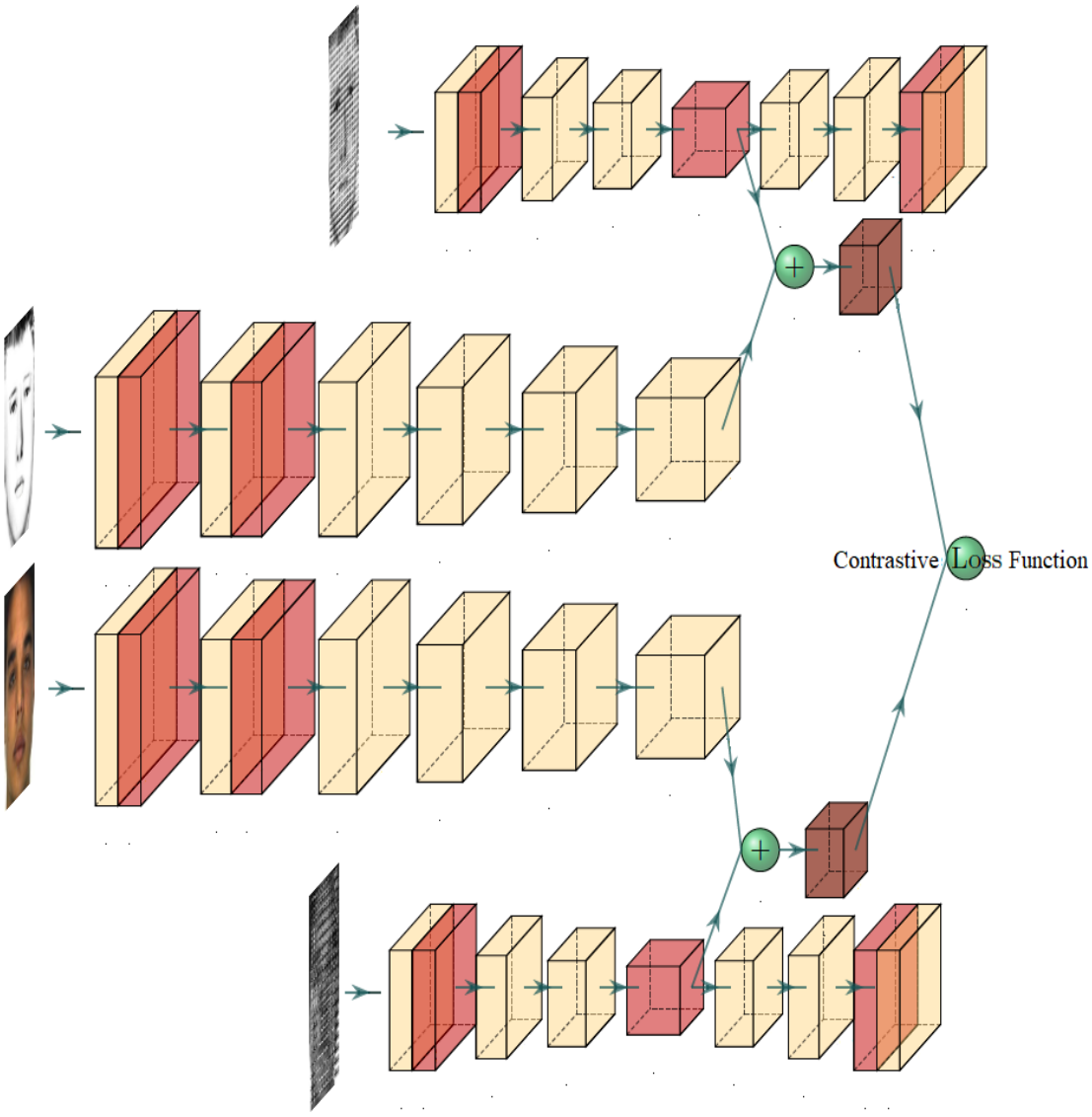


Figure 3-1 The architecture of the proposed Siamese network. The yellow blocks are convolution layers and the red ones are max-pooling layers. The brown block is the fused feature which includes features of the last convolution layer from each convolution channel and the feature from the hidden layer of the sparse auto-encoder-decoder. The input of the main network consists of the photo and the sketch of the face. The input of each sparse auto-encoder-decoder is the HAOG feature for the corresponding facial photo and the corresponding facial sketch.

However, these methods cannot reduce the influence of overfitting in this network. The deep neural network has a stronger expressive ability because of its complex structure than traditional machine learning methods have. The more complex the model, the more diversity features it can learn. In contrast, the deep neural network focuses on interpreting training data when the training data are insufficient. This leads to the overfitting phenomenon after a model is trained: the effect of training data is better than unknown data. Thus, the input data of the Siamese network derives from a pair of images which consist of two face images, one a photo and the other a sketch. The number of pairs of input image are $(N * E) * (N * E)$. N is the number of samples in each class, E the number of classes. Otherwise in the face sketch dataset, there is one facial photo and one sketch for each person. For the facial photo-sketch dataset (including the E class), there is $Image\ pairs = E^2$. Thus, the number of positive image pairs is $Positive_{image} = E * (N * N)$. If all the image pairs, positive and negative are input into the Siamese network, it causes the number of positive image pairs to be far less than the number of negative image pairs. The accuracy of recognition is affected by the imbalance input data, because the trained model has learned more about the distribution of the negative image pairs than of the positive image pairs. In order

to reduce the effect of imbalanced data, we select the negative image pairs randomly from all the different classes of image pairs. The probability ratio between positive image pairs and negative image pairs is 1:10. In general, the amount of trained datasets is not lower than the number of a model's parameters. For example, the CUFSF includes 1194 images as the largest face sketch dataset. After separating the dataset as training data and test data into 7:3, the positive image pairs and the negative image pairs in the training set are 835 and 8350, separately. The amount of input samples does not exceed 10000 image pairs. Hence, the number of training data is too small that there is a risk of overfitting.

To address the overfitting problem further, a convolution layer is used in the last layer instead of a fully connected layer. The memory of the convolution layer is smaller than that of the output, since shaping the output from a matrix into a column vector leads to a smaller number of redundant parameters than a fully connected layer would have. This further reduces the risk of overfitting. Although this method does not support learning 'distributed feature representation' for each sample from the hidden feature space, it can reduce the number of parameters and computations. Moreover, it supports a better way to select features in embedding space for recognition. Furthermore, L2 regulation, which as a penalty

function lowers the weight to reduce the complexity of the network, is used in weighting to avoid the risk of overfitting. The dropout is set as 0.2 before the last convolution layer. This diminishes the number of parameters through throwing units before connecting to the next layer of the neural network during training.

Due to the small size of facial photo-sketch datasets, the deep convolution network cannot extract enough effective and similar features from photos and sketches for recognition, A deep learning algorithm is needed to calculate probability of data distribution from the data. We propose to fuse Histogram of Averaged Oriented Gradients feature (HOAG) (Galoogahi and Sim, 2012b) to increase the accuracy of recognition. The HAOG feature utilizes the squared magnitudes to increase the angle from a histogram of averaged oriented gradients. It supports to extract weak and fine-grained features from face sketches in the feature extraction stage. Thus, this type of feature is provided as suitable for directly matching the images for face sketch recognition. Otherwise, HAOG feature use edge structure to describe the local features of images. This method reduces the effect of image rotation and image transfer by quantizing the location and orientation. The HOAG feature adopts a histogram for the local area of the image to reduce the effect of illumination change; this method focuses on

the image outline as the key feature. However, the HAOG method focuses more on using a gradient to describe the objective shape than on eliminating noise. Moreover, there are too many redundant features to help improve the recognition accuracy. We built two sparse auto-encoder networks to compress the HAOG features, which include a mass of redundant features to separately increase the calculation needed for the model. Each sparse auto-encoder network consists of six convolution layers and several max-pooling layers. For the encoder of our sparse auto-encoder-decoder network $h = f(x)$, the number of hidden nodes for hidden layers are less than the number of nodes for the input layer, to reduce the dimension of the original HAOG features. In order to learn from the input data what the remarkable feature is, we introduce a penalty function $\mathrm{L1}$ as a constraint in the encoder layer to reduce the extracted features' complexity.

$$L_1 = \sum_i |w_i| \text{(3-1)}$$

w_i is the difference of the extracted features from the images. Based on this characteristic, our network is built to learn the feature vectors by minimizing the discrepancy between the features extracted from the convolutional network and the original features which are generated using this feature descriptor.

In this sparse auto-encoder network, 'ReLU' is used as an activate function to reduce the difference between the encoder's output and the decoder's output. Moreover, the padding of each layer is all 'valid', which means that the output feature map needs to fill up '1' before sending the output to the next layer to keep the dimensionality between the encoder stage and the decoder stage unchanged. The structure of the sparse auto-encoder-decoder is shown in Figure 3-2. After training the sparse auto-encoder network, the features of the encoding layer are extracted as compressed features and these are fused with the features extracted from each channel of the Siamese network. The fused features will be used for training through the contrastive loss function.

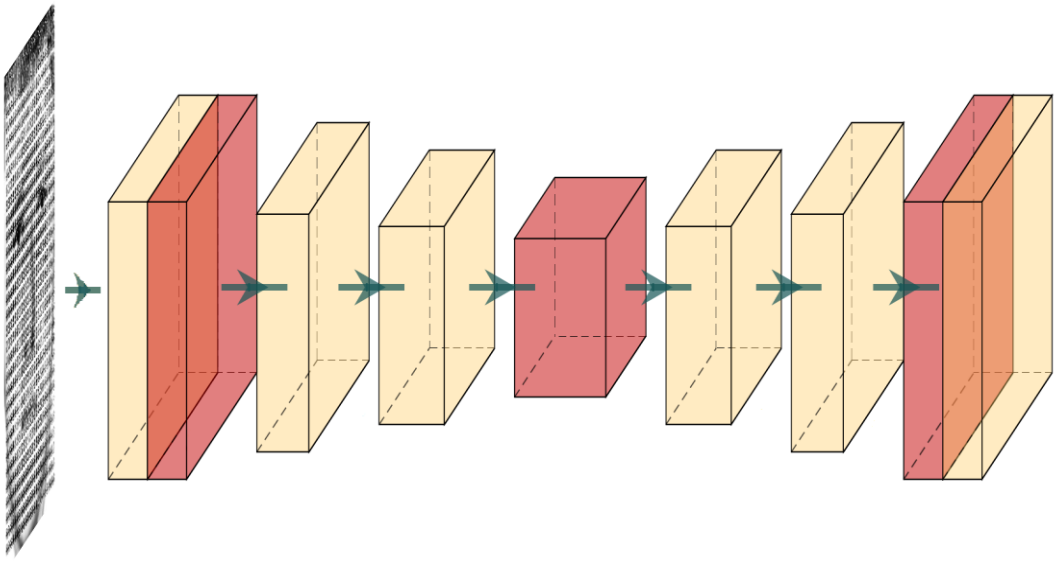


Figure 3-2 The detail of each sparse autoencoder-decoder

3.3. Loss functions

Loss functions are used to evaluate the discrepancy between the prediction and the real value, in order to optimize deep learning or machine learning models. The model's performance increases as the value of the loss function decreases. In our project, the loss function is to calculate the distance between a facial photo and a sketch when learning a mapping which projects different modality features into a common space. The aim is to separate the inter-modal sample and the intra-modal samples using a constraint condition. Three loss functions are used to train the network, for example, the contrastive loss function, Hinge loss function, and Cross-entropy loss function.

3.3.1.Contrastive loss function

The contrastive loss function is used to compare the similarities in a pair of images, which is defined as Formula (3-2):

$$L_D\left(W, \mathrm{Y}, \overrightarrow{X_1}, \overrightarrow{X_2}\right) = (1-Y)\frac{1}{2}(D_w)^2 + (Y)\frac{1}{2}(\max\{0, m - D_w\})^2 \qquad (3\text{-}2)$$

In this function, $\overrightarrow{X_1}$and $\overrightarrow{X_2}$ are the features of the images from the last layer of the two channels, based on convolution neural network. W is a set of parameters for a function which can keep invariance when the sample is map from a high

dimension to a low dimension. m is the margin to remove the unlike features of each image pairs. When the distance between the extracted features of the image pairs is large than the margin's value, the loss function as 0. It is used to represent the similarities in each image pair. Y is the label for X_1, X_2 samples. If the label Y is 0, it shows that X_1 and X_2 are the images of the same person. If the label Y is 1, X_1 and X_2 are considered as two different people. m is the threshold to divide the same sample and different samples for the input samples X_1, X_2. D_w is the Euclidean distance between X_1, X_2, $D_w = \|X_1 - X_2\|_2$.

According to the different contribution of each feature to the classification, the chi-square distance is combined with the sensitivity method in the contrastive loss function to calculate the weight of the feature between the two images in any image pair for training.

$$D_w = 0.5 \sum_1^n (Vec_{photo} - Vec_{sketch})^2 / (Vec_{photo} + Vec_{sketch}) \qquad (3\text{-}3)$$

Vec_{photo} and Vec_{sketch} are the feature vectors for a photo and a sketch, respectively. The Chi-square distance uses the contingency table method to analyse the difference between the data sets. Compared with the Euclidean distance method, the Chi-square distance represents the relative distance change

of each feature effectively using Features distance. It can effectively reflect the relative distance change of each feature quantity, especially the weight of the feature quantity under the chi-square distance which is calculated by using the sensitivity method according to the difference of the contribution of each feature quantity to the classification.

Because the range of the output distance is too large using the Chi-square distance, before comparing the similarities in an image pair, the output needs to be normalized to a value between 0 and 1. Then the normalized distance is used to compare with a margin value to determine the similarity of each pair. If the distance of a pair of images is within the margin value, the image pair represents the same person. A dissimilar image pair is without margin value. This loss function can be used to increase the distance between different people and to decrease the distance between images originating from the same person. The main idea of the Siamese network (Hadsell et al., 2006) is to separate samples of different classes sufficiently, based on the threshold of Euclidean distance. The threshold value which is set as 0.5 is defined by a number which is used to divide the dataset into positive pairs and negative pairs. However, for face sketch

datasets, due to certain factors such as high dimensionality of the feature vectors, there is no suitable value to ensure the similarity of the image pairs.

3.3.2.Hinge loss function

The Siamese network is trained for a multi-class classification task using the Hinge loss function to optimize this network. The Hinge loss function is described as Formula (3-4) below:

$$\mathrm{L_H} = \min\frac{\lambda}{2}||\omega||_2 + \sum_i \max(0,1 - \mathrm{Y}o_i^{net}) \qquad (3\text{-}4)$$

λ denotes the weight decay, ω is the weight of network. o_i^{net} is the output feature for the i^{th} sample, Y is the corresponding label for each input image's pair, the meaning is same with Formula (3-2). The Hinge loss function is used as part of the loss function which makes the distance close to a probability value. However, the Hinge loss is not differentiable at zero. The Squared L2-norm regularization, which is differentiable at zero, is used to alternate with the Hinge loss, in spite of its sparsity.

3.3.3.Cross-entropy loss function

The cross-entropy loss function, using parameters, estimates the probability between predictive value and real value. The function of the cross-entropy loss function is to maximize the log-likelihood function. The cross-entropy loss function is described as Formula (3-5) below:

$$L = \frac{1}{N}\sum_i [\mathrm{Y} \cdot \log(p_i) + (1 - \mathrm{Y}) \cdot \log(1 - p_i)] \qquad (3\text{-}5)$$

Y is the corresponding label for each input image's pair, the meaning is same with Formula (3-2). p_i is the positive probability for the i^{th} sample. Thus, this loss function performs well for optimized unbalanced samples in a multi-class classification. The output features from each convolution network are combined using a fully connected layer with a single output. However, the output dimension is too high. Dimensionality reduction is used to maintain the size of the features. In order to reduce the number of parameters, the fully connected layer that involves sigmoid as the activation function has its output size set to 4096. The sigmoid is used as an activation function, which maps the output features of the fully connected layer into the common space and measures the probability that

two image feature vectors resulting from the last layer are similar. Meanwhile, the sigmoid function increases the speed of updating the squared loss function weight.

3.4. Implementation and Experimental Results

Before generating the image pairs, a facial landmark detector (Kazemi and Sullivan, 2014) is chosen for use in face alignment in the pre-processing stage. This method determines the location of 68 specific points for each face photo and face sketch from the sparse subset of the pixels' grey values using an Ensemble of Regression Trees. This algorithm can discover a more precise position of the face sketches and face sketch attributions for our datasets than other face detection methods can, in order to align all the face images (in photos and sketches), the locations of the eyes are fixed at (100, 50) and (100, 100) after being translated, rotated and scaled. After using the facial landmark detector, all images are resized to 200*150. Due to overfitting as a result of the small size of the data, the recognition accuracy on the test set is lower than the one on the training set. In deep learning methods, we use the data augmentation method to increase the amount of data. For the existing face photo-sketch datasets the number of instances of each subject is limited. Each contains only one photo and one corresponding sketch of each subject. Moreover, each face sketch dataset is

small, none exceeding 400 pairs of images. However, some data augmentation methods are not suitable for face sketch recognition, since these rotations may produce unnecessary and negative noise which intensifies the complexity of the network. One reason is that sketch images, which consist of lines and shapes, are too simple to be used for extracting available features, given that the sketches of faces are less informative than the photos. The second reason is that some of data augmentation techniques such as vertical and titled rotation are not feasible for face recognition. A valid method is to generate more data using the existed face sketch dataset. In order to extract abundant features, we used a 3D morphable model (Bas et al., 2016b) to synthesis face images from a single image in different directions. This method uses image edges for face model fitting and synthesizes a 3D face model. Then the 2D face images are obtained after different directions of the 3D face model are obtained from different rotation angles. Despite the loss of edge information and hair space information, this method increases the number of instances for each subject. We choose four different poses, including rotated -30°, $\mathrm{rotated}\ -15^{\circ}$, rotated 15°, and rotated 30°, after synthesised a 3D face model from each face photo and face sketch. After

the data augmentation, each subject involves four generated images together with the original one.

Two resized images, each consisting of images of different modalities, are concatenated and input into the network as a single image. Each face photo was paired with one of the sketches to generate image pairs. If the photo and the sketch showed the same subject, it was labelled 0 as a positive pair. Otherwise, as a negative pair, it took the label 1. The number of positive pairs was far lower than the number of negative pairs. The data were then separated according to the subject of the input photo images; the percentages of the randomly sampled training data and test data were 80% and 20%, respectively. All instances were normalized to reduce the sensitivity and increase the speed of convergence.

Table 3-1 The hyper-parameters for each layer of the Siamese network (the padding of all the layers is the same.

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

As shown in Table 3-1, the basic Siamese network consists of five convolution layers and two max-pooling layers. In order to capture the features from sketch images, the kernel size of the second convolution layer was set to 7*7. The kernel size of the last three convolution layers was set to 5*5 to increase the nonlinear transformation. However, the number of parameters in the convolution layers was too large to avoid overfitting, so the max-pooling layer was added after each convolution layer. The max-pooling layer keeps the features from the largest filter after extracted features from several filters. It is not only to keep the edge and texture features from images, but also to reduce the dimension of the extracted

feature map. Thus, the amount of input data for the next layer using the max-pooling layer is less than without the max-pooling layer. Then the model was trained to use the Adam optimizer which adds bias-correction and momentum with a learning rate of 0.000006. The Adam optimizer performs better than the stochastic gradient descent and RMSProp optimizers. The weights were initialized randomly and a mini-batch was set as 125 for training. A gradient clip was appended in our model to avoid a gradient explosion. Several experiments certified that the gradient clip was set at 1.0. The other hypermeters kept default values, such as the exponential decay rate and epsilon.

Table 3-2 The hyper-parameters for each layer of each sparse auto-encoder. Except for the two last layers, the padding of the layers was set as valid.

Convolution1	Filter sizes: 3*3	Kernel size: 7*7	Activate function: ReLU
Down sampling	Pooling size: 7*7		
Convolution 2	Filter sizes: 3*3	Kernel size: 5*5	Activate function: ReLU
Convolution 3	Filter sizes: 3*3	Kernel size: 3*3	Activate function: ReLU
Encoding layer	Pooling size: 7*7		
Convolution 4	Filter sizes: 3*3	Kernel size: 3*3	Activate function: ReLU
Convolution layer5	Filter sizes: 7*7	Kernel size: 5*5	Activate function: ReLU
Upsampling	Pooling size: 7*7		
Decoding Layer	Filter sizes: 7*7	Kernel size: 7*7	Activate function: Sigmoid

The performance regarding the composite face sketch datasets, such as e-PRIP, PRIP-VSGC and Uom-SGFS datasets (Han et al., 2013) was evaluated. The models were trained using the three loss functions in turn and the results for different composite face sketch datasets were compared, as shown in Table 3-3 and Figure 3-3. In particular, while the Hinge loss function, cross-entropy loss function, and contrastive loss functions were used in our model, Table 3-3

indicates that the accuracy of contrastive loss with NN classification in Rank-10 was higher than 70% for most of the datasets and that the improved contrastive loss function obtained better performance than the other loss functions did.

Table 3-3 Recognition accuracy by different loss functions in Rank-10

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

We compared the performance with the ones in Kazemi et al. (2018), Mittal et al. (2015), and Galea and Farrugia (2018). In the Uom-SGFS datasets, the sketch was obtained by means of software. From the sketch dataset, the painter selected the image patch which resembled the suspect's face most closely to form a composite face sketch image. Although the features of the photographed face and the sketched face were highly similar, significant differences were found in

the automatic face identification system. The attributes for each face sketch formed a geometric mismatch with those of the corresponding photo, meaning that the recognition accuracy was lower than the state of art model.

Table 3-4 Recognition accuracy for e-PRIP datasets in Rank-10

Methods	Recognition accuracy
Improved Siamese network	85.33%
(Kazemi et al., 2018)	72.6%
(Mittal et al., 2015)	52.0%
(Galea and Farrugia, 2018)	54.9%

Table 3- 5 Recognition accuracy for Uom-SGFS datasets in Rank-10

Methods	Uom-SGFS(A)	Uom-SGFS(B)
Improved Siamese network	64.15%	81.74%
(Galea and Farrugia, 2018)	66.13%	82.67%

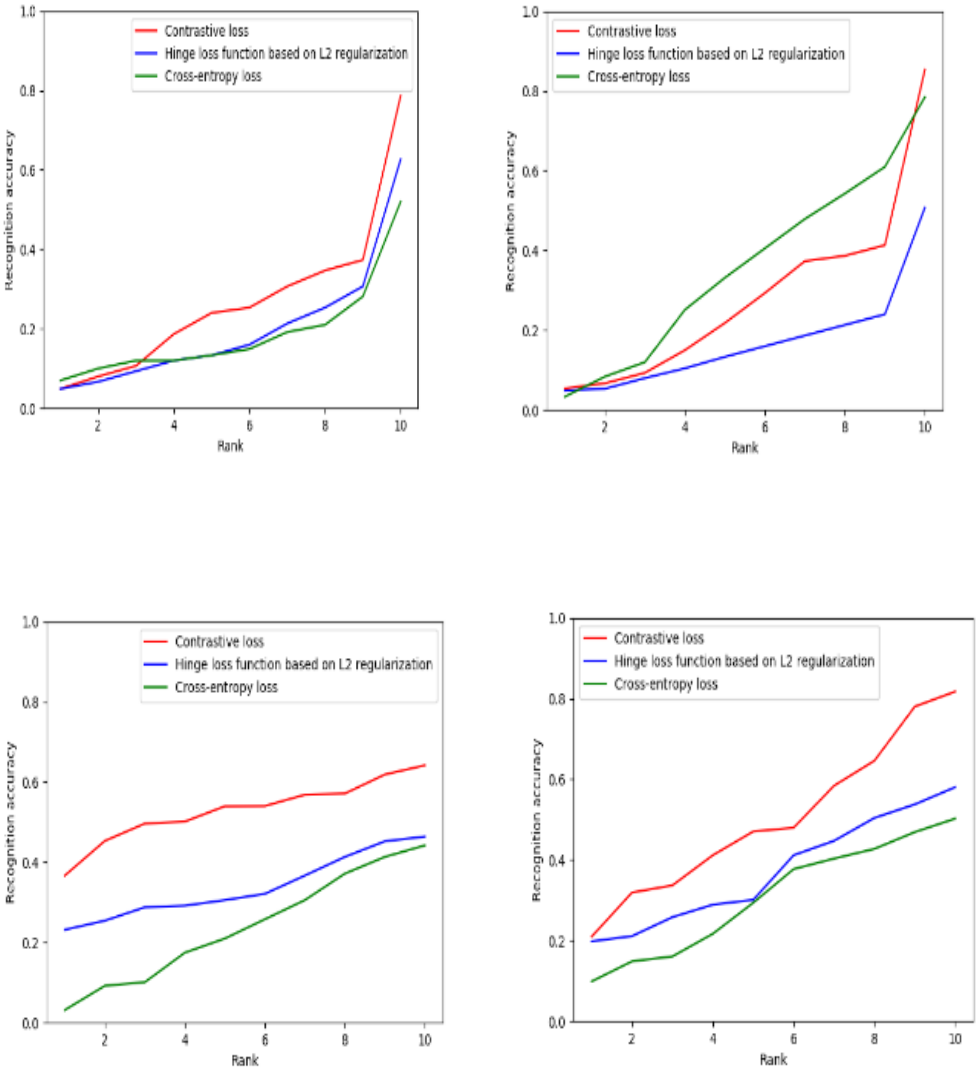


Figure 3-3 Recognition accuracy of proposed method using three loss functions for four datasets from Rank-1 to Rank-10 (Hinge loss function, Cross-entropy loss function and Improved contrastive loss function) (A) e-PRIP, (B) PRIP-VSGC (Indntikit), (C) Uom-SGFS (A) and (D) Uom-SGFS (B)

Table 3-6 Performance with (Galea and Farrugia, 2018) and (Mittal et al., 2015) for PRIP-VSGC dataset in Rank-10

Methods	Recognition accuracy
Improved Siamese network	78.67%
(Galea and Farrugia, 2018)	80.8%
(Mittal et al., 2015)	60.2%

Unlike the input of the single modal network and the original Siamese network, the input of the cross-mode attitude metric came from the two sample modalities, i.e., a photo and a sketch. The improved loss function eliminated the modal interference in the sample and mapped the distance metric for features in different modes, thus raising the level of accuracy. Moreover, the data augmentation and the regulation methods were used to increase the size of the dataset and reduce both the risk of overfitting and the complexity of the model. The experimental results of the proposed method showed that with the proposed method the accuracy rate on most datasets was higher than 70% in Rank-10. The recognition accuracy obtained using deep learning methods was sometimes no better than the accuracy of traditional methods. The cause of this phenomenon may be that important correlation information was not used in recognition, since the texture

features for face sketches were fewer than those for the corresponding facial photos. These texture feature for face photos make a kind of obstruction for cross-modal images. Moreover, deep learning methods generally need a large dataset for training to obtain a model that reflects the relationship between the photo and the sketch. However, face photo-sketch data sets tend to be too small to generate effective models. The Siamese network that was designed can share the parameters for training to extract similar features from photos and sketches of faces. After being combined with the features extracted from the last convolution layer for training the neural network, the max-pooling layer was used to reduce the number of features and to map the features into a common space. Because the data sample was not large enough, instead of the original Siamese network which was used as an end-to-end learning approach for both feature extraction and classification, we used traditional machine learning algorithms, namely, Support Vector Machine (SVM) (Shawe-Taylor and Cristianini, 2000), Random Forest (Liaw and Wiener, 2002) and XGBoost (Chen and Guestrin, 2016), in order to train classifiers for obtaining a higher rate of recognition than deep learning methods would have obtained.

3.5. Model's structure using classifiers

The structure of our proposed method is shown in Figure 4. Following the previous proposed Siamese network, we designed a neural network model that had the same parameters and would extract similar features to map the images into a new space. In the first step, the features of the photos and sketches were extracted using the Siamese network with shared parameters. However, the use of a deeper network had several negative effects. One was the gradient problem. When the value of a derivative is more than 1, it may lead to a gradient explosion in a deep neural network, or, in contrast, the gradient may disappear. Another disadvantage is that as the number of network layers increases, the greater depth and greater number of parameters increases its ability to fit beyond that of a shallow network. This means that the model is more complex and creates the problem of overfitting. The third is that the extracted features may increase the training error by losing the deep neural network. Thus, the Siamese network that we built was not very deep, and in consequence the features of the facial sketches were not rich enough to measure the distance with facial photos.

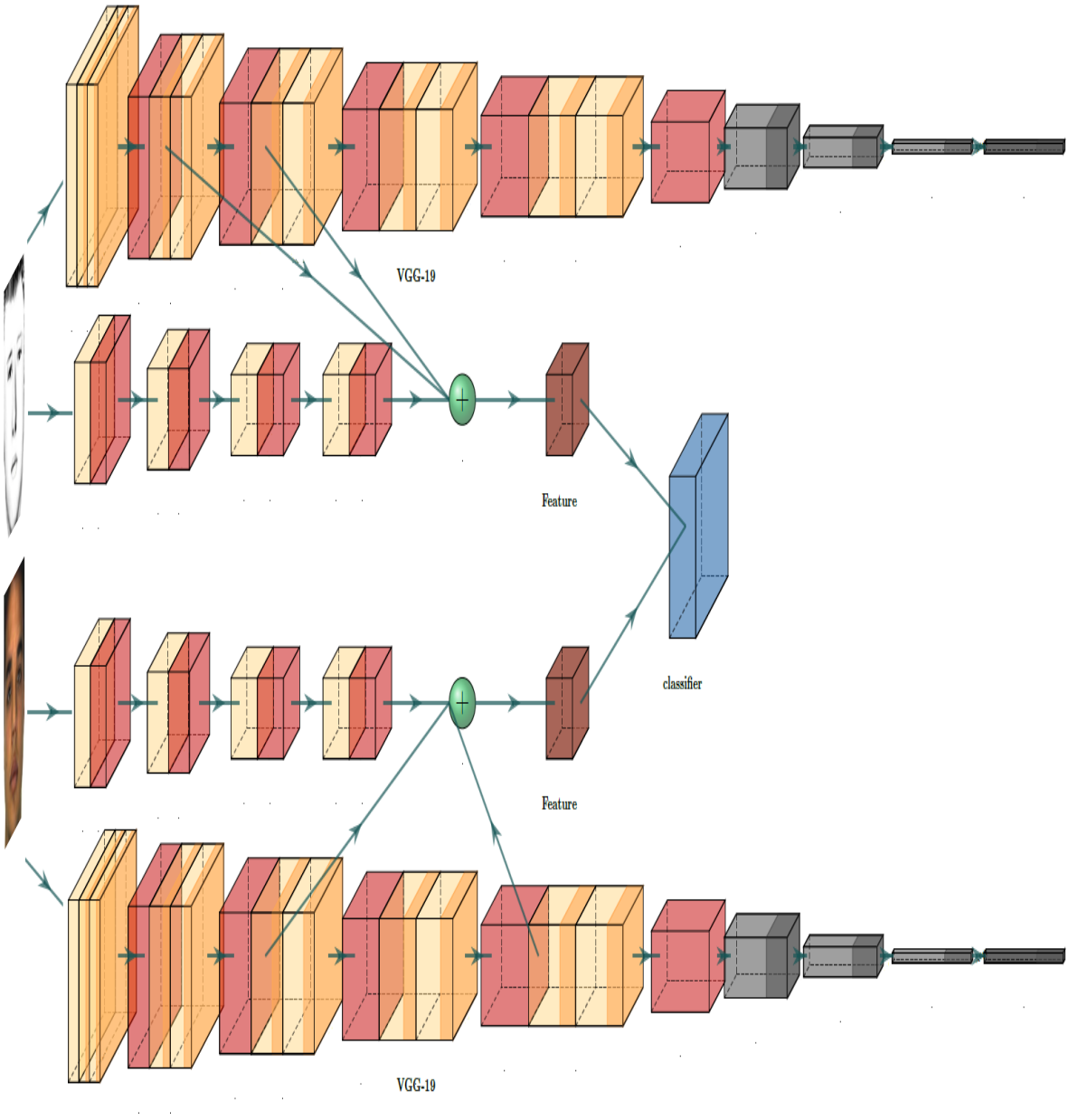


Figure 3-4 Overview of the proposed approach: The yellow block represents the convolution layers and the green block the max-pooling layers: (a) The features of similarity in the photos and sketches of faces and the features from the VGG-19 model are extracted from each model; (b) A max-pooling layer is used after merging the features; (c) Different classifiers are trained on the extracted features and evaluated on the accuracy of the recognition.

To improve the accuracy of recognition, the pre-trained model VGG-19 was used to extract features. Then the three types of feature were fused before inputting them into a max-pooling layer which was used to reduce the redundant features and the feature dimensionality. Finally, different classifiers were used on the extracted features and were evaluated for recognition accuracy. The 2-channel neural network which involves shared parameters was chosen as the basic model for extracting features from both photo images and sketch images. However, the features extracted from each channel that involved a CNN model could not be controlled, least of all the texture features extracted from sketch images, and as a result the CNN network could not yield more details. It also meant that the influence of the contrastive loss on face photo-sketch recognition was weaker than its influence on traditional face recognition. Thus, more features had to be extracted from photos and sketches, which required different feature extractors to enhance the recognition accuracy. Since the extracted features are different for different layers in the CNN model, we decided to extract more features from different convolution layers before training, to increasing the recognition rate by obtaining more diverse features. In the training stage, the aim of our network was similar to the aim of another Siamese network, which is used to extract more

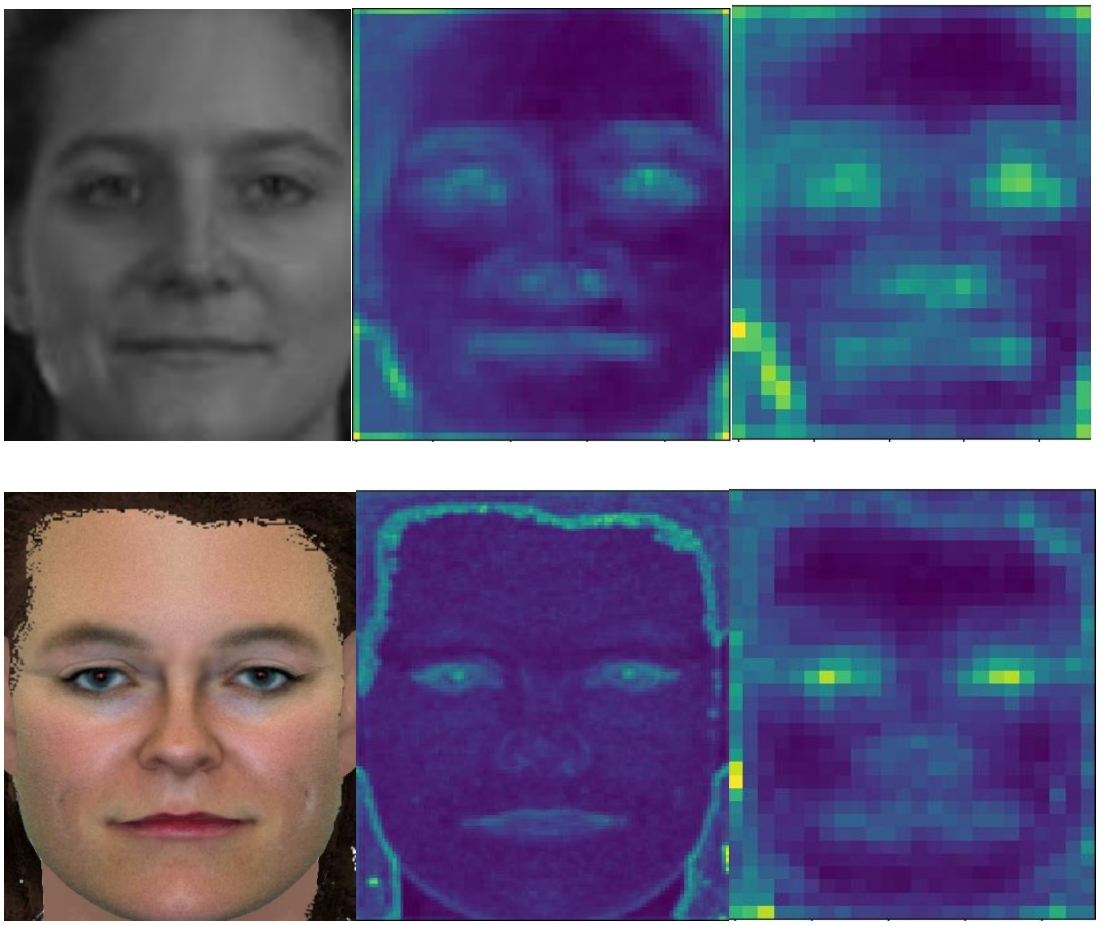
abundant features from cross-domain images for mapping into the same common space. The architecture consists of two identical channels that accept the distinct image as input and shared weights to extract features from facial photos and sketches, respectively. At the same time, related features are used to compare the distance between the face photo and the face sketch in order to reduce the disturbance from different modalities.

Because the number of instances was too small to train a reasonable model, for each channel in our Siamese network, we used the VGG-19 network as a pre-trained model to extract features from face images and sketch images. Instead of a 7*7 convolution kernel for a neural network in VGG-16 model, the VGG-19 model adopted a 3*3 convolution kernel to preserve the quality of the image features. Meanwhile, the deeper neural network was able to improve the effect of the neural network in the same perception field. The VGG-19 model as a pre-training model extracts more useful texture details and spatial features from the face photos and sketches that are deep enough to yield rich features. Thus, the performance of the VGG-19 model was better than that of the VGG-16 model. The features extracted from the VGG-19 model improved the recognition accuracy by being fused with the features obtained from the Siamese network

between face photos and sketches. To increase the quality of the features, we decided to extract two types of feature from the VGG-19 model: texture features and space information. We visualized the feature map that could be extracted using VGG-19 to obtain suitable features for our model. After visualizing the feature map, it was clear that Block2_conv2 showed more textures and directions than Block1_conv1 did. When we had processed the Block3_conv1 layer of the VGG-19 model, the performance as regards direction and colour begin to deteriorate, and more complex texture features appeared.

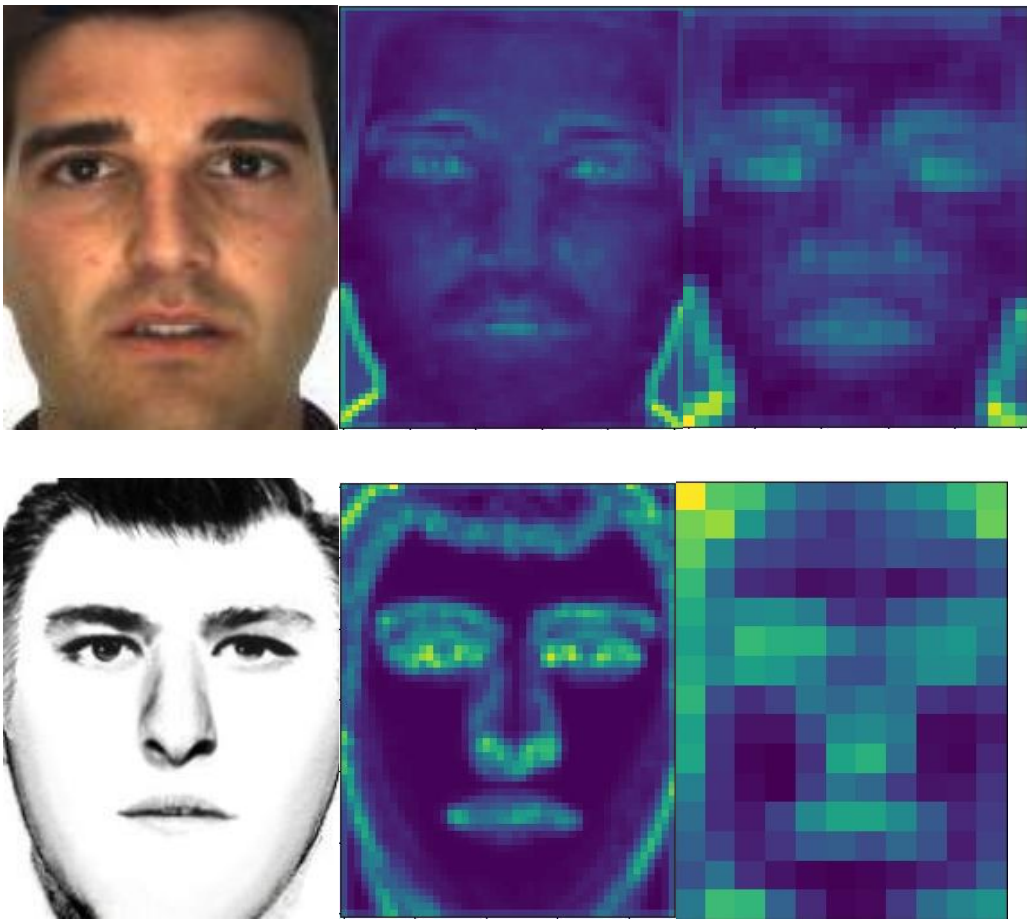
Figure3-5 and Figure3-6 show the images from the data set and the feature maps from the layers of the VGG-19 model in the Block2_pool and Block4_pool. The similarity between the feature map obtained from the Block2_pool layer and the input image is high, since the feature map shows more texture features. In addition, as the model gets deeper, more shape features are obtained from the Block2_pool layer than from the Block4_pool layer, and the extracted features contain more spatial information. Since large numbers of features exert a negative influence on the training classifiers, the features obtained from the Block2_pool and Block4_pool layers were extracted and then concatenated with the features which were obtained from each channel in the Siamese network. However, it is

not sufficient merely to reduce the number of features gained from using a small data set; it is also necessary to provide plenty of features for training classifiers to improve the recognition accuracy. A max-pooling layer was added separately to reduce the dimensionality before the features were fused. In order to avoid overfitting, it was set to obtain more features from the Block2_pool layer than from the Block4_pool layer. In each of these max-pooling layers, ReLU, which supports nonlinear mapping, was used to increase the iteration rate.



(a) (b) (c)

Figure 3-5 The Feature from different layers of VGG-19 model for Uom-SGFS (A) dataset. (a). The original face photos and sketches. (b). The photo feature from Block1_pool layer and sketch feature form Block2_pool layer. (c). The photo feature from Block2_poo layer and sketch feature form Block4_pool layer.



(a) (b) (c)

Figure 3-6 The Feature from different layers of VGG-19 model for Uom-SGFS (A) dataset. (a). The original face photos and sketches. (b). The photo feature from Block1_pool layer and sketch feature form Block2_pool layer. (c). The photo feature from Block2_poo layer and sketch feature form Block4_pool layer.

3.6. Experiment

In order to address the above-mentioned issue, we set the class weight for different pairs in the loss function to increase the influence of the positive pairs and also to keep the imbalance and increase the gradient stability. Figure 3-6 shows the heat-map images of the features which were extracted from the last layer of the Siamese network using the contrastive loss function. The weight ratio of the red area was higher than the weight ratio of the others. In the photos, the eyes are considered more important points than other areas from the last layer. Because the weight ratios after training are different for the different facial attributes in photos and in sketches, the recognition accuracy is still low, even if the effective feature was learned from the neural network.

In terms of the experimental setting, the performance obtained using the Adam optimizer, thanks to RMSProp, yielded generally better bias-correction and momentum than was obtained using the stochastic gradient descent optimizer or the RMSProp optimizer. The weight was initialized randomly, and the mini-batch size was set as 125 for training. A gradient clip was added in our model to avoid a gradient explosion. After several experiments, the gradient clip was set at 1.0. The other hyper-parameters kept their default values; for instance, the

exponential decay rate for the first-moment estimates and the exponential decay rate for the second-moment estimates and epsilon.

After an end-to-end Siamese network was trained using the contrastive loss function with chi-square distance, the similarity feature for each image pair was extracted for matching. After training the neural network, the output data from our model had the following characteristics: a) the data's dimensions were too high to classify; b) the negative data were several times greater than the positive data; c) the model's parameters were too large to avoid the overfitting problem. In consequence, we used traditional learning methods on the trained classifiers, namely, SVM, Random Forest and XGboost. In traditional machine learning, the class of new sketch can be estimated on the basis of Euclidean distance using a classifier trained on labelled data, which would overcome the limitations of deep learning methods on small datasets. That is to say. while the number of face sketch instances was too low, deep learning methods are generally not sufficient for training high-performance classifiers, although the size of the input data can be increased several times. Moreover, the unbalanced data may also lead to low recognition accuracy, because the distribution of the positive features is much lower than that of the negative features. Therefore, the above-mentioned three

traditional machine learning algorithms were used to train the classifiers on small datasets to avoid the above issues. SVM is capable of dealing with high dimensionality through its use of a kernel function for the transformation of the feature space. Otherwise, the high dimensionality of the data is also likely to affect the performance of most classifiers, because the generalization performance could be below what is required, leading to the incorrect classification of test instances that present different features. Hence, we used an optimizer to improve the generalization performance so as to increase the fitting ability of the classifier.

We compared the recognition accuracy obtained using three classifiers on the e-PRIP dataset and UoM datasets. The performance is shown in Tables 3-7 – 3-9. First, we tried to train a classifier using nonlinear SVM, which is suitable for improving the recognition accuracy of small data sets. The aim of SVM in our experimental setting, using a hyper-plane. For the features obtained from the Siamese network, was to identify whether the two images in a pair represent the same person. SVM can make a classification after mapping the features into high dimensional feature space. We used the 'RBF' kernel which reduces the complexity in high dimensional feature space to avoid having to measure the similarity between each sample in a new common space which was suitable for

classification by exploring the gradient as a parameter of SVM. Meanwhile, L1 was taken as a penalty term to reduce the sparsity of features in some sketches. Second, the random forest method was used to build an ensemble of base classifiers, and then the outputs obtained from the base classifiers were voted on to make the final classification. One advantage is that good performance can be obtained on high dimensional data without feature selection, i.e., the random forest method involves the effective self-evaluation of features. The other advantage is that two random values in a random forest increase the anti-noise capacity and avoid overfitting. Finally, the features extracted from sketches are simpler than those extracted from photos, especially for the e-PRIP dataset. For the sparse features of some facial sketches, XGboost used as an optimized boost method can lead to efficient training on the sparse feature space. The normalization in XGboost helps to reduce overfitting and increases the recognition rate.

Table 3-7 Recognition accuracy for classifiers on e-PRIP dataset at Rank-1

Methods	Recognition accuracy
(Galea and Farrugia, 2018)	54.9+3.2%
(Peng et al., 2019)	76.4%
(Saxena and Verbeek, 2016)	51.5%
Siamese net with SVM	77.8%
Siamese net with Random Forest	72.2%
Siamese net with XGboost	80.6%

Table 3-8 Recognition accuracy for classifiers on Uom-SGFS (A) dataset at Rank-1

Methods	Recognition accuracy
(Galea and Farrugia, 2018)	31.6%
(Peng et al., 2019)	64.80%
Siamese net with SVM	56.9%
Siamese net with Random Forest	65.3%
Siamese net with XGboost	63.9%

Table 3-9 Recognition accuracy for classifiers on Uom-SGFS (B) dataset at Rank-1

Methods	Recognition accuracy
(Galea and Farrugia, 2018)	52.17%
(Peng et al., 2019)	72.53%
Siamese net with SVM	82.3%
Siamese net with Random Forest	81.4%
Siamese net with XGboost	47.1%

Table 3-10 The precision for each classifier on different datasets

Methods	e-PRIP	UoM-A	UoM-B
Siamese net with SVM	51.42%	55.33%	70.27%
Siamese net with Random Forest	65.71%	63.88%	62.16%
Siamese net with XGboost	55.77%	74.28%	91.89%

Table 3-7 shows the performance of various classifiers on the e-PRIP dataset. The sketches in this dataset are generated by line using the software. Thus, the features extracted from sketches are simpler than those extracted from the corresponding photo. On this dataset, our proposed method involved the Siamese

network with XGboost and gained 61.1% accuracy, but the performance was lower than the one reported in Peng et al. (2019), which obtained 76.4% accuracy. Tables 3-8 and 3-9 show the recognition accuracy on the UoM-A dataset and the UoM-B dataset. On the UoM-A dataset, the recognition accuracy obtained with our method, which involved the Siamese network with Random Forest was 0.5% higher than the performance reported in Peng et al. (2019). For the UoM-B dataset, our method, which involved the Siamese network with XGboost, obtained 80% recognition accuracy, which was better than the other methods. Due to the data imbalance, precision was able to be more effective than recognition accuracy in evaluating the recognition performance. In general, precision is the ratio of correctly predicted positive values to the total predicted positive values. This metric highlights the correct positive predictions among all the positive predictions. Table 3-10 shows the precision obtained in performance using SVM, Random Forest and XGboost. The method that involves the combination of the Siamese network and XGboost shows high performance without using the e-PRIP dataset, especially from the UoM-B dataset on which the precision exceeds 90%. For all the datasets, the method that involves SVM shows the lowest precision. Overall, using XGboost as a classifier gives a better performance than the other

algorithms do. The method that uses SVM can avoid the complexity resulting from the high dimensionality by using a kernel function on the feature space. For recognition accuracy and precision, the performance obtained using SVM is generally worse than the one obtained using Random Forest or XGboost. In the construction process, it should be ensured that the trained base classifiers are well diversified, which requires new samples to be drawn randomly from the original training data, so that diverse base classifiers can be trained on different samples creating a good chance of improving the performance. For the line sketch, the method that involves using Random Forest shows better performance than any obtained using the other learning methods, which indicates that the Random Forest method is efficient enough to ensure that the trained base classifiers are diverse, due to the random sampling of instances and features to form multiple diverse training samples and feature sub-spaces.

3.7. Conclusion

In Chapter 3, we designed a cross modalities Siamese network to match different modality images. First, we designed a Siamese network which ensured that the input of the cross modalities attitude metric would come from the sample of two modalities. The improved loss function eliminates the modal interference in the

sample and maps the distance metric for features in different modes to increase the level of accuracy. Moreover, the data augmentation and regulation methods are used to increase the size of the dataset and reduce both the risk of overfitting and the complexity of the model. The experimental results show that using the proposed method raises the accuracy on most datasets to higher than 70% in Rank-10. Although the recognition accuracy is high, however, the use of the contrastive loss function does not lead to improved classification performance. In the next step, we focus on extracting spatial information from images to reduce the distance of features between face photos and face sketches after mapping both of these into a common space. Thus, we build a new Siamese network that increase the recognition accuracy; it involves parameter sharing and is combined with VGG-19 as a pre-trained model to extract similar features, using the contrastive loss function to reduce the data imbalance. Based on these extracted features, multiple classifiers are trained to improve recognition accuracy. Like the original Siamese network, our designed network extracts similar features from each pair of images using the shared parameters. One reason is that the parameters for each channel are used to extract the same types of feature. At the

same time, the parameters are used to extract useful features for different types of data.

Our experiment used traditional learning methods for training classifiers to increase the robustness of face photo-sketch recognition. We explored the performance obtained by adopting three traditional learning algorithms (Support Vector Machine, Random Forest and XGBoost) combined with the Siamese network for training classifiers, based on the features extracted using the Siamese network and other features obtained from the use of the pre-trained VGG-19 model. Our methods showed high recognition accuracy on the e-PRIP and UoM datasets. Especially when using XGboost, it performed well in terms of both precision and accuracy.

Chapter4:

Attention-Modulated Triplet Network for Face Sketch Recognition

4.1. Introduction

In Chapter 3, we described our improvements to the shared-weight Siamese networks to increase the similarity of the features in facial photos to those in facial sketches for the sake of recognition. One improvement was to combine the HAOG feature after reducing its dimensions, using the autoencoder-decoder model. Next, we extracted some channel features from the VGG-19 model to focus on the features of the facial photo and facial sketch that were similarly located. As shown in the feature maps after the activation function (see Figures 3-5 and 3-6, showing the sketched image), the extraction focused on the information about features on the edges. The extracted features of the photo image tend mostly to be those features that are facial attributes. One reason for the low accuracy of the recognition is that the Siamese network measures the similarity by the score of the features' distance. The feature's dimension affects the score's accuracy. In our Siamese network, in addition to the matched object, the score of other similar objects is too high to match. It is difficult to match features in the facial photo to those in the facial sketch, using the same parameters after training. Following this, we analysed the human visual system, which uses attention mechanisms for quickly screening out high-value information, such as anything related to the task

area, from huge sets of information. For example, when we see another person, we focus on the shape of his or her face and then combine the information from different regions to form an overall impression of him or her. This means that the distribution of attention in each spatial position is different for an object and for a scene. This is why we designed a model to simulate the human visual mechanism, called the attention block. The goal of the attention block is to select key information for the current task. It needs to pay more attention to this area while restraining other information to obtain more details of the target.

The contribution of the present research is to build a triplet network combined with an attention module and a spatial pyramid pooling layer, with the aim of distinguishing different classes of image and, after comparing the distance between the features of each attention module, to identify when the same class of images is involved. The attention module is used to learn the related features in similar locations from cross-modality images. It consists of two attention blocks: the channel attention block acts on both facial photos and sketches to generate the feature maps, and the two spatial attention blocks act on photos and sketches to focus on the location of the facial features. Both spatial blocks share the same structure, and the block related to the photos is trained first because large training

datasets are available. Then the spatial attention block for sketches is trained using fine-tuning, together with a smaller sketch training dataset adapted to the photo attention block. The experiments show that implementing this method achieves better results than the state-of-the-art results with composite facial photo-sketch recognition.

The contributions of this chapter are as follows:

1, We developed a triplet network combined with an attention module and a spatial pyramid pooling layer. The parameters of each channel were shared to generate the same encoding rules for extracting features before the attention module.

2, We designed an attention module which consisted of a channel attention block and a spatial attention block. The spatial attention block focused on extracting similarly shaped features from different modalities of images (photographed and sketched).

3, The spatial pyramid pooling layer (SPP layer) was introduced to reduce the effect of image noise and deal with input images of different sizes.

Attention-Modulated Triplet Network

4.2. The attention Triplet Network

Our facial photo-sketch recognition system has three parts. The first part is a triplet network. The second is an attention network which is introduced to extract similar feature vectors from both the photo and the sketched images. The third is a spatial pyramid pooling (SPP) layer used to prevent information loss due to the fixed size of the input images. The proposed triplet network consists of three branches of neural networks, as shown in Figure 4-1. In our triplet network, two of the input images are the sketch and the photo images of the same person, and the third input is the face sketch of a different person. Each image was input into a channel to extract the edge features from a shallow convolution layer and texture features from a deep convolution layer.

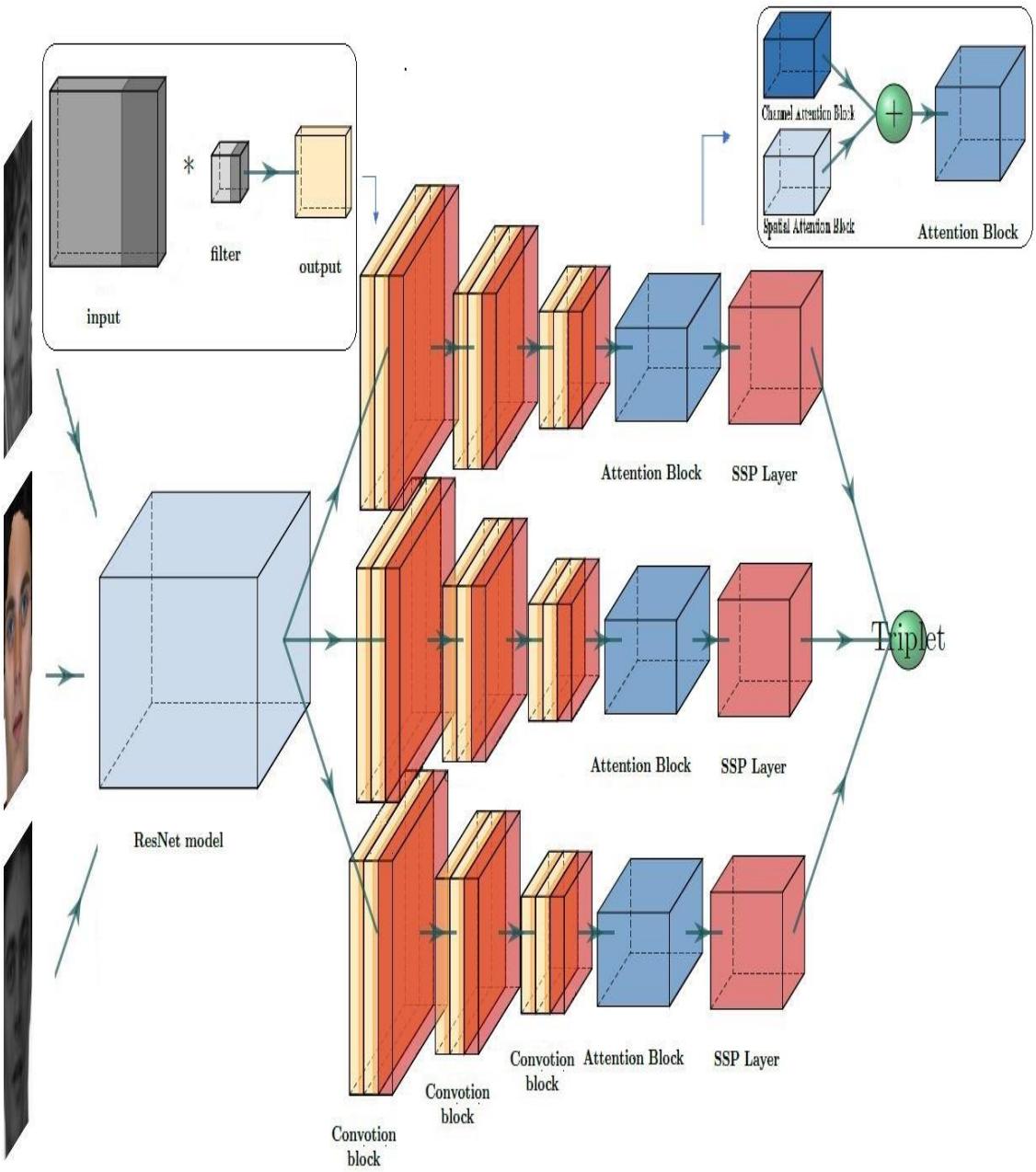


Figure 4-1 The structure of the triplet network and details of each channel. The features are extracted from the fourth convolution block from the ResNet model. The structure of each convolution block consists of two convolution layers (yellow) and a max-pooling layer (red). The kernel size of each convolution layer is 3*3, and the stride size is 2. The input image size is $n * n * 3$.

Each channel was composed of the same convolution neural network. The convolution neural network was constructed from three convolution blocks, an attention block, and an SPP layer. Each convolution block included one convolution layer and two max-pooling layers. We adopted a 7*7 size convolution kernel as an image filter on each input image. A 7*7 size convolution kernel not only reduces the number of parameters, but also reduces the space complexity. The space complexity is calculated by

$$space \sim O\left(\sum_{l=1}^{D} 7^2 \cdot C_{l-1} \cdot C_l + \sum_{l=1}^{D} M_l^2 \cdot C_l\right) \qquad (4\text{-}1)$$

For the l^{th} convolution layer, 7 was the kernel size r, D was the number of convolution layers, M was the size of the output feature map of each convolution layer and C_l was the size of the l^{th} convolution layer.

The process of making a convolution layer for an input image is as follows:

1. Features are extracted from the selected image using the convolution kernel as a filter.

2 Each value from the convolution kernel is multiplied by the corresponding values in the image.

3. All the product results are added together.

4. A feature map is generated after step 3.

The size of the feature map extracted from a convolution layer is calculated by

$$w_{n+1} = \frac{w_n + 2*p - k}{s} + 1 \qquad (4\text{-}2)$$

In (4-2), w is the width of an image, k is kernel size, s is the stride, and p is padding on the input image to resolve the problem of information loss at the edge of the image. After each convolution layer is completed, the size of the feature map, w_{n+1}, is rounded down if w_{n+1} is not an integer.

The two max-pooling layers for each convolution layer were used in the output feature map. The precise position of this efficient feature in an image is far less important than its position in relation to the other features. The max-pooling layer divides the input image into several rectangular sub-regions and calculates the maximum value for each of them. The purpose is not only to reduce the space needed for the data, but also to decrease the number of the model parameters to avoid overfitting. After creating max-pooling layer, the dimension of this feature map is calculated by:

$$\mathrm{Dim}(H_n, w_n, D_n) = (\frac{(H_{n-1}-\mathrm{k})}{(Z_{n-1}+1)}, \frac{(W_{n-1}-\mathrm{k})}{(Z_{n-1}+1)}, D_n) \qquad (4\text{-}3)$$

In (4-3), H_n, w_n is the width and height of the output in the last layer. D_n is the number of convolution kernels. z_{n-1} is the stride for the max-pooling layers; in our model, it is set as 2.

The image feature maps were extracted from each channel and used in a triplet loss function to minimize the feature differences between the pairs of images of the same person and maximize those between different persons. In the structure of the typical triplet network, each feedforward neural network maps images into an embedding space. The output of our triplet network was the L-2 distance between each positive sample and negative sample. The ideal solution for sample selection is:

$$argmin_{positive} \left\| Net\left(anchor_{sample}\right) - Net\left(positive_{sample}\right) \right\|_2^2$$

$$argmax_{negative} \left\| Net\left(anchor_{sample}\right) - Net\left(Negative_{sample}\right) \right\|_2^2$$

The distance ($distance_{same}$) between facial photo ($positive_{sample}$) and the corresponding sketch ($anchor_{sample}$) is:

$$distance_{se} = \frac{\left\|Net(anchor_{sample}) - Net(positive_{sample})\right\|^2}{\left\|Net(anchor_{sample}) - Net(positive_{sample})\right\|^2 + \left\|Net(anchor_{sample}) - Net(negative_{sample})\right\|^2}$$

The distance ($distance_{different}$) between the facial sketch ($anchor_{sample}$) and a different sketch ($negative_{sample}$) is:

$$distance_{different} = \frac{\left\|Net(anchor_{sample}) - Net(negative_{sample})\right\|^2}{\left\|Net(anchor_{sample}) - Net(positive_{sample})\right\|^2 + \left\|Net(anchor_{sample}) - Net(negative_{sample})\right\|^2}$$

After optimized using triplet loss function as:

$$L = \sum_{1}^{n} \left[\left\|Net(anchor_{sample}) - Net(positive_{sample})\right\|_2^2 - \left\|Net(anchor_{sample}) - Net(negative_{sample})\right\|_2^2\right]$$

The best effect of the triplet loss function is used to reduce the distance between the same samples and increase the distance between different samples as:

$$distance_{same} \to 0, \ distance_{different} \to \infty$$

Then the distance between each sample is as follows:

$$\left\|Net(anchor_{sample}) - Net(positive_{sample})\right\|_2^2 < \left\|Net(anchor_{sample}) - Net(negative_{sample})\right\|_2^2$$

$$\left\|Net\left(anchor_{sample}\right) - Net\left(positive_{sample}\right)\right\|_2^2 + \alpha$$
$$< \left\|Net\left(anchor_{sample}\right) - Net\left(negative_{sample}\right)\right\|_2^2$$

$negative_{sample}$ is the face sketch of a different person, $positive_{sample}$ is a face photo of the same person, i.e., the facial sketch of the same person as the input photo, and $negative_{sample}$ is a photo of different person's face images. We used the chi-squared distance to measure the feature differences between the images.

4.3. The attention network

To reduce the number of parameters, we used max-pooling layers after each convolutional layer. However, pooling layers lose information and also ignores the relationship between whole images and local regions. Meanwhile, the features extracted from photos and sketches using the trained shared-weight network are different for each facial attribute, and the recognition rate is still low. Thus, after assigning different weights for different parts using linear weighting methods, our attention module was designed as a set of neural network blocks that would highlight our targets for attention, as parts of the input based on the relationship between each pixel in each image. This highlighting would extract more information about the target from the selected attention regions and screen out

unwanted information. The main idea with so much information was to focus on the parts that would yield more useful information for the current task. There are two main tasks before using the attention mechanism to improve the efficiency and accuracy of task processing. One is to decide the important part of the input image to focus on and reduce the attention that might be given to other information, even filtering it out altogether. The other is to solve the problem of information overload by using the limited information available to determine the important parts of the input. Thus, we proposed to extract features from effective regions of the images to increase the recognition accuracy. The attention mechanism is used in each channel of the triplet network.

One part of the attention module is designed to ascertain the relationship between the images of each channel and focus on extracting the shape of the input images; the other focuses on extracting spatial information and texture features from the channel attention layer. The proposed attention module consists of a channel block and a separate spatial block.

First of all, the spatial attention in the spatial domain was controlled by treating the image features in each channel equally without the information in the channel domain. This approach limited the transformation method of spatial domain to the

extraction of the original image features. However, the features extracted from the spatial attention block cannot be interpreted when it is applied to other layers of the neural network layer. After using spatial attention blocks, the attention block for this channel extracted the image information from the pooling layer in one channel and ignored the local information in other channels.

The convolution operations produced a local receptive field. The features corresponding to the pixels with the same location in the facial photo and facial sketch may be slightly different. Such differences introduce inconsistencies between the intra-class images and the inter-class images. Our method, however, was able to adapt by focusing on the features in the same positions in the facial photo and the facial sketch, using the attention module to enhance the representation of the features for recognition. The attention module in our network paid more attention to the important parts of the images, which was useful for matching the images that were photos and those that were sketched. This model included a channel attention block and a spatial attention block in order to extract the edge features and texture features from the input images. Finally, the feature map from the spatial pyramid pooling layer was fed into a fully connected layer with L2 normalization. The input of the attention module was the feature map

F_{conv3} , which was extracted from the third convolution layer. The attention feature map which concatenated the channel attention feature map $F_{channel}$ and the spatial attention feature map $F_{spatial}$ was as follows:

$$F_{attention} = [F_{channel}, F_{spatial}] \quad (4\text{-}4)$$

4.3.1.The channel attention network

As illustrated in Figure 4-2, the channel attention of photo images uses the intra-channel relationship between the features extracted from the convolution layer to represent meaningful features for recognition. The channel attention map is computed as:

$$F_{edge} = GloPool(conv(conv(F_{conv3}))) \quad (4\text{-}5)$$

$$F_{texture} = MaxPool(conv(conv(F_{conv3}))) \quad (4\text{-}6)$$

$$F_{channel} = Sigmoid(F_{edge} \times F_{texture}) \quad (4\text{-}7)$$

To compute the edge features, F_{edge}, we used global average pooling to keep the edge information on the feature map. We fed F_{conv3} into two convolution layers followed by a global average pooling layer to get a feature vector F_{edge}.

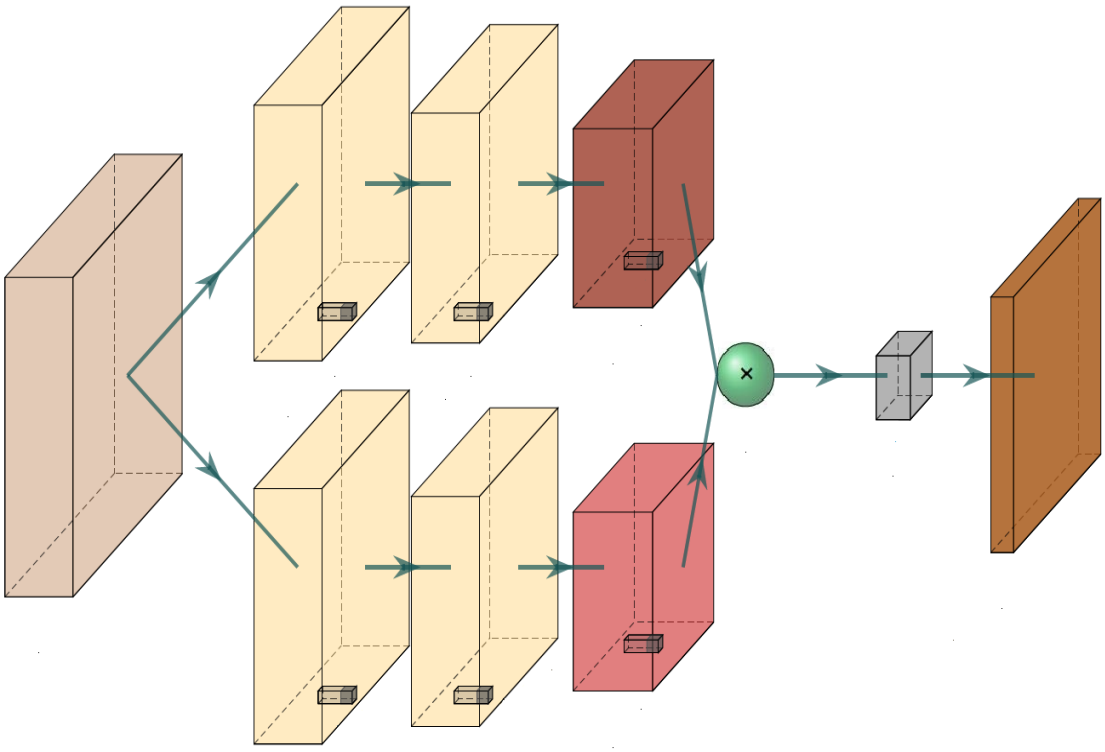


Figure 4-2 The structure of the channel attention module. The channel attention module consists of two branches. Except for the last layer of each channel, two convolution layers (yellow) and one pooling layer are included in the module. One is globalPooling layer (crimson), and red is maxPooling layer. The kernel size is $1 * 1$ for each convolution layer and pooling layer.

The output features from the two convolution layers represented the weights for each pixel on the last feature map and the information from each channel was separated, by means of a convolution kernel, into information components. Otherwise, a convolution kernel of size $n * n$ can be activated on the last output feature map $W * H$. The generated feature map is composed of (W, H). W and H are respectively the width and height of the last feature map. Since the input and output of the convolution has only $1 * n$ dimension, it does not consider the relationship between the pixels and the surrounding pixels. However, our feature

map which was input into the channel attention block included 128 channels, and the 1 × 1 convolution kernel was activated on each pixel on different channels to fuse the information. At the same time, the 1 × 1 convolution kernel can change the number of output channels to reduce the output dimensions. This not only kept the planar structure of the last image, but also increased the nonlinear character of our model after increasing the depth of the network using a 1 × 1 convolution kernel.

Meanwhile, an adaptive max-pooling layer was used to extract the texture feature $F_{texture}$ with F_{conv3} passing through two identical convolution layers. Because the contributions of key items of information are different on the output feature map from this max-pooling layer we assigned different weights to each channel to represent the correlation between the channel and the useful information. Then we multiplied the matrix element-wise F_{edge} by $F_{texture}$ to form an integrated channel feature $F_{channel}$ which contained both the edge and textural features. Finally, we used an ELU activation function to obtain a channel feature without image noise.

4.3.2.The spatial attention block for photo image

Unlike a channel attention module, a spatial attention module works on the position of a picture. The spatial attention module transfers the information about the image to a new space and keeps key information which can increase the accuracy of recognition. In a traditional convolution layer, a pooling layer, such as max-pooling and average-pooling, can compress the size of the feature map to reduce the computing time when the receptive field is large. The next convolution layer can receive more information after the pooling layer. However, a pooling layer loses some information and this can reduce recognition accuracy. Moreover, combining information directly by means of a pooling layer can make key information unrecognizable. In order to extract the key information from images and improve the recognition, we tried to transform the corresponding spatial domain information to extract its more important features, which can represent the relationships between a sketch and a photo.

We used two types of spatial module to act either on the photo image or the sketch image, as shown in Figure 4-3.

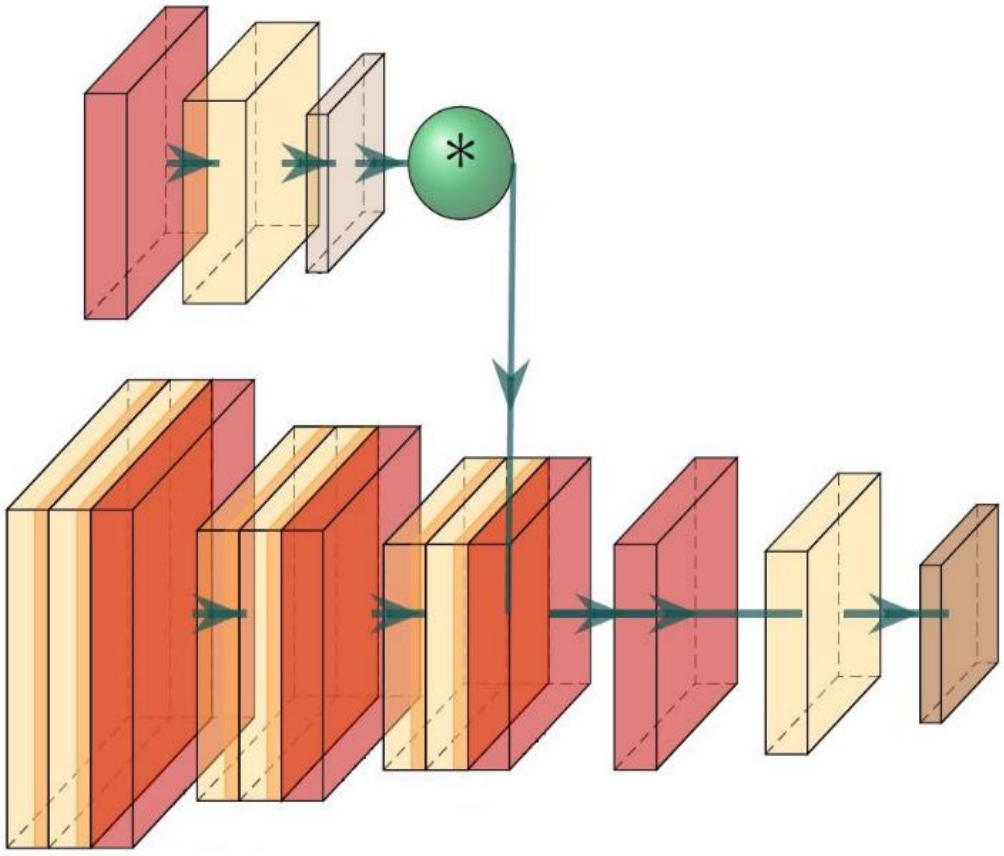


Figure 4-3 The structure of the spatial attention module for the photo and sketch images. The spatial attention module (top part and the last three layer in bottom part) consists of a stochastic pooling layer (red) and a convolution layer (yellow). The kernel size of the stochastic pooling layer is $2 * 2$ (stride is 1) and the kernel size of the convolution layers $3 * 3$ with padding size 1. Following the convolution layers, the sigmoid function (brown) is used as an activation function.

The spatial modules adopted the stochastic pooling layer (Zeiler and Fergus, 2013) combined with the overlapping pooling strategy (Prabhu and Pe'Er, 2009). The pooling map of the stochastic pooling layer had a probability p_i for each

pooling region j that was computed by a multinomial distribution after nominalizing the feature map a_i as follows:

$$p_i = \frac{a_i}{\Sigma_{k \in R_j} a_k} \qquad (4\text{-}8)$$

The sample was selected randomly with the multinomial distribution p_i from each pooling window. Unlike classic pooling strategy, this pooling strategy exploited the overlapping regions between the adjacent pooling windows to avoid loss of information from the input feature map. This method fused the multi-level features using sparse sampling to increase the robustness of the target deformation. The principle of this strategy dictates that the stride s of each filter window must be smaller than z, the size of the pooling window in the stochastic pooling layer. We obtained a pooling map F_{pool_photo} from the feature map F_{conv3_pool} which was extracted from the third convolution layer. One advantage of this method is that the edge information and texture information are extracted without distortion. Another advantage is that the image noise is reduced. Then the F_{pool_photo} was fed into a convolution layer to generate an attention map $F_{spatial\_photo}$ which was computed as:

$$F_{spatial\_photo} = Conv(F_{pool\_photo}) \qquad (4\text{-}9)$$

$$F_{pool\_photo} = [AvgPool(F_{conv3\_pool}), StochPool(F_{conv3})] \quad (4\text{-}10)$$

where [□,□] means concatenation of two feature vectors.

4.3.3.The spatial attention block for sketch image

The triplet network was designed to use shared weights to extract similar features from each channel of the network. However, the convolution layer ignored the inter-relationship from each patch of the sketch. We used the edge feature vector of the photo, which was generated from the photo spatial attention module, to focus on the same position and extract more similar features from the channel attention $F_{channel\_sketch}$ of the sketch. The structure of the spatial attention module for the sketch was as follows:

$$F_{point\_sketch} = F_{spatial\_photo} \times F_{conv3\_sketch} \quad (4\text{-}11)$$

where × denoted the elementwise multiplication between the photo spatial attention and the feature map of the sketch, so as to extract the features that were in similar positions in the two images. The sketch edge feature vector $F_{spatial\_sketch}$ was extracted using the same method as the photo spatial attention layer, to give the image more weight. Next, a spatial feature map for a sketch image was obtained after calculating the correlation using a sigmoid layer.

$$F_{spatial\_sketch} = \sigma(Conv(F_{pool\_sketch})) \quad (4\text{-}12)$$

$$F_{pool\_sketch} = [AvgPool(F_{point\_sketch}), StochPool(F_{point\_sketch})] \quad (4\text{-}13)$$

4.4. The spatial pyramid pooling

To handle photos and sketches of arbitrary sizes, we used a spatial pyramid pooling layer (He et al., 2015) after the attention module, instead of the fully connected layer in the original triplet network. Spatial pyramid pooling allows not only inputs of arbitrary aspect ratios, but also arbitrary scales. One reason for using spatial pyramid pooling was that it avoided information loss from cropping. The second was that different features could be extracted from the feature map of the attention module to increase the robustness of our method. The input feature map after the attention module was divided into N sub-windows of fixed size, and average pooling was applied on each sub-window. After this pooling, the dimensions of the feature maps for all the sub-windows were the same. The output feature map was composed of the feature maps from all the sub-windows.

4.5. Experiments

4.5.1.Pre-process method

We used the MTCNN network (Zhang et al., 2016) to extract the location of the facial image, in order to reduce the effect on recognition accuracy of such factors as position and occlusion. In the first step, all the facial photos and sketches were input into the P-Net model. P-Net is a fully convolutional network in the MTCNN model which describes the target location determined by the bounding box (BBox).

The bounding box is a rectangular box that is specified by the x and y axis coordinates in the upper-left corner and the x and y axis coordinates in the lower-right corner of the rectangle. First, the input image is divided into many sub-regions. Second, the regions are merged according to the similarities between these sub-regions. Finally, the adjacent regions are continuously merged as new candidate frames of different scales. Thus, many overlapping candidate windows are generated for the same objective. To eliminate overlapping, an NMS algorithm searches for local maxima and suppresses non-maximum elements. Bounding boxes save the optimal candidate window which has the highest confidence score. The process of the NMS algorithm is as follows:

1. Calculate the confidence score for all BBoxes after a P-Net, and sort them by their confidence score.

2. Select the highest confident score bounding box and remove it from the BBox list, after adding it into the final output BBox list.

3. Remove each BBox from the BBox list if the overlap area (IOU) with the current highest score box is greater than a certain threshold.

4. Repeat from step 1 until the BBox list is empty.

After the P-Net model with NMS algorithm reaches optimization, all the bounding boxes and facial landmark points from the P-Net model are selected as training data into R-Net part of the MTCNN model to optimize. R-Net uses a fully connected layer to retain more image features and filter out the many candidate windows that perform badly. Then, for a set of candidate windows $P = (P_x, P_y, P_w, P_h)$, a mapping f is calculated in order to choose a region proposal which is close to the ground truth objective followed: $f(P_x, P_y, P_w, P_h) \approx (G_x, G_y, G_w, G_h)$. We adopted the NMS algorithm for further optimized candidate windows after R-Net, as described in the MTCNN model. After it locates its objective in each image, it corrects all feature points to the same position to

increase recognition accuracy. We used O-Net to extract the five feature points of the facial images. In O-Net, the greater number of convolutional layers supports more supervision in identifying areas of the face. It also retains more image features by using a full connect layer. Finally, the upper-left and lower-right coordinates of the facial area and the five feature points of the face, that is, the two eyes, the nose, and the two corner points of the mouth, were selected.

4.5.2. Attention module results

We used the Grad-CAM method (Selvaraju et al., 2017) to analyze the effect of our attention model, for one thing, to show the weight distribution on each image, and, for another, to use a warm-to-cool colour spectrum to show which parts of an image received most attention. In the Grad-CAM method, each channel is weighted according to its gradient for each category on the feature map of the output from the attention model, to highlight specific areas in images, in order to represent the importance of their positions. Blue carries the least weight in an image, while green, orange and red represent steadily increasing weight. The process is as follows:

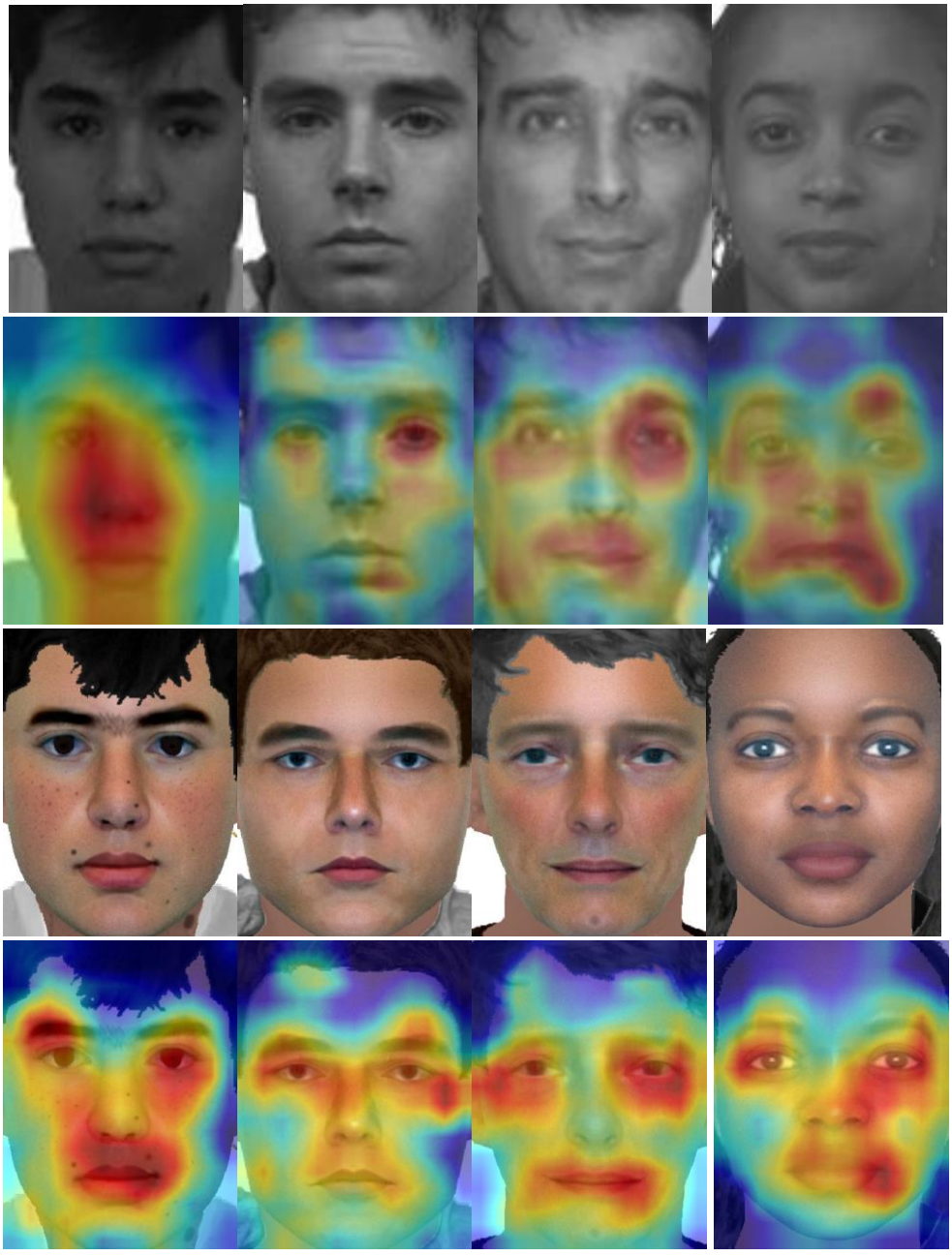
1. Reload the trained model.

2. Output a set of features from the attention block and calculate the gradient of the features.

3. Calculate the average gradient of each pixel on a specific channel.

4. Multiply each channel of the features by its weight and obtain a heatmap image.

As shown in Figures 4-4 to 4-8, the network pays more attention to certain regions of the face in the images. Grad-CAM visualizations can correctly localize them. Figures 4-5 and 4-6 show that the attention model gave more weight to spectacles than to other features. In contrast, when a facial image had no spectacles our attention model allocated average weights on all facial features, eyes, noses and mouths.



(a)

(b)

(c)

(d)

Figure 4-4 (a) original photos for UoM-SGFSA dataset (b) Grad-CAM visualizations for the original photos (c) original sketches UoM-SGFSA dataset (d) Grad-CAM visualizations for the original sketches

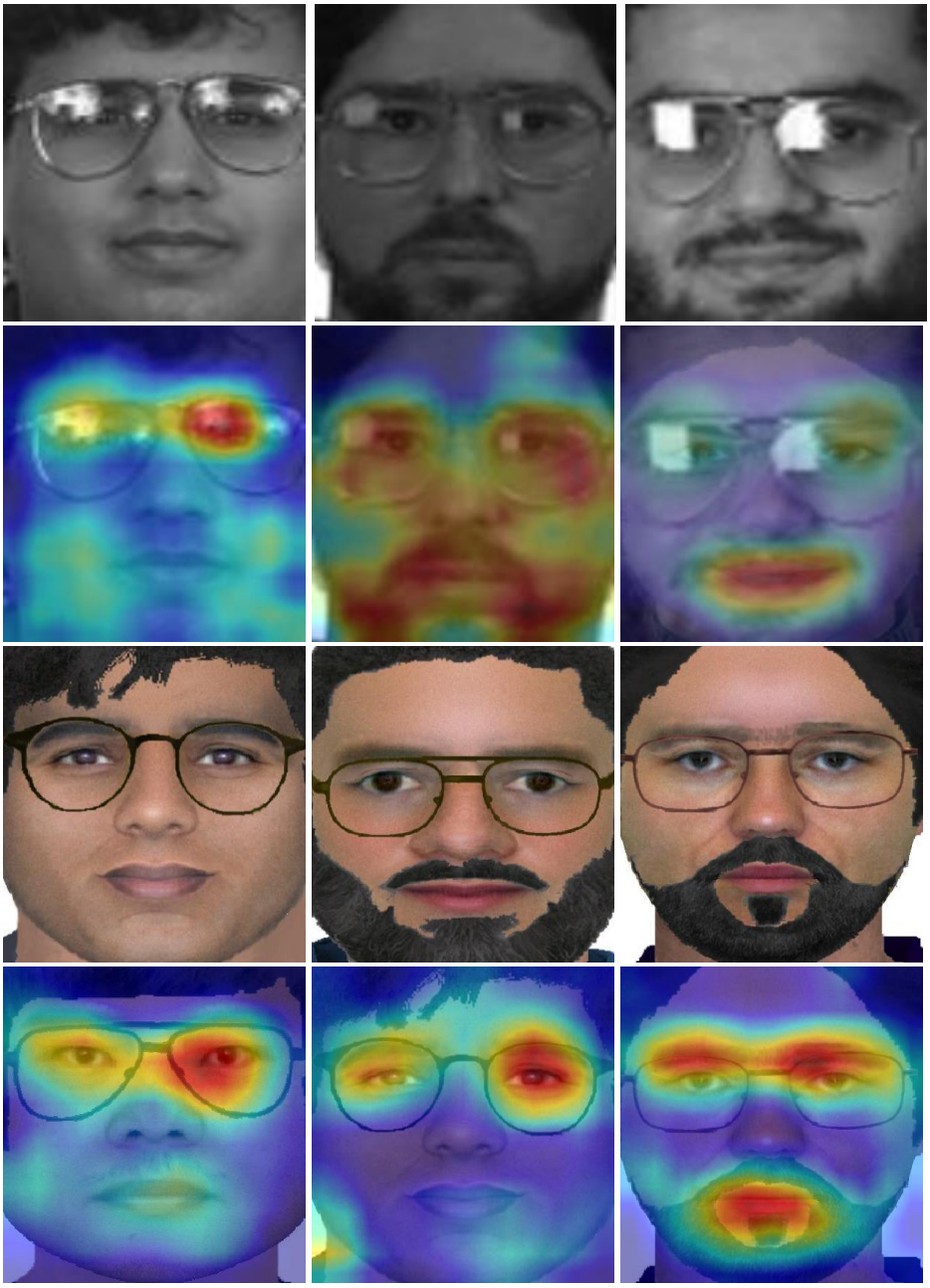
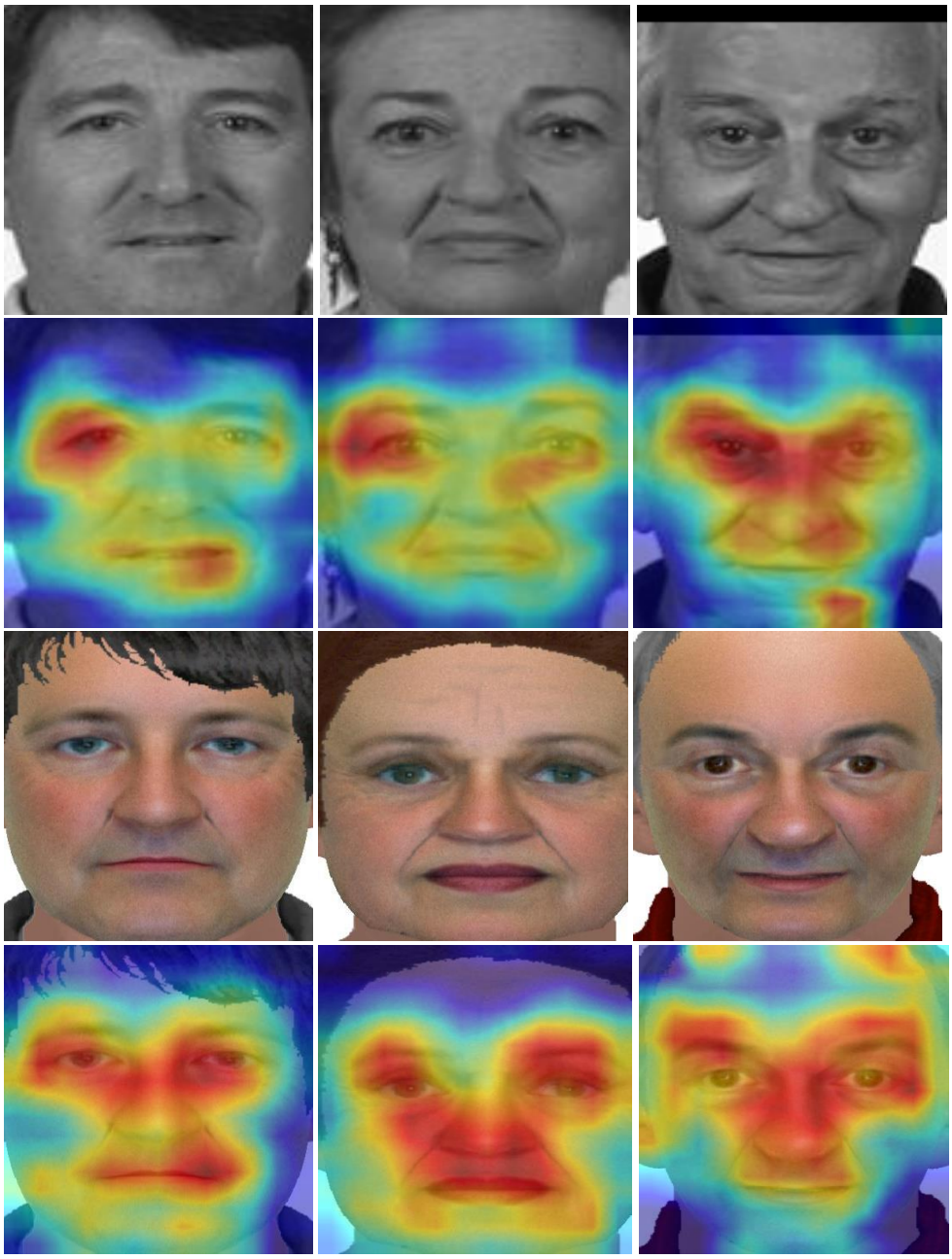


Figure 4-5 (a) Original photos in UoM-SGFSA dataset (b) Grad-CAM visualizations for the photos

(c) Original sketches in UoM-SGFSA dataset (d) Grad-CAM visualizations for the sketches



(a)

(b)

(c)

(d)

Figure 4-6 (a) Original photos in UoM-SGFSB dataset (b) Grad-CAM visualizations for the photos

(c) Original sketches in UoM-SGFSB dataset (d) Grad-CAM visualizations for the sketches

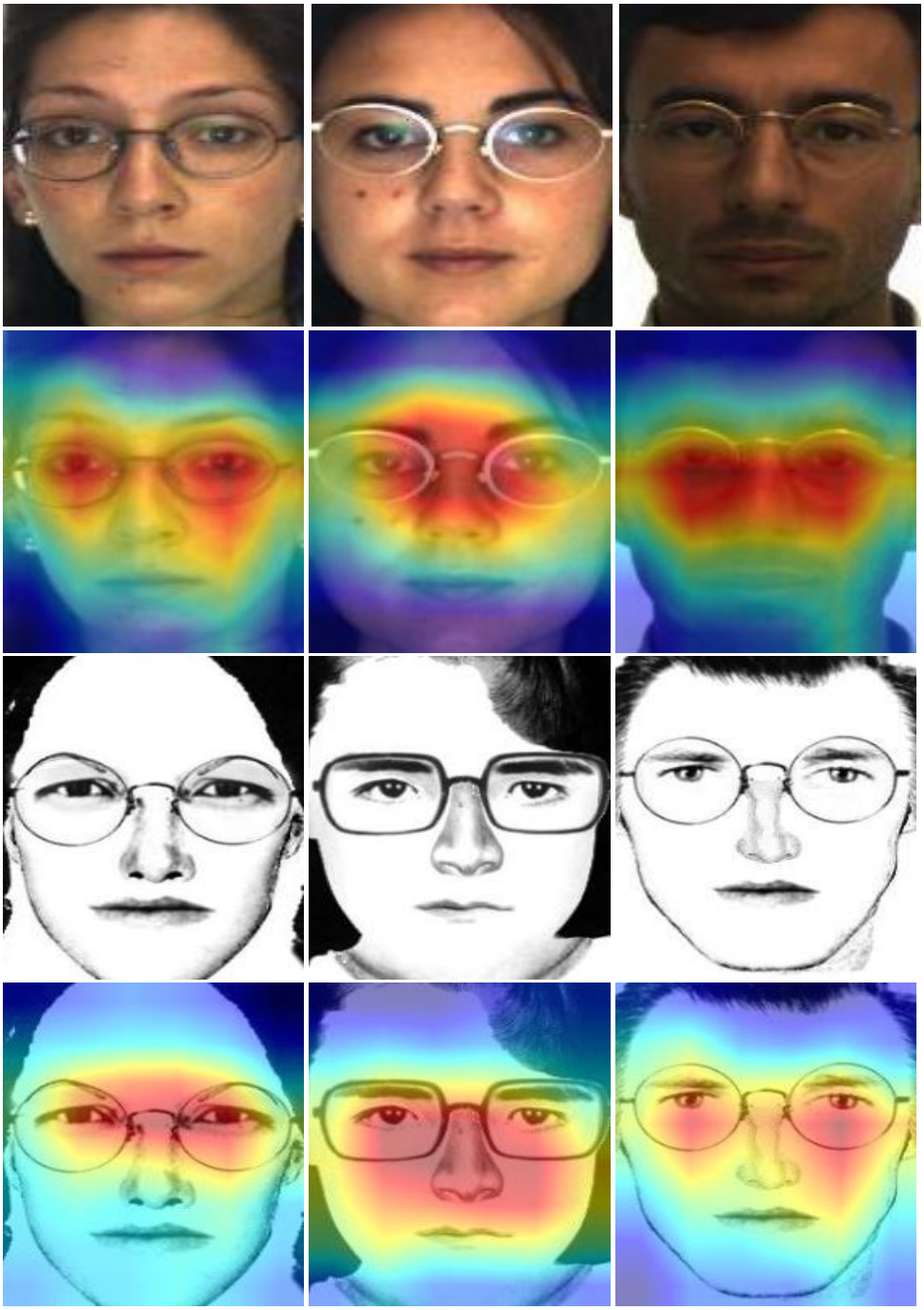
(a)

(b)

(c)

(d)

Figure 4-7 (a) Original photos in e-PRIP dataset (b) Grad-CAM visualizations for the photos (c) Original sketches in e-PRIP dataset (d) Grad-CAM visualizations for the sketches



(a)

(b)

(c)

(d)

Figure 4-8 (a) Original photos in e-PRIP dataset (b) Grad-CAM visualizations for the photos (c) Original sketches in e-PRIP dataset (d) Grad-CAM visualizations for the sketches

4.6. Testing results

The input for our model consisted of three images, each of which was entered into the corresponding channel. For the first and the last channels, the input images were facial photos, while the input of the middle channel was the facial sketch of the same subject that had been input to the first channel. To train all the datasets, we built a model using ResNet (He et al., 2016) as a pre-trained model, together with the attention module and SPP layer described above. The initial learning rate was set as 6e-5 with the Adam optimizer.

Evaluation on UoM-SGFS dataset

To obtain rank-1 accuracy, we compared different approaches to the proposed method, as well as several state-of-the-art methods on the UoM-SGFSA dataset. From the feature map generated by class activation mapping (Yin et al., 2016), we could see that the weights of our network focused on parts of the facial photos and their corresponding sketches. After we applied the channel attention module and spatial attention module, more similar features were extracted from the same position of the facial photo and the sketch. As shown in Table 4-1, the FacialNet model scored 45.50% using the shared parameter network. Our model, which

combines the attention module and SPP layer with Resnet34 as the pre-trained model, increased the accuracy to 66.70%.

Table 4-1 Experimental results on UoM-SGFSA dataset

Methods	Top-1 accuracy	Top-10 accuracy
FaceNet	45.50%	50.70%
Triplet net +Attention +SPP Layer	66.75%	90.46%
(Peng et al., 2019)	64.80%	92.13%
(Galea and Farrugia, 2017)	31.60%	66.13%

Table 4-2 Experimental results on UoM-SGFSB dataset

Methods	Top-1 accuracy	Top-10 accuracy
FaceNet	52.00%	80.10%
Triplet net +Attention +SPP Layer	81.25%	90.56%
(Peng et al., 2019)	72.53%	94.80%
(Galea and Farrugia, 2017)	52.17%	82.67%

The facial photo and corresponding sketch in Set B of the UoM-SGFS dataset resembled one another more closely than those in Set A. The metrics of the extracted features from photos and their corresponding sketches in Set B were closer with the use of the attention module than those in Set A. Table 4-2 shows

that the accuracy for Set B using our model exceeds 81%, while the accuracy of the others did not reach 75%.

Evaluation on e-PRIP dataset

Table 4-3 shows the experimental results for the e-PRIP dataset. It can be seen that, although the top-1 accuracy of our method went down to 58.85%, it is still more accurate than any other state-of-the-art method.

Table 4-3 Experimental results on e-PRIP dataset

Methods	Top-1 accuracy	Top-10 accuracy
FaceNet	50.20%	56.70%
Triplet net +Attention +SPP Layer	58.85%	84.60%
(Peng et al., 2019)		82.80%
(Galea and Farrugia, 2017)	54.90%	80.80%
(Mittal et al., 2015)	52%	60.20%

Evaluation on hand-drawn face photo-sketch dataset

The component facial sketches generated by the software were close to authentic forensic sketches. However, because the options for facial attributes in the software are limited, the recognition was not very accurate, even though the attention module and the SPP layer were added to give more weight to important

and useful parts of the images. We also tested our model on the hand-drawn sketches (Wang and Tang, 2009).

(a)

(b)

Figure 4-9 (a) The example result for hand-drawn sketch on Top-1 accuracy (b) The example result for hand-drawn sketch on Top-10 accuracy.

Table 4-4 Experimental results on CUFS dataset

Methods	Top-10 accuracy
FaceNet	75.10%
Triplet net +Attention +SPP Layer	89.60%
Triplet net + channel Attention block+ SPP Layer	79.15%
Triplet net + Spatial Attention block +SPP Layer	77.36%
Triplet net + Attention blocks +SPP Layer	89.60%
(Wan et al., 2019)	92.56%

However, the dataset contained only 188 image pairs, so the amount was too small for training. We used data augmentation to increase the number of the training data. In the tests, the top-1 accuracy of our model was higher than that

of the FaceNet model, scoring 84.27%. Table 4-4 shows the top-10 accuracy of our methods compared with some others.

4.7. Conclusion

We presented a novel approach to enhance the recognition accuracy for facial photo-sketch datasets. We built a triplet network architecture with a triplet loss function layer to learn about the feature representation. The triplet network gave us a chance to carry out end-to-end learning between the input images and the desired embedding space after extracting features from the correlated facial photos and facial sketches. To optimize the network for the final task, an essential part of learning using the triplet loss was to compare facial images by computing the Euclidean distance in the embedding space. In order to increase the learning ability of the network, all the training data for selecting the shortest distance between images of the same person and the longest between images of different persons had to be input into this model as either a positive or a negative sample. However, three straightforward samples, such as similar positive sample pairs, or widely negative sample pairs, may be selected as input data for training the generalization ability of a limited network. But a major disadvantage is that selecting widely negative sample pairs too often makes the training unstable.

Otherwise, repeated images as input data generate too much redundant information for effective training. We introduced an attention model to focus on features in the same position using a channel attention block and a spatial attention block. Attention blocks generate more distinguishable features which adapt well to the depth of our triplet network. In addition, we introduced an SPP layer to extract the features from image blocks of different scales, and to reduce the influence of distortion and noise from the input images. To verify the model's effectiveness, we tested two kinds of facial photo-sketch dataset for recognition. For component facial photo-sketch datasets, the introduced attention model improves the performance of our model described in this chapter, and moreover, the performance of our model was better than any other popular state-of-the-art methods. The highest accuracy that we obtained from the three datasets was 81.25%. We used the hand-drawn facial photo-sketch dataset CUFS to verify that the corresponding features between facial photos and hand-drawn sketches resembled one another more closely than did the corresponding features of facial photos and component facial sketches.

Chapter5:

Siamese Graph

Convolution Network

5.1. Introduction

The main challenge for facial photo-sketch recognition is the modality difference between facial photos and facial sketches. In previous chapters, we used the Siamese network and an attention-modulated triplet network. The two types of network models use the following structure (see Figure 5-1). One advantage is that these models utilize the similarity between different types of images to increase recognition accuracy. In a recognition system, the label is used to mark the different categories of data. However, the number of images is too few to learn unique features from each person. This factor leads to low recognition accuracy using a suitable model. Therefore, we convert the traditional type of label which is used to mark the different person into a binary label. For this binary label, '1' represents the same person and '0' is the different person. Rather than the traditional deep learning method for extracted images' features, the features we get are more able to reflect the similarity between images. Another is that the attention mechanism is used to focus on the key features for increasing recognition accuracy. However, a facial sketch may miss some features of the corresponding facial photo, because of the actual process of making a sketch. These missing features impair the stability of traditional recognition algorithms for facial photo-sketch datasets. Moreover, this is due not only to the difference in modality, but also the changed context. Altering

the illumination will cause different reflections of light on the face. As a result, the textural features of the facial image may change, reducing the similarity and correspondences of features of the facial image and the corresponding facial sketch. Because the spatial relationship of facial attributes in a realistic facial image is affected by rich local facial changes and special illumination, distortion can be avoided only by discarding some basic facial features. This prevents the feature extractor from extracting valid features for recognition, those that show similarities to the features in facial photos. In human beings, the overall features which reveal integral character, such as the disposition of skin colour, contours, facial attributes, and local features, are critical for increasing the accuracy of recognition and the perception of people's faces. Moreover, variations in the thickness or fineness of the drawn lines act as potential factors for confusion, and effectively increase the noise. This noise interferes with the description of detailed local attributions and irregular features such as scars. The effect of this is that features from two different people can sometimes be more alike than two features from the same person.

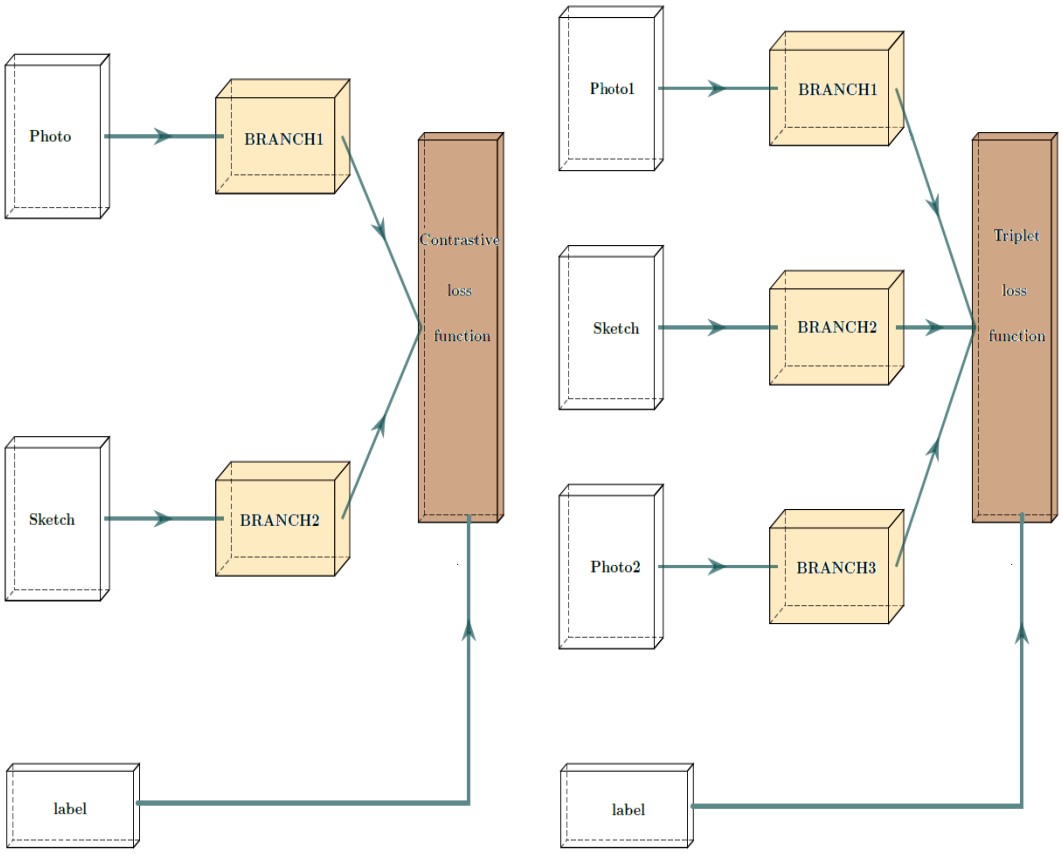


Figure 5-1 The structure of Siamese network in chapter 3 and the attention-modulated triplet network in chapter 4

Humans do not work like algorithms: our power to recognize depends not only on the attention mechanism, but also on the similar structures in images, which represent the relationship between global features and local features. In the present study, we decided to extract graphs from images to reduce the effect of uncontrollable factors, such as illumination and expression. Then we designed a Siamese graph convolution network (GCN) to learn more about the embedding

space. Finally, we used the contrastive loss function to optimize the node and edge information and compare the similarities of each pair of graphs. The contributions of this aspect of our paper are as follows:

(1) We used a CNN and two types of super-pixel methods to generate a graph structure from facial photos and facial sketches. First, a full convolution network was used to extract image edges from photos and sketches. The super-pixel methods were then used to cluster similar pixels into small regions. After the features were extracted from each cluster region, a graph representing facial contour information was built.

(2) A Siamese GCN was designed to transfer a graph structure into an embedding space which retains the intrinsic structural properties of graphs. In addition, this Siamese GCN was able to capture the topology of the graph and the relationship between nodes with shared weights and to keep a similar graph structure and node information for recognition.

(3) We combined a deep graph matching method with the GCN and the MoNet networks to extract more similar cross-modal graph features than had been extracted by the original weight-shared Siamese network. The method used the

contrastive loss function to measure the graph distance based on Euclidean distance. It was able to reduce the difference between two graphs of the same class but different modalities.

5.2. Related work of Graph convolution network

Unlike images and tables, graphs not only present the connections between objects, but also the topology relationships between objects' local features. The graph uses nodes and edges to represent the relations between objects (Wang et al., 2019) (Garcia and Bruna, 2017) (Knyazev et al., 2019a). The nodes represent objects. The attributes of two nodes are stored in the edges which connect the two nodes. These attributes in edges show a kind of mathematical relationship, such as Euclidean distance, between any two objects. However, the length for each edge is difference to build regularize data structure using gridding. Thus, as unstructured data, the relationship between any two nodes is not regulated as images are (i.e., the number of neighbours around each node is not fixed), and a graph is not orderly as a regular topology is. In addition, the spatial feature is extracted by calculating the weighted sum of the center pixel and adjacent pixels from the CNN model. because the number of adjacent vertices of each vertex may be different, the convolution kernel cannot keep translation

invariance on non-Euclidean structure data. Thus, the CNN model which adopt the same size's convolution kernel cannot extract spatial feature from graph structure data. These reasons mean that the graph network cannot use a traditional convolution operation to extract features. To extract them, one strategy is to do linear mapping following a fully connected operation. This method, however, loses certain advantages of convolution, such as weight sharing and local connection. Moreover, the number of parameters is too great to be trained. To extract information from a graph, we subjected it to a convolution operation. The core method of graph convolution is that the features of the graph are extracted using propagation in the graph network features and messages from the node, by means of matrix multiplication and linear mapping of the graph. The target of the GCN model is to use the relationship between nodes and edges to extract spatial features from an existing graph. Due to the influence of a node's neighbours and other points which relate to it, each node in a graph regulates its state until the final balance to describe the structure of various object by means of their attributes. The first GCN was proposed by Bruna et al. (2013). This method used the convolution method to extract features from non-Euclidean space. The authors proposed two models, one based on spatial space and the

other on spectral space. For each layer, the graph information was extracted from several filters and saved as neurons. The number of neurons in each layer, called Kth layer clustering, is the cluster's result of the last layer. The advantage of this GCN is that it extracts features from various items of graph structure data, especially from weakly connected graphs. However, this design cannot achieve a shared weight strategy for different positions on the graph. Alternatively, Defferrard et al. (2016) proposed an approximate smooth filter in the spectral domain using Chebyshev polynomials. A set of parameters that included shared-weight parameters from the neighbours of the same order and unshared-weight parameters from the neighbours of a different order were added to the convolution kernel in the GCN that had been proposed to reduce the complexity of the parameters. The property of the graph was to be locally stationary, because the relationship between the nearest nodes was stronger, and fewer parameters are needed to train a hyperscale graph. However, the model cannot distribute different weights evenly between different neighbours in a same-order neighbourhood using fewer parameters. Yet GCNs can be used to extract information on first-order neighbours in the graph (Kipf and Welling, 2016). The method introduces a graph Laplacian regularization term in the GCN model for

semi-supervisory classification. This network is a variant of the traditional convolution algorithm on graph structure data for processing the latter directly. The essence of this GCN network is that the features of each node are composite, in that the feature weighting of the node and the node's neighbours are propagated through the topology. However, this model is unable to capture of its spatial information. Meanwhile, for each node, the principle of graph convolution filters is similar to that of the filters of the CNN model.

In the traditional CNN-based methods, the image is a regular grid structure. However, some of the values of the pixels that make up images is too similar to extract features. A graph convolutional network allows the data of an image to be treated as a kind of non-Euclidean structure. In general, the image can be transformed into a graph by constructing a k-NN similarity graph using image pixels as nodes. Peng et al. (Peng et al., 2017) have proposed a graphical representation based on Markov networks for facial photo-sketch recognition. They use Markov networks to select a set of the nearest image patches from overlapping photo image patches and overlapping sketch patches based on a coupled metric of representational similarity. The advantage of this is that the Markov network extracts spatial features for recognition. The deep sparse graph

neural network (DSGNN) (R.Wu et al., 2017) extracts an undirected graph G from a facial photo image for recognition. The graph nodes are the divided blocks for each facial image. Undirected edges are generated using Euclidean distance to calculate the correlation between pairs of image patches. After the features from the deep sparse graph neural networks have been learned, the recognition accuracy of DSGNN on the LFW dataset (Huang et al., 2008) reaches 99.5%. However, this method of extracting features using the CNN model to generate the graph structure data is sensitive to the effects of occlusion and illumination. Wang et al. (Wang et al., 2019) use the GCN model to predict a new node from an existing graph model. They used a KNN to build a graph structure after extracting facial features using a CNN model. Then the similarity nodes are clustered by GCN, using the weighted average between adjacent nodes and neighbour nodes. This method supposes that if, after inference from a graph, two facial images have the same ID, they have connectivity. Knyazev et al. (Knyazev et al., 2019b) build a hierarchical multigraph network to improve the accurate graph classification of image datasets. In the first step, the graph for an image is built by super-pixels of the images. This method builds a three-layer graph convolution network on graph

data to extract node information from the low-resolution image dataset with increasing recognition accuracy.

5.3. The proposed method

In a two-dimensional image, each pixel can be treated like a node in a graph. The local information of a node can no longer be described as a simple rectangular grid. The graph is generated by the correlation between the nodes which calculates the k^{th} nearest neighbours from one node to another. The described node contains the position information of the pixels and the corresponding textural information. This method exhibits more powerful and accurate node embedding, according to the information on the neighbour nodes. For example, textural feature calculations multiply the value of the pixels in a given area. Even if the similarities between a photo of a person and the corresponding sketch are high, the features in the photo are markedly different because of the illumination. One thing to remember is that the extracted texture may change radically when the resolution of the image changes. Another is that a 2D image cannot reflect the true texture of the surface of a 3D object. Thus, previous algorithms is difficult to distinguish. For this reason we used a graph to build the semantic relationship behind an image, rather than the content of the image itself.

This chapter, reports the use of graph structure data as input to reduce the modality gap between photos and sketches for facial recognition, based on a Siamese network. The architecture of our model is shown in Figure 5-2. The input graph structure data were generated from images using super-pixel methods (Vedaldi and Soatto, 2008) (Achanta et al., 2012). First, the holistically-nested edge detection (HED) method (Xie and Tu, 2015) was used to generate an edge image that simplifies the image information. Then the super-pixel method was used on the edge image to segment the image into regions. A graph was generated by taking the centre of each region as a node and the distance between each pair of nodes as the edge feature. Next, a set of input data composed of two graphs, one from the sketch image and one from the photo image, was input into our Siamese network model. When we added more graph convolution layers to the model, the final state of each node involved the hidden state of a good number of neighbouring nodes. This made the process of backpropagation very complicated. Although some methods are intended to improve model efficiency through rapid sampling and subgraph training, they still cannot be extended to the deep architecture of large graphs. Thus, each channel in the Siamese network model consists of two graph convolution layers to extract features from the graph

in an embedding space. Finally, Euclidean distance was used in the contrastive loss function to measure the distance between any two reconstructed graphs for recognition. The aim of our model was to measure the degree of similarity between the graphs in order to increase the accuracy of recognition.

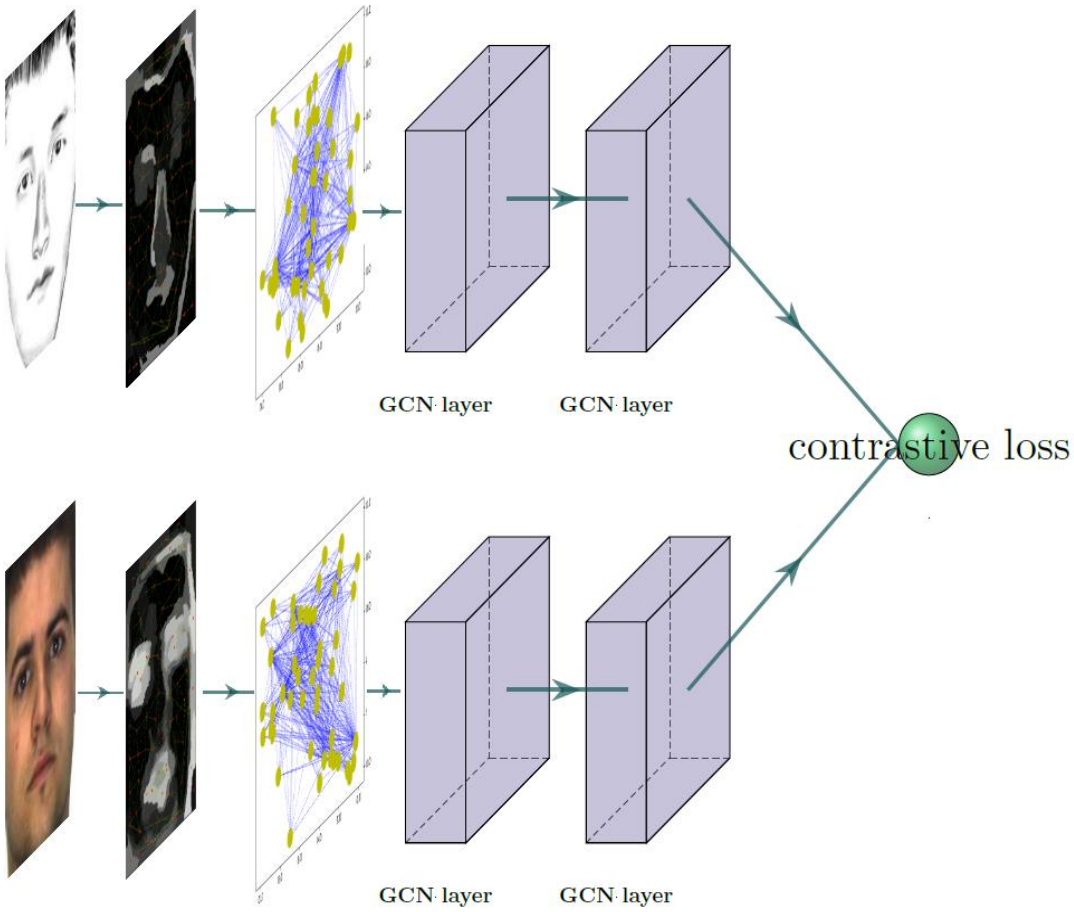


Figure 5-2 Architecture of the Siamese graph network model. Each branch of our model consists of two graph layers for extracting graph features.

5.3.1.Graph structure data for images

Photos are generated by the principles of optical imaging. A facial photo uses the relationships between pixels to delineate all the features of an accrual human face in a two-dimensional space. In contrast, a facial sketch uses geometric deformation and varied line density to represent the illumination and the characteristics of a face. Since the representations in photos and sketches are different, traditional convolutional networks mainly designed to extract feature from natural images. Therefore, this kind of CNN model can capture only the local structure of a sketch, but cannot fully extract its colour and textural information. Even if a large-sized convolution kernel is used to obtain more spatial structure information, instead of a small-sized convolution kernel. It fails to represent the model features that a photo does, because the dimension of extracted feature is too high for training. Therefore, we decided to build a graph structure based on image features and structural information. The first step was to transfer information from an image to a graph. In general, an image can be transferred into a 2D matrix as a regulated graph structure $G(V, E)$. Elements of the 2D matrix are considered to be nodes $v = \{1, 2, \dots, N\}$ of the regulated graph. In this graph the distance E between the nodes was treated as an attribute of the edge. The

nearer nodes all weigh more than the further nodes. However, training a graph neural network by this method is greedy of resources. The essential function of a graph convolution network is to extract spatial features from graphs from the relationship between the node and the neighbours. Thus, after several convolution operations the representations of nodes with similar characteristics will converge on one point, because the nodes with the same or similar pixel values can be converted to a node. To reduce the number of graph nodes, we tried to cluster the pixels with the same and similar values into the same region, as nodes of the graph. However, the regions of pixels with the same values are different from nodes where the pixels have similar values, because of the distinct modes of representation in facial photos and facial sketches. We used an edge detection method to extract the contours of the facial images and reduce the background noise, as is shown in Figure 5-3. It not only kept the structural properties of the image, but also reduced the amount of weakly relevant information. Traditional edge detection methods, such as Sobel, Prewitt, Canny, and HOG, use local region changes, including colour changing and illumination, to search image edges. However, sketches use lines of different widths to represent textural features, which means that some facial details cannot be

represented. The low-level features extracted by traditional methods do not reflect the actual edges of the sketch. Moreover, it is difficult to extract colour, illumination, and gradients from sketched images to detect edges because textural features in sketch images have weak distribution patterns at the edge. CNN-based methods (He et al., 2017) increase the recognition accuracy by using the kernels of large receptive fields to extract global features and details from images and pooling layers. Large receptive fields and pooling layers in low convolution layers remove more details than are removed in high convolution layers. Hence, the low convolution layers are used to extract the edge features, and the high convolution layers focus on the global semantic features. Thus, we were able to use the deep learning method on facial photos and sketched images to obtain the image contours from a high convolution layer which included more semantic information than a low convolution layer does. Meanwhile, the HED network (Xie and Tu, 2015) combined multi-scale features with a multi-level feature to map several multiple side output layers on the main convolutional network. It obtained a set of edges of different scales. The drawback of the HED network is that this model adopts many downsampling layers and does not fully fuse the multi-scale features, so its edge detection results in rough and fuzzy lines.

Figure 5-3 The image edge detection for face photos and face sketches using HED network. The examples are from e-PRIP dataset and UoM-SGFS dataset.

Bearing in mind the correlations between the pixels, the pixel colours, and the similarity of brightness in the image edge, we used super-pixel methods, such as Quickshift (Vedaldi and Soatto, 2008) and SLIC (Achanta et al., 2012) to cluster adjacent pixels with similar features in the same region. Super-pixel methods cluster pixels with the same or similar values in facial photos and sketches to generate a representative region as a node. In general, these segmentation

areas can be identified as attributes of the facial image. Then the features are extracted from each small region to build a graph. The process of super-pixel methods is similar to the K-means clustering algorithm, as follows:

Input: Facial photo or sketch

1. Based on K- meaning clustering method (Lloyd, 1982), select K super-expected centres as seed on the image

2. Fine-tune the position of the seed and determine the range for each seed.

3. Choose the closest nodes in the surrounding space of each seed as the region of seeds in the same category.

4. Calculate the average value of all the pixels in the K^{th} super-pixels region.

5 Repeat steps 3 and 4 until convergence is reached.

After implementing the super-pixels' algorithm, we build a three layers CNN model to extract feature from each super-pixel region. Each super-pixel region was used as a node in an undirected graph structure. Then, the distance between each node using extracted feature which were obtained CNN model, is calculated using Euclidean distance and mapped as a graph edge. The image was mapped as a

weighted undirected graph $G(V, E)$. In graph G, $v = \{1, 2, \dots, N\}$ was applied to the regions using super-pixels for the image. E was a set of edge for adjacent regions. In our method, the weights of the corresponding edges $W(v_i, v_j)$ showed the differences between the features of the region. The generated graph data is shown in Figure 5-4.

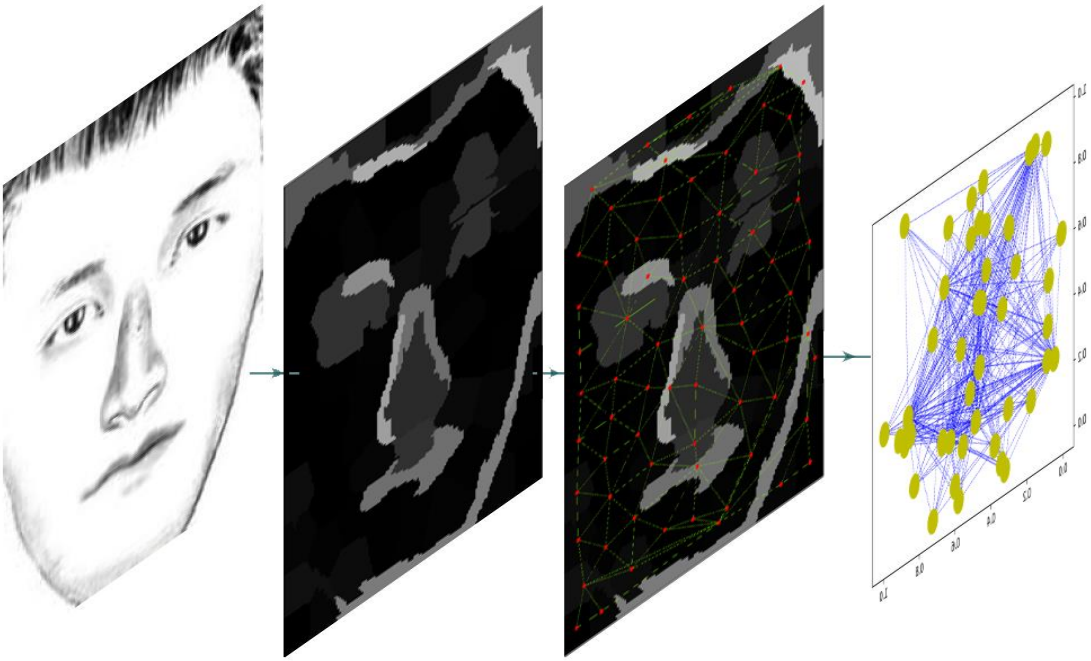


Figure 5-4 The pipeline of creating graph structure data from an image. The first step is to extract image edges using the holistically-nested edge detection method. Next, a superpixel segmentation of the edge image is generated. Then, a region adjacency graph is built based on the superpixel segmentation.

5.3.2.Graph convolution network

For recognition purposes, we adopted two strategies: GCN (Defferrard et al., 2016) and MoNet (Monti et al., 2017) for use on our graph data. As input, we used undirected graph data $G(V, E)$ with N nodes V and edge E.

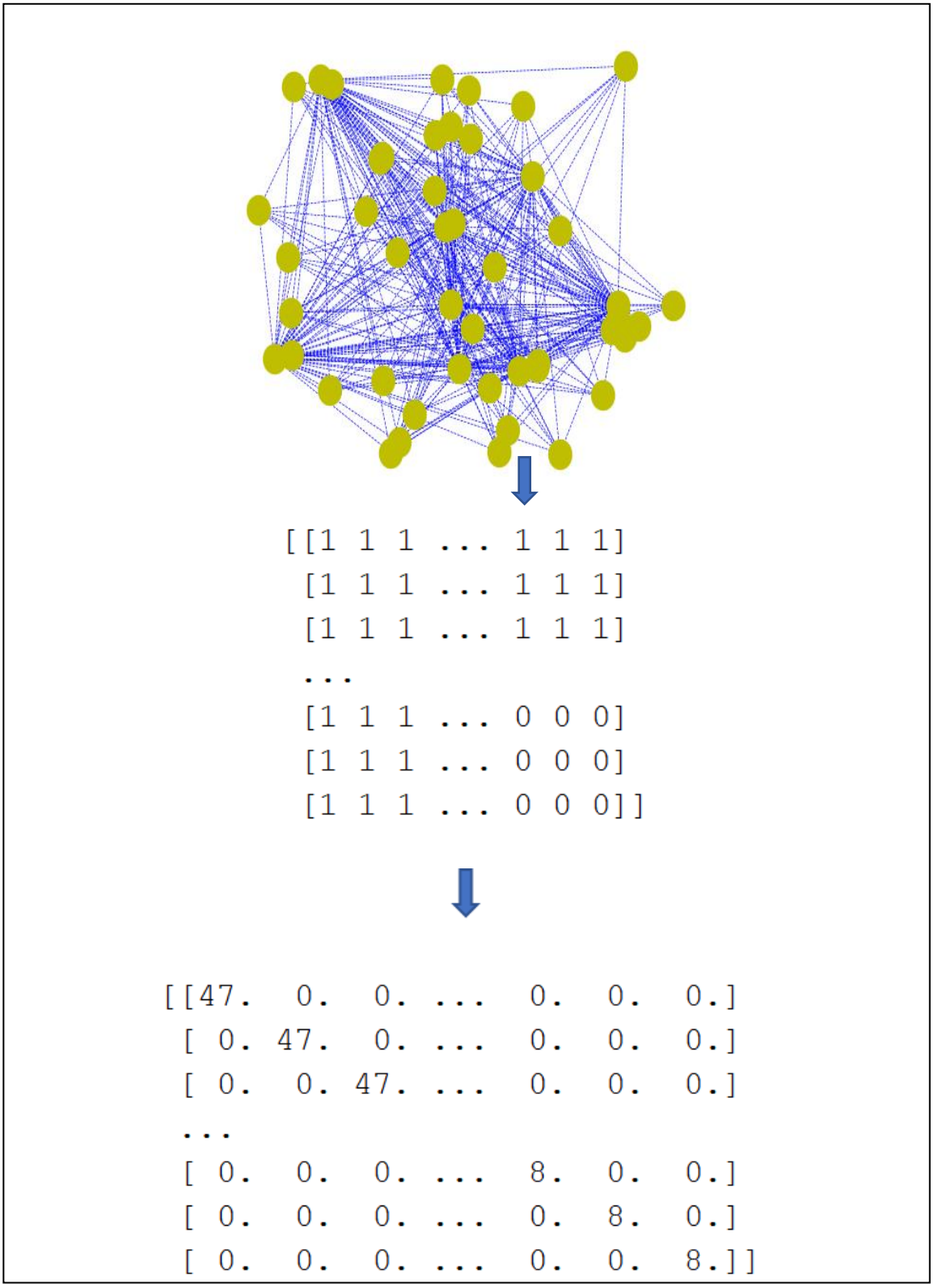


Figure 5-5 The calculation process from graph to generate adjacency matrix A and degree matrix D

This graph is represented by three matrices: one is the adjacency matrix A which is used to represent the relationship between the nodes. If the value of an element is 0, the two nodes of the corresponding row and column are not related. Otherwise, it means the two nodes are connected. The second is the degree matrix D. This is a diagonal matrix, with the diagonal elements $D = \sum_i A_{ij}$. The degree of each node refers to the number of nodes that are connected to it. The third matrix is the feature matrix X, which represents the node features. To undertake the convolution operation for the graph, the graph needed to be transferred into the adjacency matrix A. However, when we calculated the new features using a filter for each node in the graph, we added a self-loop to it so that it could add its own features in the adjacency matrix A. The adjacency matrix A with all possible self-loops is shown as:

$$\tilde{A} = A + I_N \qquad (5\text{-}1)$$

After this, the adjacency matrix was normalized using the inverse of the degree matrix:

$$\hat{A} = \widetilde{D^{-1}}\tilde{A} \qquad (5\text{-}2)$$

The normalization method avoided a gradient explosion or gradient disappearance in the training stage. Then we used a graph convolutional layer on the graph-structure data to extract their features. Figure 5-5 represent process from graph to generate adjacency matrix A and degree matrix D. The core of GCN for convolution is to use the Fourier transform on a graph, as follows:

$$\mathrm{f} * g_\theta = U\mathrm{diag}(U^T g_\theta)U^T f \qquad (5\text{-}3)$$

where g_θ was the convolutional kernel and $U^T g_\theta$ and $U^T f$ represented respectively the Fourier transformation of g_θ and f, induced from the Laplace matrix of the graph. $U = (U_1, U_2, \ldots U_n)$ were the orthonormal eigenvectors of the Laplacian matrix $\tilde{A}$. The core of GCN is that the eigenfunction of the Laplacian matrix was transferred to the eigenvector of the Laplacian matrix calculated from graph G. The node feature of each layer in the GCN network was composed by the convolution of signals. Then an activation function was used to perform a nonlinear transformation and obtained a matrix that aggregated the features of adjacent vertices to generate a new representative node. According to the principles of a convolutional network, GCN uses some overlapping convolution layers to provide multi-order neighbourhood information for updating. In our model, we adopted GCN to extract the first-order neighbourhood information from graphs

directly using graph topology. From the l^{th} GCN layer, the extracted message $f_{message}^{l+1}$ could be represented as follows:

$$f_{message}^{l+1} = w_0^l f_i^l + \sum_{E_i \in N(V_i)} w_1^l f_j^l \qquad j \in N(v_i) \qquad (5\text{-}4)$$

where $N(V_i)$ was a set of nodes which connected with V_i in graph $G(V, E)$, and w_0^l and w_1^l were weights of nodes. The process for two graph convolution layers in each channel in the Siamese network is:

1. Input a graph $G(V, E)$

2. Calculate the adjacency matrix A, degree matrix D and feature matrix X from G.

3. Calculate $\hat{A}$ using (5-2)

4. Implement two GCN layers with ReLU as the activating function.

5. Obtain an embedding graph.

The original Siamese network uses the contrastive loss function to calculate the similarity between two graphs in an embedding space. Instead of mapping the graph in a vector space, we used the graph's matching networks (Li et al., 2019)

to update the nodes of our graph network model. It received and clustered the information from the neighbouring nodes of each selected node, and fused the local graph structural information. This method not only aggregated messages on the edges of each graph, but also changed the way that the nodes in each propagation layer were updated, using a cross-graph matching vector. This cross-graph matching vector measured the degree to which the nodes in one graph matched several nodes in another graph.

Another strategy was to build graph layers according to the MoNet method (Monti et al., 2017). Spectral graph convolution depends on the specific feature function of the Laplace matrix, so it is not easy to transfer the spectral graph convolution network model that has been learned to another graph with different feature functions. However, the space-based method alternates the convolution to the combination of graph signals in the neighbourhood of the node, and defines a learnable filter in the vertex domain. This method is designed with a universal patch operator that integrates signals to the neighbourhood of its nodes. MoNet introduces pseudo-coordinates of nodes to determine the relative positions of a node and its neighbours in D-dimensions and thus increases the power of the

model to be generalized. The convolution calculation of a node x is defined in MoNet as follows:

$$(f * g)(x) = \sum_{j=1}^{J} g_j D_j(x) f \qquad (5\text{-}5)$$

where f is the signal on the graph, g is the convolutional kernel with dimension J, g_j is the j^{th} element of g, and $D_j(x)f$ is a weighted sum of the signal on x's neighbouring nodes. Here the weight depends on the pseudo-coordinates of every neighbour. After the GCN layer, a pooling layer was used to coarsen the graph so that the original information could be transferred to the nodes of the new graph.

5.4. Settings of the experiment

We used the MTCNN model (Zhang et al., 2016) to detect the location of facial images and crop all facial images to the same size, 128*128. After extracting the image edges, the contours of the images were represented by grayscale images. Next ,we built and tested four models, namely, two Siamese networks based on GCN using SLIC and Quickshift, and two Siamese networks based on MoNet, one using SLIC and the other using Quickshift. In detail, we used SLIC (Achanta et al., 2012) and Quickshif (Vedaldi and Soatto, 2008) respectively to extract

super-pixel regions, and based on which build graphs for all the facial photos and facial sketches. For SLIC, the generated super-pixels regions were compact. Unlike other super-pixel algorithms, SLIC uses a simple clustering algorithm (a 'greedy algorithm') to obtain a clear boundary and improve the computing speed. We extracted N<100 super-pixels; each super-pixel region could be represented as a node, while an edge value was computed as the spatial distance between the super-pixel regions. The Quickshift algorithm was employed, involving an approximation of the kernelized mean-shift method. The first step was to calculate the average offset of the current point. Then the point was moved to a new place using the average offset. After this, the point continued to move until it balanced. One advantage of the Quickshift algorithm is that it requires the fewest parameters. This algorithm need only to set the kernel size. In our experiment, the kernel size was 2. Then we built GCN and MoNet as layers in our Siamese network. For MoNet, we used Equation (5-6), below, to compute the distance between two nodes:

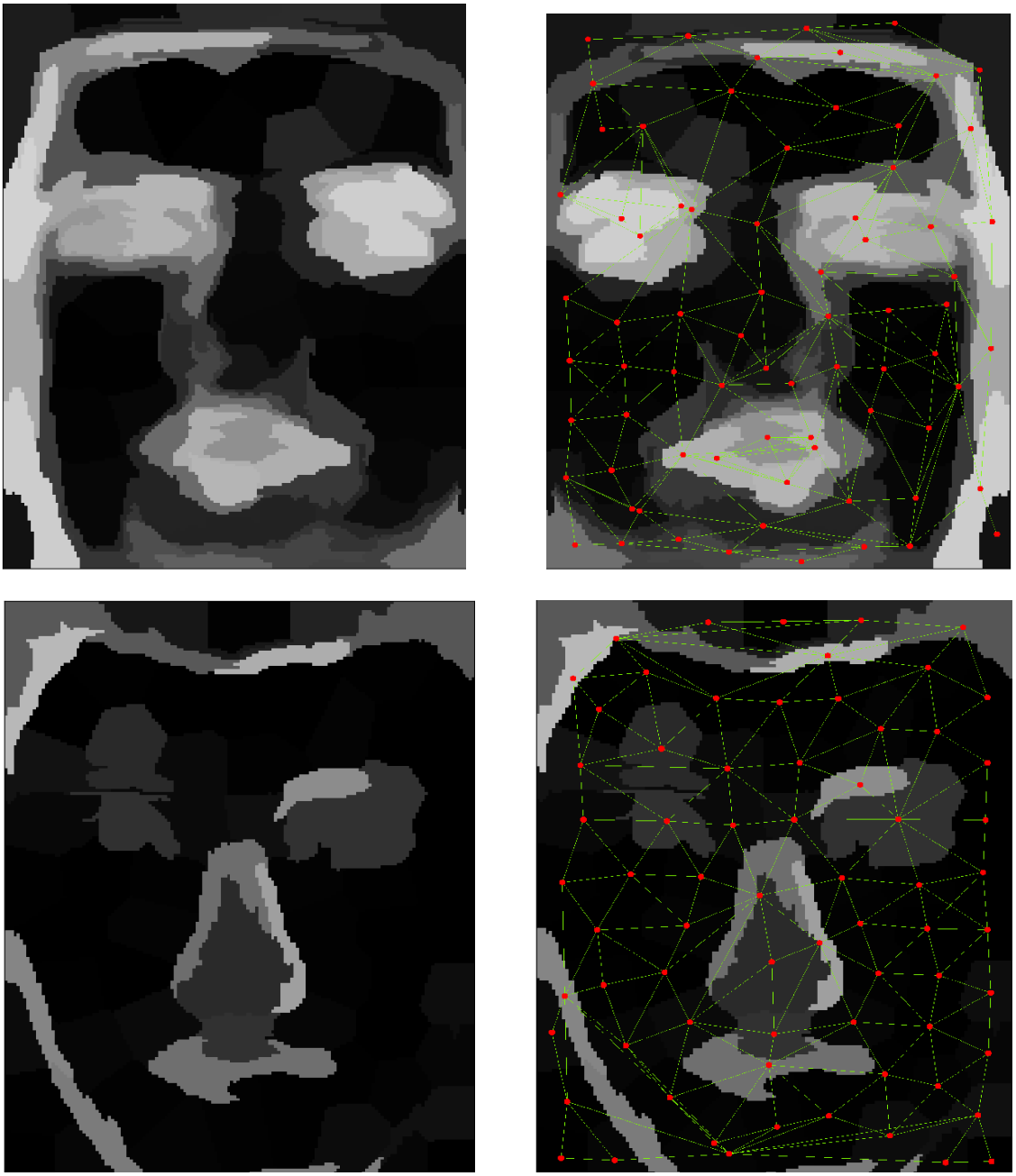
$$u(x,y) = \left(\frac{1}{\sqrt{\deg(x)}}, \frac{1}{\sqrt{\deg(y)}}\right) \qquad (5\text{-}6)$$

$\deg()$ is the degree for each node in a graph.

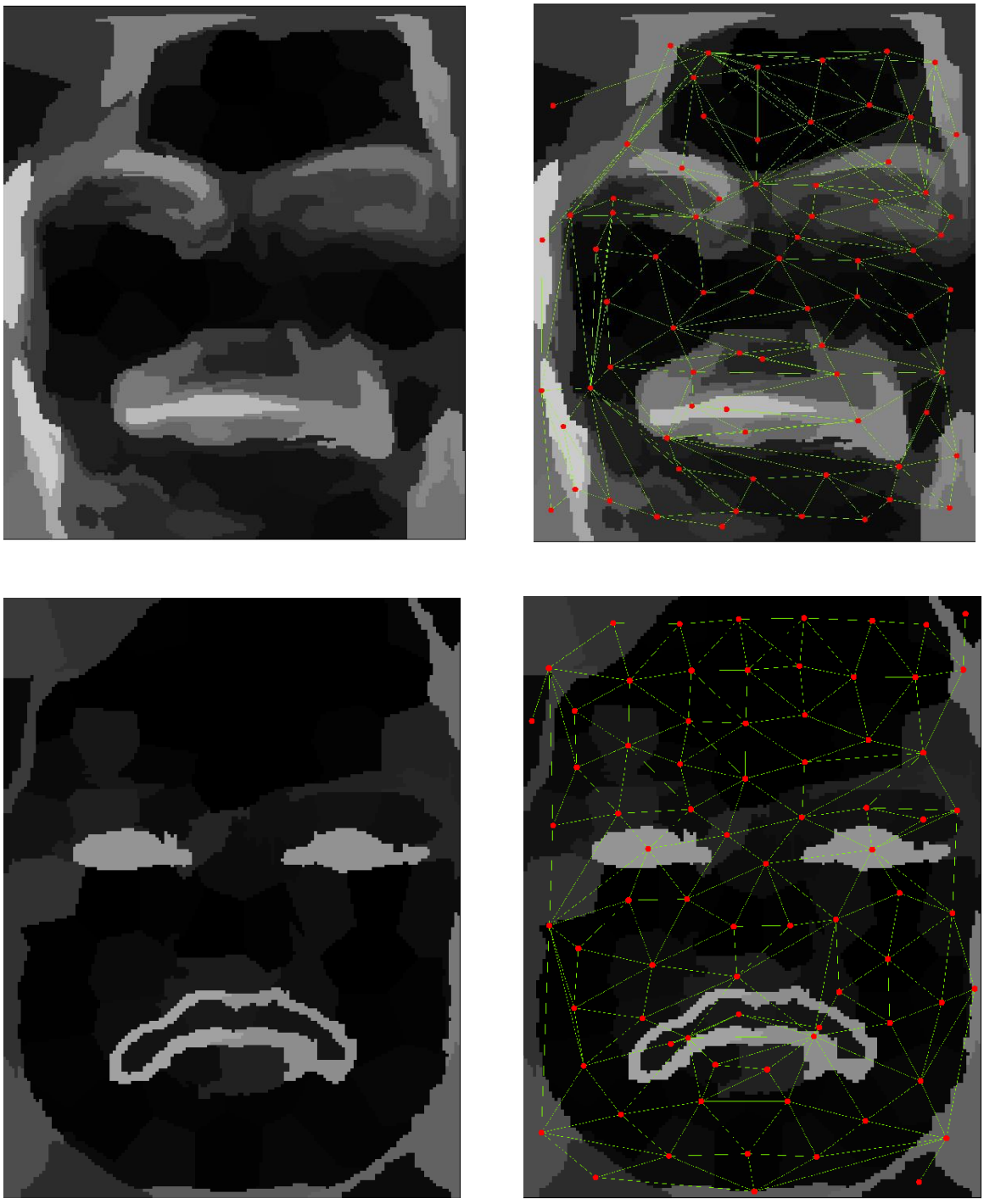
In our model, two graphs, one from a sketch and the other from a photo were input in the Siamese network. Then the loss function in our model was contrastive loss which compared the similarities in pairs of input graph data.

$$Loss = \frac{1}{2N}\sum_{1}^{N}((1-y_n)d_n^2 + (y_n)\max(L - d_n^2, 0)) \qquad (5\text{-}7)$$

where y_n is the label for each input pair. $y_n = 0$ represents a pair in the same class, while $y_n = 1$ represents a pair in different classes. L is a margin to measure the distance between data in the same class and in different classes. d_n is the difference between the input graphs. In order to increase the distance between different class data, we used squared Euclidean distance to measure the difference between samples. We trained this Siamese network using the Adam optimizer. The learning rate was set as 1e-6.



(A)



(B)

Figure 5-6 The results of superpixels algorithms (SLIC (A) and Quickshift (B)) for face photos and face sketchs using HED network

5.5. Results

Because composite facial sketches are widely applied for facial recognition in forensics, we tested our models by means of three composite facial photo-sketch datasets which had different characteristics (the UoM-SGFSA, UoM-SGFSB, and e-PRIP datasets).

We compared the performance of our models with some state-of-the-art models (Galea and Farrugia, 2017, Peng et al., 2019, Mittal et al., 2015). Tables 5-1 to 5-3 compare the performance of the three different composite facial sketch datasets. The Siamese model with GCN and Quickshift has the best Top-1 accuracy on datasets UoM-SGFSA and e-PRIP, with recognition accuracy of 74.16% and 55.28%, respectively. However, the recognition accuracy of our models was lower than that of Peng et al. on the UoM-SGFSB dataset.

Table 5-1 Experimental results on UoM-SGFSA

Methods	Top-1 accuracy	Top-10 accuracy
(Peng et al., 2019)	64.80%	92.13%
DCNN (Galea and Farrugia, 2017)	31.60%	66.13%
Siamese GCN(Quickshift)	74.16%	76.66%
Siamese MoNet (Quickshift)	64.17%	74.17%
Siamese GCN (SLIC)	68.33%	72.25%
Siamese MoNet (SLIC)	66.65%	73.33%

Table 5-2 Experimental results on UoM-SGFSB

Methods	Top-1 accuracy	Top-10 accuracy
(Peng et al., 2019)	72.53%	94.80%
DCNN (Galea and Farrugia, 2017)	52.17%	82.67%
Siamese GCN(Quickshift)	65%	80.83%
Siamese MoNet (Quickshift)	62.5%	80%
Siamese GCN (SLIC)	60.83%	77.5%
Siamese MoNet (SLIC)	59.1%	79.17%

Table 5-3 Experimental results on e-PRIP dataset

Methods	Top-1 accuracy	Top-10 accuracy
DCNN (Galea and Farrugia, 2017)	54.90%	80.80%
(Mittal et al., 2015)	52%	60.20%
Siamese GCN (Quickshift)	55.28%	73.9%
Siamese MoNet (Quickshift)	50.4%	67.48%
Siamese GCN (SLIC)	47.15%	63.4%
Siamese MoNet (SLIC)	48.78%	61.78%

From the tables, we can see that the performance using the Quickshift method was better than that using the SLIC method. The SLIC algorithm uses K-means clustering to obtain super-pixel regions under an average distribution of cluster centres. Ignoring the image edge information as it does, its segmentation results for the super-pixel blocks are inaccurate. Different regions are classified as belonging to the same super-pixel block, producing under-segmented super-pixel blocks. We can also see that the performance of GCN is better than that of MoNet. The GCN trains all the nodes in the graph to obtain a new graph representation in the embedding space. A new graph presentation was generated from the last graph convolution layer, embedding the optimized node. However, the

representation of each node was affected by all the related nodes using GCN kernel which is extract the relationship based on nodes' degree. The convolution kernel is too focused on the local nodes' relationship to learn representation from our graph data. Otherwise, we try to build a deep GCN model to learning information from data. The experiments show any node in the graph almost contains the information of the whole graph. It means the representation for each node converges to a similar value. These may have made the graph convolution layer worse than the actual convolution layer.

Table 5-4 Experimental results on CUFSF dataset

Methods	Top-1 accuracy
(Wan et al., 2019)	80.80%
DCNN (Galea and Farrugia, 2017)	82.80%
(Mittal et al., 2015)	52%
Siamese GCN(Quick shift)	82.25%
Siamese Monet (Quick shift)	80.75%
Siamese GCN (SLIC)	77.5%
Siamese MoNet (SLIC)	75.5%

We used graph structure on the CUFS and CUFSF datasets to obtain the relationship between the pixels on the image edge while avoiding the effect of image distortion. However, HED could not extract suitable image edges to build the graph data. Its performance scored 87.71% and 82.25% for the CUFS dataset and the CUFSF dataset, respectively. Although the results all exceed 80%, the dataset was too small for recognition. The results from the hand-drawn facial photo-sketch datasets do not show the full capacity of our model.

Table 5-5 Experimental results on CUFS dataset

Methods	Top-1 accuracy
Siamese GCN (Quick shift)	87.71%
Siamese MoNet (Quick shift)	85.9 %
Siamese GCN (SLIC)	82.4 %
Siamese MoNet (SLIC)	78.9%

5.6. Comparison of our methods

In this part, we compare our three methods in operation on hand-drawn photo-sketch datasets of faces and composite photo-sketch datasets of faces. The performance of the attention-modulated triplet network for Uom-SGFS datasets

exceeded 90%. It could do so because the generated sketch was less deformed than other types of generated sketch. Thus, the similarities between the colour sketches and the corresponding photos in the background colour, skin colour and shape of facial attributes were all high. The attention-modulated triplet network which gave more weight to local information about the whole image than other methods favoured the extraction of similar features, such as facial shapes and facial attributes.

Because there were too few images in the e-PRIP datasets for adequate training in the Siamese network and triplet network, the overall Top-1 accuracy was not good. The parameters of the Siamese network with classifiers were lower less than those of the Siamese network and triplet network. The recognition accuracy of the Siamese network with classifiers is the best of the three. Otherwise, for graph neural networks, the line sketch that is generated is too simple to convey the different relationships between the image patches after super-pixel segmentation.

In general, we proposed several methods for small datasets using the deep learning method. Although the recognition accuracy is not higher than that of most traditional machine learning methods, we employed restrictions, attention

modules, and graph data to learn the efficient features of the Siamese and triplet networks. We also tested our models on the CUFS dataset. However, the recognition accuracy did not reach 90% as the traditional method did. After analysis, the number of images was not enough for training. Moreover, the redundant parameter in the model may lead to difficulty in capturing the data, because of the huge number of calculations required.

Table 5-6 Recognition accuracy for Uom-SGFS(A) datasets

Methods	Top-1 accuracy	Top-10 accuracy
Improved Siamese network		64.15%
Siamese net with SVM	56.9%	
Siamese net with Random Forest	65.3%	
Siamese net with XGboost	63.9%	
Triplet net +Attention +SPP Layer	66.75%	90.46%
Siamese GCN(Quickshift)	74.16%	76.66%
Siamese MoNet (Quickshift)	64.17%	74.17%
Siamese GCN (SLIC)	68.33%	72.25%
Siamese MoNet (SLIC)	66.65%	73.33%

Table 5-7 Recognition accuracy for Uom-SGFS(B) datasets

	Top-1 accuracy	Top-10 accuracy
Improved Siamese network		81.74%
Siamese net with SVM	82.3%	
Siamese net with Random Forest	81.4%	
Siamese net with XGboost	47.1%	
Triplet net +Attention +SPP Layer	81.25%	90.56%
Siamese GCN(Quickshift)	65%	80.83%
Siamese MoNet (Quickshift)	62.5%	80%
Siamese GCN (SLIC)	60.83%	77.5%
Siamese MoNet (SLIC)	59.1%	79.17%

Because the number of images in the e-PRIP datasets is not enough for adequate training using the Siamese network and triplet network, the Top-1 accuracy for them are all not good. The parameters of the Siamese network with classifiers are less than the Siamese network and triplet network. The recognition accuracy for the Siamese network with classifiers is better than the others. Otherwise, for

graph neural networks, the line sketch is too simple to capture the different relationships between the image patches after superpixel segmentation.

Table 5-8 Recognition accuracy for e-PRIP datasets

	Top-1 accuracy	Top-10 accuracy
Improved Siamese network		85.33%
Siamese net with SVM	77.8%	
Siamese net with Random Forest	72.2%	
Siamese net with XGboost	80.6%	
Triplet net +Attention +SPP Layer	58.85%	84.60%
Siamese GCN (Quickshift)	55.28%	73.9%
Siamese MoNet (Quickshift)	50.4%	67.48%
Siamese GCN (SLIC)	47.15%	63.4%
Siamese MoNet (SLIC)	48.78%	61.78%
(Kazemi et al., 2018)	72.6%	
(Mittal et al., 2015)	52.0%	60.20%
DCNN (Galea and Farrugia, 2018)	54.9%	80.80%

Table 5-9 Experimental results on UoM-SGFS datasets

Method	Top-1 accuracy (UoM-SGFSA)	Top-10 accuracy (UoM-SGFSA)	Top-1 accuracy (UoM-SGFSB)	Top-10 accuracy (UoM-SGFSB)
Improved Siamese network		64.15%		81.74%
FaceNet	45.50%	50.70%	52.00%	80.10%
Triplet net +Attention +SPP Layer	66.75%	90.46%	81.25%	90.56%
Siamese GCN (Quickshift)	74.16%	76.66%	65%	80.83%
Siamese MoNet (Quickshift)	64.17%	74.17%	62.5%	80%
Siamese GCN (SLIC)	68.33%	72.25%	60.83%	77.5%
Siamese MoNet (SLIC)	66.65%	73.33%	59.1%	79.17%
(Peng et al., 2019)	64.80%	92.13%	72.53%	94.80%
(Galea and Farrugia, 2018)	31.60%	66.13%	52.17%	82.67%

Table 5-10 Experimental results on CUFS

Methods	Top-1 accuracy
Triplet net +Attention +SPP Layer	89.60%
Siamese GCN (Quick shift)	87.71%
Siamese MoNet (Quick shift)	85.9 %
Siamese GCN (SLIC)	82.4 %
Siamese MoNet (SLIC)	78.9%
(Wan et al., 2019)	92.56%

In general, we proposed several methods on small datasets using the deep learning method. Although the recognition accuracy is not higher than most traditional machine learning methods, we adopt restrictions, attention modules, and graph data representation to learn efficient features on the Siamese network and triplet network. We also test our models on the CUFS dataset. However, the recognition accuracy does not reach 90% achieved by the traditional methods. After analysis, the number of images in all datasets is not enough for training using deep learning methods. Our models adopt many convolution layers to extract the similar feature from cross-modal images. Thus, these models are all

complex to extract features. the number of the parameters is too much after training the model.

5.7. Conclusion

In this chapter, we presented a Siamese network based on graph structural data for facial photo-sketch recognition. This model constructed two graph convolution layers for each channel to learn a set of graphs on an embedding space. In order to reduce the modality gap between the facial photos and sketches, we used a super-pixel method on the contour images obtained from the HED model to extract similar structural graph data from the sketch and the corresponding photo. Experiments showed greater similarity between the graph data of the facial photos and of the sketches if we used the Quickshift method and not SLIC. We tested our methods on composite facial photo-sketch datasets and hand-drawn facial photo-sketch datasets. With the composite facial photo-sketch datasets, the Top-1 recognition accuracy for the UoM-SGFSA dataset was better than the state-of-the-art methods, reaching 74.16%. With the hand-drawn facial photo-sketch datasets, the performance was better than it was with the composite facial photo-sketch datasets. The graph convolution network is a connection model that captures graph dependencies through messages passing between graph nodes

Unlike standard neural networks, graph neural networks can represent information using neighbourhoods of any size. Meanwhile, GCN captures global information from the graph to represent a good nodal feature. In addition, the super-pixels methods that we adopted to build graphs reduced the redundant information of an image. Using the same graph for a facial photo and facial sketch reduces the impact of inconsistent features from the same position caused by cross modalities. For one thing, GCN uses transductive learning to update the state of each node. To obtain the graph embedding space, this method requires all the nodes to participate in the training stage. Thus, the generated graph needs to be aligned with the optimized node embedding. However, the representation of each node is affected by its relationship with other nodes. A node in the new graph which is generated by GCN needs to reconcile all the information from many related nodes at each forward transmission. This brings great computational demands, especially for a large graph. Moreover, in the GCN model, the convolution kernel does not have spatial localization. It does not calculate the sum of weights for the features on the centre node of the K^{th} neighbours after each convolution kernel.

Chapter6:

Conclusion

6.1. Conclusion

This thesis has focused on improving the recognition accuracy for facial photo-sketch recognition. One challenge of this project is the representation for face photo and face sketch are different. First, the captured photos show differences of illuminations, such as side lighting, top lighting, backlighting, and highlighting. Moreover, the illumination is different in different locations. Second, different facial photos may be closely similar. The human face is a kind of non-rigid model. Thus, the captured images of human faces are very different from different angles. Another challenge is that this project involved a kind of cross-modal image recognition. Although the descriptions of the cross-modal images referred to the same objects, the different objects or similar objects could be recognized as attributes of the same one, because of the pose, illumination, and angles. In the facial photo-sketch recognition project, a sketch is drawn by an artist in line with her/his experience and the descriptions of eyewitnesses. If the characteristics of the sketch and those drawn from human memory resemble each other closely, people can recognize the person in the sketch. However, due to memory gaps and the effects of time, sketches tend to offer little basis for recognition. To automatically identify the subject of a sketch from a facial photo dataset, we

designed three deep learning models to discover the complex mapping relationship between sketches and photos using information from the images.

In the first method, we designed a Siamese network that combined with a sparse encoder-decoder network to find the mapping of the joint features. First, artists have their own unique draughtsmanship, with the result that most convolution networks are difficult to train because of the limited numbers of images. We used the structure of the Siamese network for training. One advantage was that this architecture can learn abundant features from small dataset, due to the input of the Siamese network in which the image pairs consisted of a facial photo and a facial sketch. To train all the image pairs for recognition, we made the number of the input image pairs n^2 ,and made n the number of facial photos or facial sketches. Thus, this model increased the number of training datasets to avoid the problem of overfitting. Second, a sparse encoder-decoder network can learn effective HOG features without noise. Finally, instead of Euclidean distance to train the model, we adopted chi-square distance in the contrastive loss function. The experimental results show that our framework was adequate for a composite sketch dataset. Besides, it reduced the influence of overfitting by using data augmentation and modifying the network structure. Then we tried to combine it

with VGG-19 as a pre-trained model in the Siamese network to extract that are good for classification. We explored the performances obtained with three traditional learning algorithms (Support Vector Machine, Random Forest and XGBoost) combined with the Siamese network for training classifiers, based on features extracted using the Siamese network and other features obtained by means of the pre-trained model. The Random Forest displays high performance, especially on Uom-SGFS (B), it exceeds 80%.

Our second model was a novel triplet model. In spite of the abundant features extracted using deep learning methods, the limited number of datasets and the weak convergence of loss functions resulted in unsatisfactory recognition. In the Siamese networks, the extracted features of photos and their corresponding sketches are inconsistent because of the deformation of the photo images when they are turned into sketches. Thus, an attention model was designed on the image space to extract the features from the same location in the photo and the sketch, so that when the photo and sketch were mapped into a common feature space the cross-modal differences between them were reduced. The designed attention model consisted of a channel block and two separate spatial blocks for images and sketches separately. The first was designed to elicit the relationship

between the images of each channel and focused on extracting the shape of the input images. The other focused on extracting spatial information and textural features from the channel attention layer. Moreover, a spatial pyramid pooling layer was introduced into the network to deal with images of different sizes. Our proposed solution was tested on composite facial photo-sketch datasets, and it performed better than the state-of-the-art results. Set B in UoM-SGFS dataset, in particular, scored more than 81%.

The third method was based on graph convolution neural networks. We used the image detection (HED) and super-pixel methods to draw a graph from an image. HED is used to reduce the influence of noise. This method extracts an accurate edge prediction map using continuous integration and learning. After this, we used a super-pixel method to cluster irregular pixel blocks with a certain visual meaning as nodes of a graph, such as textural features, colour, and illumination. In order to reduce complex computation, super-pixel methods use only a small number of super-pixels instead of a great many pixels to delineate the features of an image. Each segmented super-pixel region is taken as a node, and each pair of adjacent regions forms an edge of the graph. After this, we built a graph convolution network (GCN) based on the Siamese architecture. A GCN updates

the state of a node by aggregating the vectors of its neighbours' nodal features using convolution operations. The representation of the node captures structural information from the neighbours in the k-hop network after several iterations. A GCN keeps the image's features and captures the semantic relationship between facial attributes. However, the hidden representation of each node tends to converge to the same node after several layers for training. Thus, we adopted a two-layer GCN model for each channel in our Siamese network. Experiments showed that the GCN performed well on several facial photo-sketch datasets, both seen and unseen. We also showed that the model performance based on the graph structural representation of the data using the Siamese GCN was more stable than a model performance using the Siamese CNN model.

6.2. Future work

This thesis proposed four methods for facial photo-sketch datasets using the attention mechanism. Because the dataset was too small, we proposed a self-attention module to compress the model. However, the parameters of this module are too large to train a suitable model for face photo-sketch dataset. And the deformation between the photo and the sketch leads to that the extracted features are difficult to utilize on recognition directly. One idea would be to develop a light

attention module for making the number of parameters smaller than that in our proposed module. Meanwhile, this attention module can reduce the deformation effect using distortion correction.

Colour sketches tend to be more common than line sketches and hand-drawn sketches. Colour information should be an important feature of recognition, but the brightness of colours is heavily influenced by illumination. Devising a CNN model that could reduce the effect of illumination would be a good way to increase recognition accuracy.

An improved GCN model based on the one in the present study would describe the best relationship between each node. GCN can capture the features that are more conducive to recognition. These representative features increase the recognition accuracy than the features from the CNN model. In addition, GCN model is easier to learning each node' features from face photos and face sketches respectively, after designing a new weight strategy for training,

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. 34, 2274–2282.

2. Ahonen, T., Hadid, A., Pietikäinen, M., 2004. Face recognition with local binary patterns, in: European Conference on Computer Vision. Springer, pp. 469–481.

3. Alex, A.T., Asari, V.K., Mathew, A., 2013. Local difference of gaussian binary pattern: robust features for face sketch recognition, in: 2013 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, pp. 1211–1216.

4. Ba, J., Mnih, V., Kavukcuoglu, K., 2014. Multiple object recognition with visual attention. ArXiv Prepr. ArXiv14127755.

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

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

7. Bhatt, H.S., Bharadwaj, S., Singh, R., Vatsa, M., 2012. Memetic approach for matching sketches with digital face images.

8. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R., 1994. Signature verification using a" siamese" time delay neural network, in: Advances in Neural Information Processing Systems. pp. 737–744.

9. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y., 2013. Spectral networks and locally connected networks on graphs. ArXiv Prepr. ArXiv13126203.

10. Cao, B., Wang, N., Gao, X., Li, J., Li, Z., 2020. Multi-margin based decorrelation learning for heterogeneous face recognition. ArXiv Prepr. ArXiv200511945.

11. Chelali, F.Z., Djeradi, A., Djeradi, R., 2009. Linear discriminant analysis for face recognition, in: 2009 International Conference on Multimedia Computing and Systems. IEEE, pp. 1–10.

12. Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

13. Chen, Z., Yao, S., Jia, Y., Liu, C., 2018. Face sketch-photo synthesis and recognition: Dual-scale Markov Network and multi-information fusion. J. Vis. Commun. Image Represent. 51, 112–121.

14. Cheng, Z., Zhu, X., Gong, S., 2018. Low-resolution face recognition, in: Asian Conference on Computer Vision. Springer, pp. 605–621.

15. Chopra, S., Hadsell, R., LeCun, Y., 2005. Learning a similarity metric discriminatively, with application to face verification, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference On. IEEE, pp. 539–546.

16. Chugh, T., Singh, M., Nagpal, S., Singh, R., Vatsa, M., 2017. Transfer learning based evolutionary algorithm for composite face sketch recognition, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference On. IEEE, pp. 619–627.

17. Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference On. IEEE, pp. 886–893.

18. De Ridder, D., Kouropteva, O., Okun, O., Pietikäinen, M., Duin, R.P., 2003. Supervised locally linear embedding, in: Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003. Springer, pp. 333–341.

19. Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering, in: Advances in Neural Information Processing Systems. pp. 3844–3852.

20. Galea, C., Farrugia, R.A., 2018. Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning. IEEE Trans. Inf. Forensics Secur. 13, 1421–1431.

21. Galea, C., Farrugia, R.A., 2017. Matching software-generated sketches to face photographs with a very deep CNN, morphed faces, and transfer learning. IEEE Trans. Inf. Forensics Secur. 13, 1421–1431.

22. Galea, C., Farrugia, R.A., 2016. A large-scale software-generated face composite sketch database, in: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of The. IEEE, pp. 1–5.

23. Galoogahi, H.K., Sim, T., 2012a. Inter-modality face sketch recognition, in: 2012 IEEE International Conference on Multimedia and Expo. IEEE, pp. 224–229.

24. Galoogahi, H.K., Sim, T., 2012b. Inter-modality face sketch recognition, in: Multimedia and Expo (ICME), 2012 IEEE International Conference On. IEEE, pp. 224–229.

25. Gao, S., Zhang, Yuting, Jia, K., Lu, J., Zhang, Yingying, 2015. Single sample face recognition via learning deep supervised autoencoders. IEEE Trans. Inf. Forensics Secur. 10, 2108–2118.

26. Gao, X., Zhong, J., Li, J., Tian, C., 2008. Face sketch synthesis algorithm based on E-HMM and selective ensemble. IEEE Trans. Circuits Syst. Video Technol. 18, 487–496.

27. Garcia, V., Bruna, J., 2017. Few-shot learning with graph neural networks. ArXiv Prepr. ArXiv171104043.

28. Gauthier, J., 2014. Conditional generative adversarial nets for convolutional face generation. Cl. Proj. Stanf. CS231N Convolutional Neural Netw. Vis. Recognit. Winter Semester 2014, 2.

29. Gerbrands, J.J., 1981. On the relationships between SVD, KLT and PCA. Pattern Recognit. 14, 375–381.

30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems. pp. 2672–2680.

31. Güçlütürk, Y., Güçlü, U., van Lier, R., van Gerven, M.A., 2016. Convolutional sketch inversion, in: European Conference on Computer Vision. Springer, pp. 810–824.

32. Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference On. IEEE, pp. 1735–1742.

33. Han, H., Klare, B.F., Bonnen, K., Jain, A.K., 2013. Matching composite sketches to face photos: A component-based approach. IEEE Trans. Inf. Forensics Secur. 8, 191–204.

34. Han, S., Lee, I.-Y., Ahn, J.-H., 2016. Two-Dimensional Joint Bayesian Method for Face Verification. J. Inf. Process. Syst. 12.

35. He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

36. He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37, 1904–1916.

37. He, R., Wu, X., Sun, Z., Tan, T., 2017. Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition. ArXiv Prepr. ArXiv170802412.

38. Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. ArXiv Prepr. ArXiv170307737.

39. Hoffer, E., Ailon, N., 2015. Deep metric learning using triplet network, in: International Workshop on Similarity-Based Pattern Recognition. Springer, pp. 84–92.

40. Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

41. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E., 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.

42. Huang, G.-B., Wang, D.H., Lan, Y., 2011. Extreme learning machines: a survey. Int. J. Mach. Learn. Cybern. 2, 107–122.

43. Huang, S., 2016. Evaluation and Comparison of Cognitec's FaceVACS, Human Age Estimation and The State-Of-The-Art.

44. Huang, X., Lei, Z., Fan, M., Wang, X., Li, S.Z., 2013. Regularized discriminative spectral regression method for heterogeneous face matching. IEEE Trans. Image Process. 22, 353–362.

45. Huo, J., Gao, Y., Shi, Y., Yang, W., Yin, H., 2017. Heterogeneous Face Recognition by Margin-Based Cross-Modality Metric Learning. IEEE Trans. Cybern.

46. Iranmanesh, S.M., Dabouei, A., Kazemi, H., Nasrabadi, N.M., 2018. Deep Cross Polarimetric Thermal-to-visible Face Recognition. ArXiv Prepr. ArXiv180101486.

47. Jaderberg, M., Simonyan, K., Zisserman, A., 2015. Spatial transformer networks, in: Advances in Neural Information Processing Systems. pp. 2017–2025.

48. Jiao, L., Zhang, S., Li, L., Liu, F., Ma, W., 2018. A modified convolutional neural network for face sketch synthesis. Pattern Recognit. 76, 125–136.

49. Johnson, J., Alahi, A., Fei-Fei, L., n.d. Perceptual Losses for Real-Time Style Transfer and Super-Resolution: Supplementary Material.

50. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X., 2012. Multi-view discriminant analysis, in: European Conference on Computer Vision. Springer, pp. 808–821.

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

52. Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874.

53. Khalil-Hani, M., Sung, L.S., 2014. A convolutional neural network approach for face verification, in: 2014 International Conference on High Performance Computing & Simulation (HPCS). IEEE, pp. 707–714.

54. Kiani Galoogahi, H., Sim, T., 2012. Face sketch recognition by Local Radon Binary Pattern: LRBP.

55. Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. ArXiv Prepr. ArXiv160902907.

56. Klare, B., Jain, A.K., 2010. Sketch-to-photo matching: a feature-based approach, in: SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, pp. 766702–766702.

57. Klare, B.F., Jain, A.K., 2013. Heterogeneous face recognition using kernel prototype similarities. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1410–1422.

58. Klum, S.J., Han, H., Klare, B.F., Jain, A.K., 2014. The FaceSketchID system: Matching facial composites to mugshots. IEEE Trans. Inf. Forensics Secur. 9, 2248–2263.

59. Knyazev, B., Lin, X., Amer, M.R., Taylor, G.W., 2019a. Image Classification with Hierarchical Multigraph Networks. ArXiv Prepr. ArXiv190709000.

60. Knyazev, B., Lin, X., Amer, M.R., Taylor, G.W., 2019b. Image Classification with Hierarchical Multigraph Networks. ArXiv Prepr. ArXiv190709000.

61. Koch, G., Zemel, R., Salakhutdinov, R., 2015. Siamese neural networks for one-shot image recognition, in: ICML Deep Learning Workshop.

62. Kukharev, G., Matveev, Y., Forczmański, P., 2016. An Approach to Improve Accuracy of Photo–to–Sketch Matching, in: International Conference Image Analysis and Recognition. Springer, pp. 385–393.

63. Laws, D.R., 2020. Offender Classification and Registration, in: A History of the Assessment of Sex Offenders: 1830–2020. Emerald Publishing Limited.

64. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G., 2016. Labeled faces in the wild: A survey, in: Advances in Face Detection and Facial Image Analysis. Springer, pp. 189–248.

65. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

66. Lei, Z., Li, S.Z., 2009. Coupled spectral regression for matching heterogeneous faces, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On. IEEE, pp. 1123–1128.

67. Lei, Z., Liao, S., Jain, A.K., Li, S.Z., 2012. Coupled discriminant analysis for heterogeneous face recognition. IEEE Trans. Inf. Forensics Secur. 7, 1707–1716.

68. Li, K., Wu, Z., Peng, K.-C., Ernst, J., Fu, Y., 2018. Tell me where to look: Guided attention inference network. ArXiv Prepr. ArXiv180210171.

69. Li, S., Yi, D., Lei, Z., Liao, S., 2013. The casia nir-vis 2.0 face database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 348–353.

70. Li, S.Z., Lei, Z., Ao, M., 2009. The HFB face database for heterogeneous face biometrics research, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 1–8.

71. Li, Y., Gu, C., Dullien, T., Vinyals, O., Kohli, P., 2019. Graph matching networks for learning the similarity of graph structured objects. ArXiv Prepr. ArXiv190412787.

72. Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

73. Lindeberg, T., 2012. Scale invariant feature transform.

74. Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S., 2005. A nonlinear approach for face sketch synthesis and recognition, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, pp. 1005–1010.

75. Liu, W., Wen, Y., Yu, Z., Yang, M., 2016. Large-margin softmax loss for convolutional neural networks., in: ICML. p. 7.

76. Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137.

77. Luettgen, M.R., Karl, W.C., Willsky, A.S., Tenney, R.R., 1993. Multiscale representations of Markov random fields. IEEE Trans. Signal Process. 41, 3377–3396.

78. Martinez, A.M., Kak, A.C., 2001. Pca versus lda. IEEE Trans. Pattern Anal. Mach. Intell. 23, 228–233.

79. Melekhov, I., Kannala, J., Rahtu, E., 2016. Siamese network features for image matching, in: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 378–383.

80. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G., 1999. XM2VTSDB: The extended M2VTS database, in: Second International Conference on Audio and Video-Based Biometric Person Authentication. pp. 965–966.

81. Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. ArXiv Prepr. ArXiv14111784.

82. Mittal, P., Jain, A., Goswami, G., Singh, R., Vatsa, M., 2014. Recognizing composite sketches with digital face images via SSD dictionary, in: IEEE International Joint Conference on Biometrics. IEEE, pp. 1–6.

83. Mittal, P., Vatsa, M., Singh, R., 2015. Composite sketch recognition via deep network-a transfer learning approach, in: Biometrics (ICB), 2015 International Conference On. IEEE, pp. 251–256.

84. Mnih, V., Heess, N., Graves, A., 2014. Recurrent models of visual attention, in: Advances in Neural Information Processing Systems. pp. 2204–2212.

85. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M., 2017. Geometric deep learning on graphs and manifolds using mixture model cnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5115–5124.

86. Oh, B.-S., Oh, K., Teoh, A.B.J., Lin, Z., Toh, K.-A., 2017. A Gabor-based network for heterogeneous face recognition. Neurocomputing.

87. Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition., in: Bmvc. p. 6.

88. Peng, C., Gao, X., Wang, N., Li, J., 2017. Graphical representation for heterogeneous face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 39, 301–312.

89. Peng, C., Gao, X., Wang, N., Tao, D., Li, X., Li, J., 2016a. Multiple Representations-Based Face Sketch–Photo Synthesis. IEEE Trans. Neural Netw. Learn. Syst. 27, 2201–2215.

90. Peng, C., Wang, N., Gao, X., Li, J., 2016b. Face Recognition from Multiple Stylistic Sketches: Scenarios, Datasets, and Evaluation, in: European Conference on Computer Vision. Springer, pp. 3–18.

91. Peng, C., Wang, N., Li, J., Gao, X., 2019. DLFace: Deep local descriptor for cross-modality face recognition. Pattern Recognit. 90, 161–171.

92. Qi, Y., Song, Y.-Z., Zhang, H., Liu, J., 2016. Sketch-based image retrieval via Siamese convolutional neural network, in: Image Processing (ICIP), 2016 IEEE International Conference On. IEEE, pp. 2460–2464.

93. Radman, A., Suandi, S.A., 2018. Robust face pseudo-sketch synthesis and recognition using morphological-arithmetic operations and HOG-PCA. Multimed. Tools Appl. 1–22.

94. Roth, V., Steinhage, V., 2000. Nonlinear discriminant analysis using kernel functions, in: Advances in Neural Information Processing Systems. pp. 568–574.

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

96. Saxena, S., Verbeek, J., 2016. Heterogeneous face recognition with cnns, in: Computer Vision–ECCV 2016 Workshops. Springer, pp. 483–491.

97. Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 815–823.

98. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

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

100. Shawe-Taylor, J., Cristianini, N., 2000. Support vector machines. Introd. Support Vector Mach. Kernel-Based Learn. Methods 93–112.

101. Shi, H., Wang, X., Yi, D., Lei, Z., Zhu, X., Li, S.Z., 2017. Cross-modality Face Recognition via Heterogeneous Joint Bayesian. IEEE Signal Process. Lett.

102. Sun, Y., Chen, Y., Wang, X., Tang, X., 2014a. Deep learning face representation by joint identification-verification, in: Advances in Neural Information Processing Systems. pp. 1988–1996.

103. Sun, Y., Liang, D., Wang, X., Tang, X., 2015a. Deepid3: Face recognition with very deep neural networks. ArXiv Prepr. ArXiv150200873.

104. Sun, Y., Wang, X., Tang, X., 2015b. Deeply learned face representations are sparse, selective, and robust, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2892–2900.

105. Sun, Y., Wang, X., Tang, X., 2014b. Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1891–1898.

106. Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1701–1708.

107. Tang, X., Wang, X., 2004. Face sketch recognition. IEEE Trans. Circuits Syst. Video Technol. 14, 50–57.

108. Tang, X., Wang, X., 2002. Face photo recognition using sketch, in: Proceedings. International Conference on Image Processing. IEEE, p. I–I.

109. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, \Lukasz, Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems. pp. 5998–6008.

110. Vedaldi, A., Soatto, S., 2008. Quick shift and kernel methods for mode seeking, in: European Conference on Computer Vision. Springer, pp. 705–718.

111. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., 2016. Matching networks for one shot learning, in: Advances in Neural Information Processing Systems. pp. 3630–3638.

112. Wan, W., Gao, Y., Lee, H.J., 2019. Transfer deep feature learning for face sketch recognition. Neural Comput. Appl. 1–10.

113. Wang, C., Zhang, X., Lan, X., 2017. How to train triplet networks with 100k identities?, in: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 1907–1915.

114. Wang, F., Xiang, X., Cheng, J., Yuille, A.L., 2017. Normface: L2 hypersphere embedding for face verification, in: Proceedings of the 25th ACM International Conference on Multimedia. pp. 1041–1049.

115. Wang, N., Tao, D., Gao, X., Li, X., Li, J., 2013. Transductive face sketch-photo synthesis. IEEE Trans. Neural Netw. Learn. Syst. 24, 1364–1376.

116. Wang, X., Tang, X., 2009. Face photo-sketch synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31, 1955–1967.

117. Wang, X., Tang, X., 2004. Random sampling LDA for face recognition, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference On. IEEE, p. II–II.

118. Wang, Z., Zheng, L., Li, Y., Wang, S., 2019. Linkage based face clustering via graph convolution network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1117–1125.

119. Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision. Springer, pp. 499–515.

120. Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemom. Intell. Lab. Syst. 2, 37–52.

121. Woo, S., Park, J., Lee, J.-Y., So Kweon, I., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19.

122. Wu, R., Kamata, S., Breckon, T., 2017. Face recognition via deep sparse graph neural networks, in: BMVCW.

123. Wu, X., Song, L., He, R., Tan, T., 2017. Coupled Deep Learning for Heterogeneous Face Recognition. ArXiv Prepr. ArXiv170402450.

124. Xie, S., Tu, Z., 2015. Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 1395–1403.

125. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning. pp. 2048–2057.

126. Yi, D., Lei, Z., Li, S.Z., 2015. Shared representation learning for heterogenous face recognition, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops On. IEEE, pp. 1–7.

127. Yin, W., Schütze, H., Xiang, B., Zhou, B., 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Trans. Assoc. Comput. Linguist. 4, 259–272.

128. Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., Hospedales, T.M., 2017. Sketch-a-net: A deep neural network that beats humans. Int. J. Comput. Vis. 122, 411–425.

129. Zagoruyko, S., Komodakis, N., 2015. Learning to compare image patches via convolutional neural networks, in: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference On. IEEE, pp. 4353–4361.

130. Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23, 1499–1503.

131. Zhang, L., Lin, L., Wu, X., Ding, S., Zhang, Lei, 2015. End-to-end photo-sketch generation via fully convolutional representation learning, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, pp. 627–634.

132. Zhang, S., Gao, X., Wang, N., Li, J., Zhang, M., 2015. Face sketch synthesis via sparse representation-based greedy search. IEEE Trans. Image Process. 24, 2466–2477.

133. Zhang, W., Shu, Z., Samaras, D., Chen, L., 2017. Improving heterogeneous face recognition with conditional adversarial networks. ArXiv Prepr. ArXiv170902848.

134. Zhang, W., Wang, X., Tang, X., 2011a. Coupled information-theoretic encoding for face photo-sketch recognition, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On. IEEE, pp. 513–520.

135. Zhang, W., Wang, X., Tang, X., 2011b. Coupled information-theoretic encoding for face photo-sketch recognition, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On. IEEE, pp. 513–520.

136. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y., 2017. Range loss for deep face recognition with long-tailed training data, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 5409–5418.

137. Zhou, H., Kuang, Z., Wong, K.-Y.K., 2012. Markov weight fields for face sketch synthesis, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On. IEEE, pp. 1091–1097.

138. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. ArXiv Prepr. ArXiv170310593.