Semi-Automatic Generation of Users' Desired Face Image

(ユーザが望む顔画像の半自動生成)

山梨大学大学院

医工農学総合教育部

博士課程学位論文

2020 年 9 月

許　彩娥(XU CAIE)

Abstract

Face image synthesis has many potential applications including public safety, such as video surveillance and law enforcement. For example, one important application is in assisting the police to create the face image of suspects based on the memories of witnesses or victims. However, drawing an image based on descriptions of what is in one's mind is not an easy task for the majority of people. To synthesize user's desired face image, this thesis introduces three methods for this goal in a semi-automatic way. While the first method uses traditional hand-crafted feature, the second method employs the state-of-the-art deep neural network technique, and the third method proposes a novel generative neural network model to enhance the detailed features of faces.

The first method features a user-friendly system that can create a facial image based on user's feedback. It can synthesize a user desired face image without questioning the user on the explicit features of the face in his or her mind. Through a dialogic approach based on a relevance feedback strategy to translate facial features into input, the user only needs to look at several candidate facial images and judge whether each image resembles the face that he or she is imagining. The experimental results show that the proposed technique succeeded in generating images resembling a face a user had imagined or memorized. However, there are some disadvantages under this method. The result sometimes is blurring and doesn't look like the desired face exactly. Furthermore, it cannot synthesize color face image. Such drawbacks are mainly due to the feature representation. Another problem of this technique is that it fails to generate face image of completely new features inherently because of its synthesis algorithm.

Recently, with the rapid development of deep learning technology, various research areas and applications, such as computer vision, robotics, big data analysis, and pilotless automobiles, have achieved major advancements. The field of face image generation and synthesis is no exception, as it has also undergone significant developments. It especially benefits from the emergence of the Generative Adversarial Network (GAN), which is a type of neural network architecture for the generative purpose. The second method employs landmark face representation to improve feature space and uses GAN technology to compensate for the low

image quality of the first method. It also allows the users to generate their desired face images in a semi-automatic way. The second method introduces a novel algorithm to create completely new face image. The second method can take full advantage of the high image quality while compensating for the lack of user intervention of state-of-the-art GAN technology. The experiment results demonstrated that the second method can generate image with much higher quality than that of the first method.

Although most of the created images by the second method can well resemble the geometric features of the user's desired face, the resulting images fail to preserve the detail texture features of faces. For example, the wrinkles on the faces were not reproduced in the resulting images. Face texture features are an important personal characteristic, especially wrinkle is an important feature which is closely related to the person's age. GAN-based methods for face aging also have received large attention thanks to its advantage of being able to generate exceptional realistic images. But these models mainly rely on age label or age group label. In many cases, the label is not corresponding to the real age, it is vague and inaccurate.

To the best of my knowledge, none of the existing GAN can generate face images that preserve texture features well. To generate a face image with the user's desired texture information, as the third method, a novel framework called High-Frequency Generative Adversarial Network (HF-GAN) was proposed for synthesizing face image with texture details. High-frequency features of face image are extracted using technology of edge detection and then drawn on black background to create a high-frequency feature image. The trained generator samples face image from high-frequency feature image during runtime. For training stage, the model is guided by adversarial loss, classification loss, reconstruction loss, and perceptual loss. Further, attention mechanism was added to the generator and discriminator in order to enhance the features at face area when generating face image. The results show that the proposed method can generate face images from high-frequency features and can produce the user's desired texture information to a certain degree.

In summary, this thesis proposed a semi-automatic approach to face synthesis. The first method proposed the idea of generating face image in the user's mind based on relevance feedback. Through a dialogic way, the user only needs to look at several candidate face image and judge them according to similarity. The second

method improved the image quality by using the state-of-the-art deep learning technology. To solve the shortcomings in the second method, a generative neural network model was explored. The proposed third method can generate face images that capture both geometry and texture information and can produce the user's desired texture information to a certain degree.

Keywords: Face image synthesis, face image generation, relevant feedback, optimum-Path forest, deep learning, conditional generative adversarial network, landmark, face aging, self-attention.

Table of Contents

Chapter 1 Introduction

1.1 Background

Face image synthesis has been wildly used in various applications including identity matching and confirmation, law enforcement, entertainment, and so on, especially, in public safety, such as video surveillance and law enforcement. For example, identifying criminals benefit from the automatic face synthesis technique for creating the suspect portrait according to the description of the eyewitness. Moreover, a similar technique can be used for giving concrete form to imagined ideas of romantic 'types' and translate other imagined faces into explicit images. However, drawing an image based on descriptions of what is in one's mind is not an easy task for the majority of people. There are a great many studies for face synthesis in the past few decades including component-based [1-9], sketch-based [10, 11], iteration-based [12-16], and deep-learning based methods. The most representative one is Generative Adversarial Network (GAN)-based methods [17-32]. Motivated by the above-mentioned potential applications, three methods were developed to synthesize the image of a face from a user's imagination and memory through some simple user interactions and a generative model based on GAN was explored to improve the quality of resulting images.

1.2 Based on Principal Component Analysis (PCA)

Although the montage approach to face image synthesis [1-9] allows users to create face images by selecting face components in sequence, it involves the time-consuming task of choosing the right parts from a wide array of options. It is known that the composition of face parts is a more important factor in the perception of a face than the individual parts [10][11]. However, it can be very difficult to adjust the positions of individual parts to achieve a desired composition. Several methods have been developed for synthesising face images according to sketches [10][11]. Such methods, however, require the user to provide a sketch, which is not always a possibility and are still difficult for average person.

Motivated by the above-mentioned potential applications and the limitation of montage and sketch-based synthesis technologies, a system that can generate an image of a face from a user's imagination and memory based on the user's feedback was developed. In this system, a set of example images are used to train an Optimum-Path Forest (OPF) [33] algorithm to classify the face images based on their relevance to the face in the user's mind. The training process is conducted through a relevance feedback approach. All the user need to do under this method is to indicate whether the image of the face shown bears a general resemblance to the face that he or she is imagining, thereby eliminating the need to evaluate individual parts and features separately (as is the

case with the montage approach) or visualise or verbalise specific characteristics (as is the case with caricatures). The details of this method will be described in Chapter 3.

1.3 Based on Conditional Generative Adversarial Network

Recently, with the rapid development of deep learning technology, various research areas and applications, such as computer vision, robotics, big data analysis, and pilotless automobiles, have achieved major advancements. The field of face image generation and synthesis is no exception, as it has also undergone significant developments. In particular, the emergence of the generative adversarial network (GAN), which is a type of neural network architecture for the generative model first proposed by Goodfellow et al. in 2014 [17], brought about a major breakthrough in the field of face image generation. GAN consists of two networks: the generator that creates as realistic data as possible and the discriminator that attempts to distinguish fake samples from real ones. The two networks compete with each other during the training process, resulting in a generator that can produce realistic data.

Since the very first GAN model [17] demonstrated its ability to generate face images, various improved models have been developed. The face images generated with [17] are fair random draws, not cherry-picked, and are poor-quality grayscale images. To gain some control over the generated results, Mehdi and Simon proposed the conditional generative adversarial network (CGAN) in the same year, which allows inputting a condition to the model in addition to the noise [20]. This model set a solid foundation for the emergence of various variants of GAN. In 2015, Jon Gauthier *et al.* proposed the use of CGAN for convolutional face generation [21], which added to CGAN the capability of generating face images with specific attributes, such as race, age, and emotion, by varying the conditional information. Grigory *et al.* proposed the Age-GAN [22] for automatically simulating face aging based on CGAN; Age-GAN particularly emphasizes the preservation of the original person's identity in the aged version of his/her face image. The Two-Pathway Generative Adversarial Network (TP-GAN) [34] was proposed by Rui et al. in 2017 for realistic face synthesis. It is mainly used for reconstructing face images from a partial view corresponding to different poses. The most recent variant, Style-GAN [29], which was proposed by Tero *et al*., led to an automatically learned, unsupervised separation of high-level attributes, and it can synthesize high-quality face images with varying high-level attributes, such as different hairstyles and expressions. However, it considers stochastic variation and does not have the ability to control such attributes. Xing et al. focused on high-level face-related analysis tasks and proposed gender-preserving GAN (GP-GAN) [27], which could synthesize corresponding face images from landmarks; the feature points represent the geometric information of the overall shape and the individual parts of the face. Although controlling the geometric features, such as the pose, the shape of the face, and individual facial parts, is possible with GP-GAN, it requires the landmarks as input. The application of GP-GAN is therefore limited without providing users a method to create the landmarks of their desired face.

Bontrager *et al.* [32] proposed an approach based on Wasserstein GAN [18] and interactive evolutionary computation [15] to produce an image resembling a given target. The user is asked to evaluate a set of images resulting from GAN, and a genetic algorithm is used to modify the latent vector based on the user's evaluation. This is the first work that demonstrated the potential of using the evolutionary algorithm to generate face images similar to the target faces. However, their evaluation experiment reported that the average score of the results was only 2.2 out of 5. Furthermore, the method cannot provide control over detailed facial features.

The experimental results from PCA based method, which is proposed in the first method, show that the proposed method succeeded in generating images resembling a face a user had imagined or memorized. However, the result sometimes is blurring and doesn't look like the desired face exactly. The proposed method cannot synthesize colour face image. Such drawbacks are mainly due to the feature representation. A global feature space based on PCA is employed, which fails to capture the personal detail well, causing the generated face quite similar to the average face. Another problem of this technique is that the results are synthesized from the linear interpolation of the top K of user favored face images and it fails to generate face image of completely new feature. Taking into account those mentioned disadvantages of the method based on PCA, this thesis explored the potential of using generative neural network model to improve the face representation.

To the best of my knowledge, none of the existing GAN models can provide users with easy control over detailed facial features, such as the shapes and positions of individual parts of the face. As mentioned at the beginning of the thesis, the ability to control detailed facial features is required in many applications. In Chapter 4, a method combining Gender Preserving Generative Adversarial Network (GP-GAN) and relevance feedback was proposed for interactive face image generation. Similar to the first method, an optimum-path forest (OPF) classifier is used to define the desired facial features represented as landmarks which are face geometric information and are used to localize and represent salient regions of the face, such as overall shape, eyes, eyebrows, nose, and mouth. The classifier is iteratively updated based on the relevance feedback of users. The landmarks of the desired face are then used as the input to GP-GAN to generate realistic face images with the desired features. In this way, the proposed method can take full advantage of the high image quality while compensating for the lack of user intervention of state-of-the-art GAN technology. Experiment showed that the proposed method can generate a result similar to the target face in the user's memory, and has higher quality compared to the results generated with the first method.

1.4 Generative Adversarial Network for synthesizing faces with the user's desired texture features from high frequency features

Although most of the resulting images generated with the second method can well resemble the geometric features of the reference images, the generated images fail to preserve the details of texture features. For example, the wrinkles on the faces were not reproduced in the resulting images.

Face texture features are an important personal characteristic, especially wrinkle. It is closely related to the person's age. There are many applications related to face aging, such as cross-age face recognition, finding lost children, biometrics, cosmetology and so on. There have been large efforts in this field in the last two decades. Face aging, also called age synthesis or age progression, is defined as the rendering of a face image aesthetically with natural aging and rejuvenating effects on the individual face [30]. Age estimation means to label a face automatically with the exact age (years) or the age group (age range) of the individual face [35].

GAN-based methods for face aging simulation have received a lot of attention motived by the fact that GAN can generate exceptional realistic images. Perarnau *et al.* [30] reconstructs and modifies real image of faces conditioning on arbitrary attributes. Antipov *et al.* [31] proposed an Age-cGAN for automatic face aging. It can preserve the original person's identity in the aged version of his/her face. Wang *et al.* [22] proposed an IPCGANs framework, which synthesizes a face lies in given age group instead of a specific age. These previous GAN-based models fail to captures long-range dependencies of entire image since they rely heavily on convolution. To fix this problem, Cheng *et al.* [36] proposed Self-attention technology. Zhang *et al.* [37] developed Self-Attention GAN (SAGAN) by introducing self-attention mechanism into convolutional GAN to generate high-resolution details from feature locations. All existing models, however, mainly rely on age labels or age group labels. In many cases, the age labels do not correspond to the real age, it is vague and inaccurate.

As the third method proposed by the thesis, In Chapter 5, High-Frequency Generative Adversarial Network (HF-GAN) is introduced for synthesizing faces with the user's desired texture features. Meanwhile, to preserve semantic and perceptual characteristics of real image, perceptual loss was introduced for guiding model. Besides, a self-attention mechanism was added to HF-GAN model so as to focus on the face features when generating the image. By inserting self-attention into the generator and discriminator, the former can generate face images with fine details at every location, the later can more accurately distinguish the real data and generated images.

1.5 Relationship between the proposed methods

To generate the face image that is a face image in the user's memory or imagination, three methods are proposed. In the first method, the basic concept of using relevance feedback approach is proposed. By applying

PCA to extract face features from the training dataset, global feature space is constructed. OPF classifier is employed for quickly retrieving the best nodes that reflect the user's feedback. A new candidate is computed by interpolating top k best nodes. A final face image is synthesized from the principal components. With the first method, the results have some quality problems such as blurring. It can only synthesize grayscale face image and fails to a create completely new face image since the result is synthesized from the linear interpolation of the top k user favored face images. To overcome these disadvantages, the second method improves the first method from three aspects: face representation, algorithm for creating candidates, and face image generation. The second method employed landmark features for face representation. New candidates are created by moving the best node alone a direction vector toward the user's desired face image. Advanced deep learning technology, GAN model, is used for face generation to achieve high quality result. The second method largely improves the quality of the results, can generate colour face images, and can create new face images that is not in training dataset. However, it fails to capture the details of texture features such as wrinkles. Aim to generate face images with the user's desired texture features, the third method explored a generative model. Meanwhile, it can preserve semantic and perceptual characteristics of real image by introducing perceptual loss. And by adding a self-attention mechanism, the face features can be more clearly captured when generating the image.

1.6 Structure of the thesis

This thesis is organized as follows.

In Chapter 1, the research background and research goals were briefly introduced. Then, I concisely explain the three methods proposed to achieve this target. Finally, the structure of the thesis is described.

In Chapter 2, the related works of face image generation, including traditional methods and deep learning approaches, are reviewed. I will clarify the relationship between my methods and those existing methods, trying to answer why the three methods are necessary; how to take full advantage of related technologies; how to overcome the disadvantages; how to make the improvement.

In Chapter 3, the first method, synthesising user' desired face image based on PCA, is explained in detail. The method includes three major components: extracting primary features, training an OPF classifier based on relevance feedback, and synthesising face images. The feature space is constructed by applying PCA to 1,000 sample images. An Optimum-Path Forest (OPF) classifier is then dynamically trained based on the user's feedbacks. Based on the trained OPF classifier, the candidates of the user's desired face images are retrieved and interpolated to synthesize a new face images supposed to be the user imagined one.

In Chapter 4, I make a detailed introduction to the second method: generating users' desired face image using CGAN and relevance feedback. Compare to the first method, two core parts have been improved: the feature representation and the technology of face generation. Besides, I also proposed a novel algorithm for creating the new candidate landmarks of the desired face. The desired face image is generated from CGAN given the new candidate landmarks.

In Chapter 5, I propose a method for generating face images with user's desired texture from high-frequency features. The purpose is to simulate the aging through synthesising face with wrinkles rather than with specific age tags. A novel model called HF-GAN was designed. The self-attention mechanism is added to GAN model. It is guided by adversarial loss, classification loss, reconstruction loss, and perceptual loss to enhance the texture features related to ages.

In Chapter 6, I conclude and summarize the whole thesis. Also, the future work of the research will be discussed.

Chapter 2 Related Works

While a large number of works have focused on facial recognition and identification, to the best of the authors' knowledge, there are few studies that have been conducted on the synthesis of face images. Facial image synthesis is most prominently used in the field of law enforcement, such as generating the suspect's face image based on the description of the eyewitness or crime victim. This work was originally performed by the artist who creates suspect's face image by drawing or sketching after consulting with the witness or crime victim, and hence the results very much rely on the skills of professionals. In the last two decades, a number of computer-based systems for face synthesis have been developed. The most widely used systems are: FACES [1] and Identikit [2] in US, and E-FIT [3] and PRO-fit [4] in UK. Early systems mainly used component-based approach. For example, E-FIT asks the user to select individual features in isolation and then synthesizes a face image using selected features. Considering that composition of facial parts is even more important than individual parts in face perception, some systems, such as EFIT-V [15] and EvoFIT [14], allow users to interact with the system by judging whether the whole face is similar instead of independently selecting the most similar individual component. Recently, significant developments in machine learning technologies, such as GAN technology, led to major advancements in the field, which has made face image generation a current research hotspot. In the remainder of this chapter, the existing face image synthesis approaches are classified into four categories: Component-based, Sketch-based way, deep learning-based approach, and evolutionary algorithm, and roughly introduced.

2.1 Component-based method

Component-based method is also called feature-based or montage based method [3][5], which requires the user to look through a dataset of face components (eyebrows, eyes, noses, mouth, etc) in order to search for each part separately based on resemblance and composite a face image using the selected parts. Individual facial features (eyebrows, eyes, noses, mouth, etc) are selected one at a time from a large feature database and then overlaid to make the composite image. A very typical application requiring user intervention in face image synthesis is assisting the police in investigations. Electronic facial identification (E-FIT) [3] is a face image synthesis system that can produce the facial composites of wanted criminals based on eyewitness descriptions. The core concept in E-FIT is the technique of Montage synthesis. Both FACES [1] and PRO-fit [4] are also component-based methods, which contain a larger number of face components for user to select. Selected components are arranged together to produce face image. PRO-fit contains face components of different races in individual databases while FACES combines them in one single database. FACES 4.0 is the latest and most advanced version. It features with an expanded database of 4,400 facial features, hair flip, and facial details

such as moles, scars, and tattoos [6]. Identikit [7] was introduced in U.S. in 1959 for synthesizing face image by superimposing facial features drawn on transparent acetate paper. It was later optimized into multiple versions, but all follow component-based principle, selecting individual components and then fusing them into a whole face image [8][9]. Identikit 7 [2] is the newest release which can quickly produce high-resolution face image. Facial components are more detailed and produce clear face image. Identikit 7 expanded the editing tools making it possible to easily perfect the generated face image. Component-based method essentially relies on the selection of individual features in isolation. Finding the ideal parts is time consuming. Furthermore, making sure the whole synthesized face image is or is not a desired one is usually difficult even if one can ensure that each picked part is satisfactory. In my research, the similarity of synthesized face image was judged based on the entire face instead of individual components, thus it can make up for the disadvantages of the component-based method.

2.2 Sketch-based method

Unlike Component-based method, to consider the entire face for face synthesis, some researches devote to explore facial image synthesis from a sketch. Wu and Dai [10] proposed a method to synthesize face images by querying a face image database using different parts of a face sketch. The corresponding face parts with the highest degrees of resemblance are patched together to form the final image. Users can adjust the size, shape, and colour of face parts to make the resulting face more resemble the desired face. Xiao *et al.* [11] developed a method enabling bidirectional photo-sketch mapping, which can synthesize a face sketch from a photo and, conversely, a photo from a face sketch. However, all sketch-based methods require the user to draw a sketch, a talent that not everyone has. The average people can’t draw sketches well.

2.3 Deep learning-based method

With the rapid development of deep learning technology, various research areas and applications, such as computer vision, robotics, big data analysis, and pilotless automobiles, have achieved major advancements. Generating face image using deep learning has gained wide attention in recent years. Compared with the traditional methods, deep learning-based state-of-the-art technologies are superior in terms of high-quality results. Especially the emergence of Generative Adversarial Networks (GANs) [17] has caused a sensation in image generation. A large variety of GAN network models for various image synthesis applications have been proposed.

2.3.1 Face image generation using GAN

Inspired by game theory, the original GAN [17] aims to build generative models via an adversarial process. As shown in Fig. 1, it consists of a generator network G that reconstructs the data distribution and a discriminator network D that estimates the probability of the generated sample coming from real data rather than G, and in which G and D are trained simultaneously. The generator receives a random noise $\mathcal{Z}$ as input and generates a distribution. The discriminator receives the real data distribution or the generated distribution from the generator. G is trained in a way that maximizes the probability for D to make a mistake.

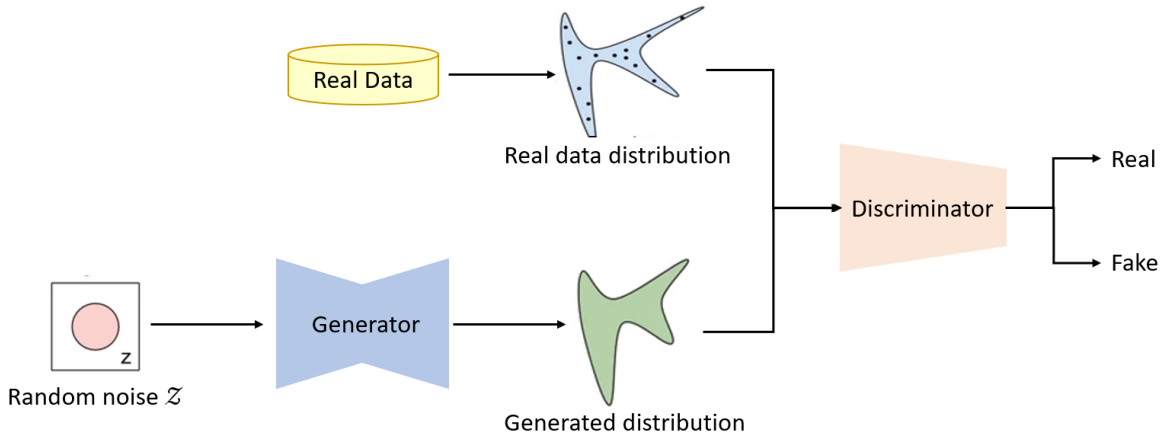


Figure 1. Architecture of generative adversarial network

Essentially, the goal of GAN is to generate a distribution that is as close to a real data distribution as possible. Therefore, minimizing the distance between two distributions is critical. How to measure the difference between fake data distribution and real data distribution? The objective functions, also called loss functions, are used to measure. A GAN is trained in a way to find a set of model parameters that can produce a distribution closely matching with the real data distribution.

The adversarial loss, denoted as $\mathcal{L}_{adv}$ in this thesis, is common to all GANs. By minimizing the adversarial loss, generator is trained to generate images that are realistic. In practice, discriminator tries to maximize adversarial loss while the Generator tries to minimize it. It is basically of the form defined in equation (1):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{\mathcal{Z}\sim p_{\mathcal{Z}}(\mathcal{Z})}[\log(1 - D(G(\mathcal{Z})))]. \quad (1)$$

The equation (1) is called minimax loss. $G(\mathcal{Z})$ is generated instance from the generator given noise $\mathcal{Z}$. $D(x)$ is the probability estimated by the discriminator that a real data instance x is real and $D(G(\mathcal{Z}))$ is the probability estimated by the discriminator that a fake instance is real. $\mathbb{E}_{x\sim p_{data}(x)}$ is the expected value over all real data instances and $\mathbb{E}_{\mathcal{Z}\sim p_{\mathcal{Z}}(\mathcal{Z})}$ is the expected value over all generated fake instances $G(\mathcal{Z})$. The formula (1) derivers from the cross-entropy between the real and generated distributions. The adversarial loss is also

called Jensen Shannon Divergence (JSD) divergence that is a method of measuring the similarity between two probability distributions in probability theory and statistics. WGAN [18] provides an alternative to traditional GAN training. It improved the quality of generated image by applying Wasserstein distance instead of the adversarial loss used in the original GAN. Radford et al. [19] proposed Deep Convolutional GAN (DCGAN) by adding a set of constraints on the network which makes the training of the network more stable. The largest advantage of DCGAN is that the trained discriminators can be used for unsupervised classification tasks.

Using face images as the real data, Goodfellow *et al.* demonstrated that the traditional GAN [17] can be trained to generate various face images from noises. However, the quality of generated face image is low and there is no control over the output.

2.3.2 Face attribution control with various conditional GAN models

To compensate for the disadvantages of lack of controllability, Conditional Generative Adversarial Nets (CGAN) [20] and its other variants [21-27] were proposed in succession. CGAN, an extension of GAN, as shown in Fig. 2, generates images of some particular attributes by feeding a conditional data to GAN model so as to gain some control over the results.

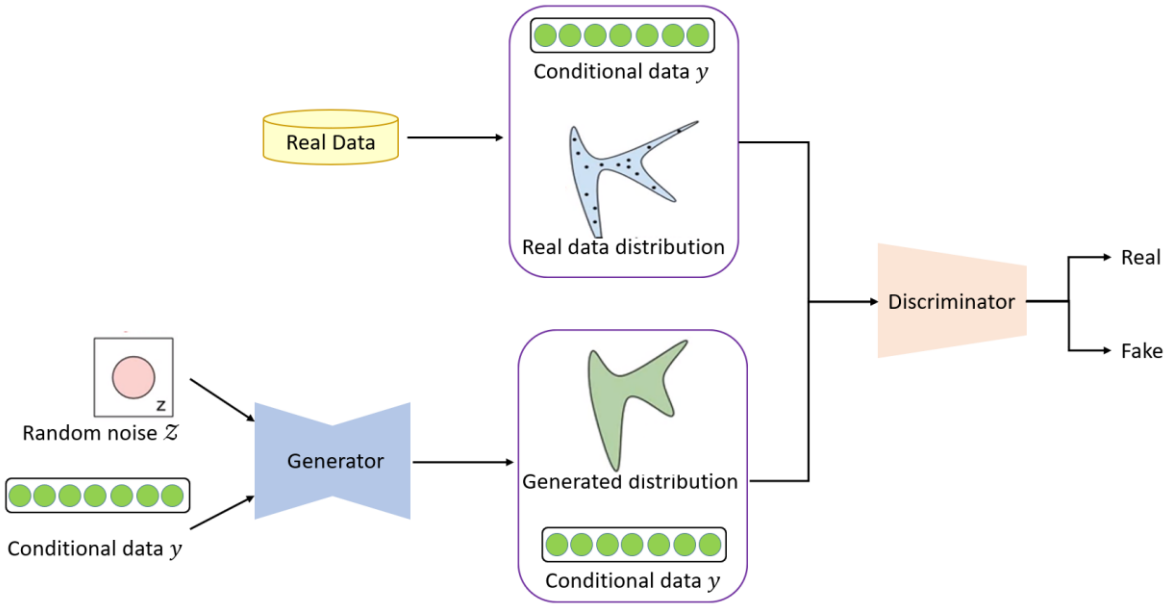


Figure 2. Architecture of conditional generative adversarial network

The conditions can be any attribute related to the target face image, such as sex, age, facial pose, with/without glasses and so on. The same year, Jon [21] applied CGAN to a conditional setting to face generation. Face image with specific attributes can be obtained by varying the conditional data for CGAN model.

On the CGAN basis, Age-GAN [22] is developed for generating face images of different ages, with a particular emphasis on preserving a person's identity in the aged version of his/her face.

Researchers have explored the potential of substituting the attribute input of CGAN with more specific control information, such as a target image, resulting in the so-called image-to-image translation technology. The most representative models of image-to-image translation are Pix2Pix-GAN [23], Cycle-GAN [24], and Star-GAN [25]. Pix2Pix-GAN [23] is trained using a training set of aligned image pairs, one is real image and another one work as conditional data. It not only learns the mapping from input to output image, but also a loss function to train this mapping. Pix2Pix-GAN [23] is effective for synthesizing photos from label maps, reconstructing objects from edge maps and colorizing images. However, for many tasks, paired training data is not available. Cycle-GAN [24], an approach to the translation of an image from a source domain to a target domain, can be trained in the absence of paired examples. However, it has limited scalability and robustness in handling more than two domains, since different models should be built independently for every pair of image domains. Choi *et al.* [25] proposed Star-GAN for image-to-image translations of multiple domains using only one model. It allows simultaneous training with multiple datasets of different domains within a single network. Given a face image and several target attributes such as blonde hair, dark skin, and age label, Star-GAN can generate the multi-attribute transfer results: a face image with blond hair, a face image with dark skin and a face image with desired age. Choi *et al.* [26] proposed Star-GAN v2 in 2020, which tackles two issues: (1) diversity of generated images and (2) scalability over multiple domains and improved results.

Xing *et al.* [27] exploited to use landmarks [29], which are the feature points characterizing the geometric features of faces, as the control information in CGAN to synthesize corresponding face images. Their model, called GP-GAN, can generate a face image that is similar to the image from which the landmark was extracted. In the second method proposed by this thesis, landmarks are used as the features for training the classifier based on the user's feedback, and GP-GAN is used for synthesizing realistic face images from the landmarks.

To enable the combination of different attributes, Style-GAN [28], featuring with a multi-resolution structure, is proposed recently. By modifying the input of each level separately, Style-GAN succeeds in controlling the output from coarse features (pose, face shape) to fine details (hair color). Unlike other GAN models that feeds noise directly, Style-GAN first map the input to an intermediate latent space, which then feed to adjust the “style” of the results. The noise is added at each convolution layer. In this way, it achieves unsupervised separation of high-level attributes and create stochastic variation in the generated images.

CGAN based methods can generate face images with specific attributes by varying the conditional information. However, these attributes are tagged to the data when preparing the training dataset, so there is no way to reflect the user's intention at the execution time. Although the recent Style-GAN achieves unsupervised

separation of high-level attributes, it is basically realized stochastically and provides no user control to the detailed features of the generated face images.

2.3.3 Age attribution control in GAN

On the CGAN basis, Perarnau *et al.* [30] proposed a method to reconstruct and modify real face image conditioned on arbitrary attributes. In this way, face aging is implemented by changing in a binary way, which is simply making face look older or younger without particular age categories. The generated face images thus fail to preserve the original person's identity. Antipov *et al.* [31] proposed Age-cGAN for automatic face aging. It can preserve the original person's identity in the aged version by introducing an approach for "Identity-Preserving" optimization of GAN's latent vectors. Identity-Preserving is used to optimize initial latent vectors that comes from the code of input face image. The Age-cGAN can generate high-quality synthetic images within required age categories. However, it is difficult to prepare labelled faces of the same person across a long age range since different persons have their own aging speed. Wang *et al.* [22] proposed the so called IPCGANs framework, which synthesizes a face lying in a given age group instead of a specific age. Meanwhile, the synthesized faces have the same identity with the input face image. But these models mainly rely on age labels or age group labels. In many cases, the age labels do not correspond to the real age, it is vague and inaccurate, especially in the era of advanced beauty industry. It is also computational expensive to use the pixelwise and Identity-Preserving optimization objectives. Therefore, as the third method, a High-Frequency Generative Adversarial Network (HF-GAN) for synthesizing faces with the user's desired texture features is proposed in this thesis. The method is based on the assumption that high-frequency textures, such as wrinkles, are the very important features related to age. Besides, to preserve semantic and perceptual characteristics of a face, a perceptual loss is adopted to enforce the preserving of high-level features in addition to the local texture details.

Although GAN-based methods have been successful for face generation, however, some experimental results show that the generated images also consist unnecessary background information. One possible explanation for this problem is that the most GAN-based models rely heavily on convolution. Convolution only focus on dependencies in a local neighbourhood and has particularly close relationship with the size of convolution kernel. It fails to capture long-range dependencies of entire image. From that point, previous GAN-based models fail to capture global dependencies in images. Of course, the capacity of the network can be improving by increasing the size of the kernel. But this way will increase the computational complexity. Cheng *et al.* [36] proposed self-attention to balance the ability to preserve long-range dependencies and the computational complexity. Zhang *et al.* [37] developed Self-Attention GAN (SAGAN) by introduce self-attention mechanism into convolutional GAN, which allows attention-driven, long-range dependency modelling for image generation tasks. Unlike traditional convolutional GAN which focuses on processing the

local neighborhood in convolutional layers, the SAGAN can generate details using cues from all feature locations. In the third method, self-attention mechanism is also added to HF-GAN model to generate high-resolution details from all feature location in image rather than only from local points in image. For the generator and discriminator of HF-GAN model, the self-attention is inserted between the residual blocks of network. In this way, the generator can generate face images with fine details at every location. The discriminator can also more accurately distinguish the input images. Most importantly, face part in the image is focused on in the whole process.

2.3.4 Summary on GAN models for face image generation

In this section, GAN models for face generation were summarized from the perspective of input/output, learning methods, training objectives, and backbone. The details are shown in Table1.

From the perspective of input, GAN models can be roughly divided into two categories: 1) Generate face image from noise $\mathcal{Z}$ sampled from a normal or uniform distribution. 2) Translate original face image into target one with desired attributes. The former means that, conceptually, noise $\mathcal{Z}$ represents the latent features of the generated image. GAN perform multiple transposed convolutions to upsample $\mathcal{Z}$ to generate face image. It is unknow that what is the semantic meaning of each bit in $\mathcal{Z}$. Training GAN learn the semantic meaning of noise. The goal of latter is to learn the mapping between an input image and an output images such as the cases of style transfer and season transfer.

The learning method here focuses on whether the paired data is required or not when training a GAN model. It is referred as supervised learning and unsupervised learning here. Supervised learning, such as the case of CGAN [20] and Pix2Pix GAN [23], needs paired data for training the model. When training CGAN, it is required that conditional data, label of real data, should be paired with the corresponding real data. For training Pix2Pix GAN, the conditional data is image, which is also required to be paired with the real data. Unsupervised learning can train a model in the absence of paired examples such as original GAN [17] and Cycle-GAN [24].

Training objectives play a key role in training the model. In addition to adversarial loss, the more commonly used loss functions are classification loss, reconstruction loss, and perceptual loss. Classification loss represents the cost paid for inaccuracy of predictions in classification (predicts which class an identified image belongs to). It is associated with classifying and generating images with a specific target label. Classification loss for the discriminator is denoted as $\mathcal{L}_{cls}^{D}$ and Classification loss for the generator is denoted as $\mathcal{L}_{cls}^{G}$. By minimizing the classification loss, generator is trained to generate images that are classified to its correct target class. However, both adversarial loss and classification loss fail to guarantee that generated images preserve the content of its input images. To resolve this problem, reconstruction loss is proposed and denoted as $\mathcal{L}_{rec}$. Although the loss functions mentioned above optimize the results of the model to a certain extent, however,

these loss functions cannot capture high-level perceptual and semantic differences between real data and generated image. Perceptual loss, denoted as $\mathcal{L}_p$ is used to capture perceptual information. The value of perceptual loss is the error between high-level image feature representations extracted from a pre-trained convolutional neural network. Usually, the pre-trained convolutional neural network is VGG network [38] pretrained on the ImageNet dataset [39]. The high-level features of generated image and input image are extracted using pretrained VGG network and then L1 distance between their features is computed that is used to optimize the generator.

For the design of network architecture, Convolution neural network (CNN) [40] is undoubtedly the most popular deep learning architecture and is the basis of various deep learning models. It employs the convolution operation which first performs a linear operation by applicating filters to input, and then activates neural unit and performs the pooling operation. A typical CNN model consists of an input, an output, and multiple hidden layers including a series of convolutional layers and fully connected layers. It can do good job in terms of accuracy and automatic detection of e important features. But CNN fails to learn the position and orientation information and needs big dataset for training. And it cannot keep spatial relationship between features. Later, U-Net was proposed by Olaf *et al.* [41] for biomedical image segmentation. It is symmetric and consists of a contracting path and a symmetric expending path. The former is same as a typical convolutional network for encoding through repeated convolutions and downsamplings and the latter is for decoding through upsampling operators. In addition, U-Net uses the skip connections between the contracting path and the expending path to preserve low-level features. As a result, there are a large number of feature channels in expending path that allow the network to propagate context to higher resolution layers. It is generally known that the results will be better and accuracy will be enhanced as the depth of the network increases. Increasing depth of network usually causes vanishing gradients and network crash. To solve degradation problem, residual neural network (ResNet) is proposed in [42]. The ResNet is easier to optimize and can gain accuracy from considerably increased depth by adding residual block to normal CNN. In a network with residual blocks, it adopts shortcut connections to feed residual mapping to layers that jump over some layers. the Dense Convolutional Network (Dense-Net) [43] is another network of particular interest in addition to U-Net [41]. Dense-Net is residual neural network with several parallel skips. The Dense-Net has several significant advantages. It overcomes the gradient disappearance problem to a certain extent, enhancing feature forward propagation, reusing features, and can provide a good representation of the image and reduces the number of parameters.

Table1. Summary of deep learning-based methods for synthesizing face images

Model	**Input**	**Output**	**Training method**	**Quality**	**Interaction**	**Training objectives**		**Backbone**	
						Generator	**Discriminator**	**Generator**	**Discriminator**
GAN [17]	Noise $\mathcal{Z}$	Image	Unsupervised	Low	N	$\mathcal{L}_{adv}$	$\mathcal{L}_{adv}$	Conv	Conv
CGAN [20]	Noise $\mathcal{Z}$ + label	Image	Supervised	Low	N	$\mathcal{L}_{adv} + \mathcal{L}_{cls}^{G}$	$\mathcal{L}_{adv} + \mathcal{L}_{cls}^{G}$	Conv	Conv
DCGAN [19]	Noise $\mathcal{Z}$	Image	Unsupervised	Low	N	$\mathcal{L}_{adv}$	$\mathcal{L}_{adv}$	Conv	Conv
Age-GAN [22]	Image + label	Image	Supervised	Low	N	$\mathcal{L}_{adv} + \mathcal{L}_{cls}^{G}$	$\mathcal{L}_{adv} + \mathcal{L}_{cls}^{G} + \mathcal{L}_{p}$	ResNet	Patch GAN
Pix2Pix-GAN [23]	Image	Image	Supervised	High	N	$\mathcal{L}_{adv} + \mathcal{L}_{rec}$	$\mathcal{L}_{adv}$	UNet	Patch GAN
Cycle-GAN [24]	Image	Image	Unsupervised	High	N	$\mathcal{L}_{adv} + \mathcal{L}_{rec}$	$\mathcal{L}_{adv}$	ResNet	Patch GAN
Star-GAN [25]	Image+ label	Image	Supervised	High	N	$\mathcal{L}_{adv} + \mathcal{L}_{rec} + \mathcal{L}_{cls}^{G}$	$\mathcal{L}_{adv}$	ResNet	PatchGAN
Style-GAN [28]	Noise $\mathcal{Z}$ / Image + Style	Image	Unsupervised	High	N	$\mathcal{L}_{Wasserstein-1}$	$\mathcal{L}_{Wasserstein-1}$	DeNet+ResNet	PatchGAN
GP-GAN [27]	Image	Image	Supervised	High	N	$\mathcal{L}_{adv} + \mathcal{L}_{rec} + \mathcal{L}_{p} + \mathcal{L}_{cls}^{G}$	$\mathcal{L}_{adv} + \mathcal{L}_{cls}^{D}$	UDeNet	PatchGAN
IECGAN [32]	Noise $\mathcal{Z}$	Image	Unsupervised	Low	Y	$\mathcal{L}_{Wasserstein-1}$	$\mathcal{L}_{Wasserstein-1}$	Conv	Conv

2.4 Evolutionary method

The first evolutionary method for facial image composition is given by Davies and Valentine [12]. Then Stuart Gibson *et al.* [13] introduced a method for photographic quality facial composition using evolutionary algorithms. Unlike traditional component-based methods, it combines random samples from a facial appearance model with an evolutionary algorithm to drive the search procedure to convergence.

Charlie, Peter and Derek [4] developed EvoFIT system that is an interactive computerized facial image composition system. EvoFIT presents a number of possible faces and asks the witness or user to select those that look most like the target face image. In this method, for each iteration, the presented possible faces are used to generate a new set of faces using genetic algorithm. After many iterations, the system gradually synthesizes face image which closer to the target face. There are also tools available in EvoFIT system to improve the likeness on demand, such as to change the perceived age, weight and also to change individual components of the face—eyes, nose, mouth, etc [14]. The evolutionary algorithm was also explored by Stuart *et al.* [13] who uses both local and global models, allowing a witness to evolve plausible, photo-realistic face images in an intuitive way. EFIT-V system [15] synthesizes face image based on whole face principles. The witness is shown a number of randomly generated face image and is asked to select the one that he/she thinks is most similar to the target one. The system employs a genetic algorithm to breed new face images based on the selected images.

There are some limitations to the traditional evolutionary methods such as genetic algorithm for solving the problem of face composition. It is sensitive to the initial population. It has stochastic property being a non-deterministic class of algorithms. Another bigger limitation is that the optimal resolution obtained is rather a sub-optimal one but globally one although genetic algorithm can indeed provide an optimal solution. And, usually the convergence speed is slow, which needs dozens, even more than a hundred times iterations.

Gibson *et.al*. [13] report an evaluation of genetic algorithm system by conducting trials based on simulated witness behaviour. Unfortunately, the result shown Select Multiple Mutate algorithm (SMM) [13] required 150 iterations, and the Follow-The-Leader algorithm (FTL) [13] required 350 iterations to produce a satisfactory face image. For example, a subject obtained a result after viewing 162 faces, over 27 iterations and took approximately 20 minutes using SMM algorithm.

Philip et al. proposed an approach [32] based on Wasserstein GAN [18] and the genetic algorithm [16] to produce user-desired images. The user is asked to evaluate a set of images resulting from GAN, and a genetic algorithm is applied to modify the noise input based on the user's evaluation. Their paper showed some examples of using the method for generating face images resembling target faces. However, the average score of the generated images is only 2.2 out of 5, which is the result of the user evaluation experiment.

In this thesis, a relevance feedback technology and OPF classifier are employed to quickly learn user's intention through an iterative approach. During iteration, OPF classifier is trained based on relevance feedback. Trained OPF classifier is used to select candidate face images for the synthesizing final results. The experiment shows that on average, it took 6.5 iterations for the subjects to arrive at the final results.

Chapter 3 Based on Principal Component Analysis (PCA)

3.1 Proposed method

As depicted in Fig. 3, the first method includes three major components: extracting primary features, training an Optimum-Path Forest (OPF) classifier based on relevance feedback, and synthesising face images that do not already exist in the database.

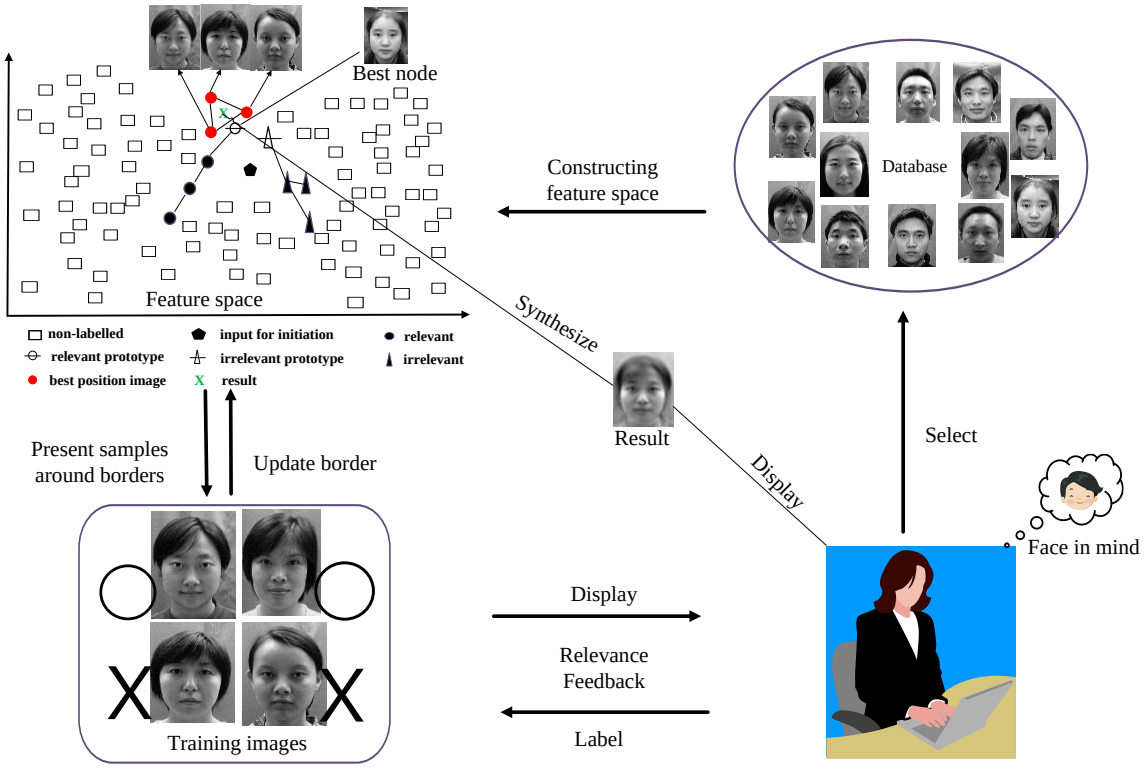


Figure 3. Overview of the proposed system

1,000 sample images were used and these images were converted to a feature space for training an OPF algorithm to classify whether a face image resembles the face in the users' minds based on their relevance feedback. The ultimate purpose of the proposed method is not to classify those sample face images or to retrieve a particular face from these sample face images but to synthesise a new image resembling the face in the user's mind. The trained OPF classifier defines the positions in the feature space that correspond to the desired face images.

To train the OPF classifier, the system defines an initial classification boundary by letting the users evaluate an initial dataset consisting of face images of different sexes and ages. Then, the system shows the user multiple unevaluated images (i.e. cases that have not been judged by the user to resemble or not resemble the picture in his or her mind) that lie near the classification boundary and has the user label them according to whether they resemble or do not resemble the face in his or her mind. Based on these labels, the system updates the classification boundary.

Then, the system interpolates K cases in the positions farthest from the classification boundary on the relevant side and produces the final synthesis. If the results satisfy the user, the search process is complete; otherwise, the user repeats the labelling process on unlabelled cases near the classification boundary.

3.2 Constructing the feature space

3.2.1 Feature representation

Various feature representations have been studied in the context of face recognition in the past few decades. Recent research results have demonstrated that deep learning can be used to learn the face representation, which is effective for both face identification and verification [44, 45]

However, since the purpose of the proposed method was to synthesise the target face image, a feature representation that could not only discriminate faces but could also be used to generate a face image is needed. The feature vector space needed to be compact enough to allow for the interactive relevance feedback process. For this purpose, the pixel-level image feature is used similar to face hallucination method [46].

The basic idea is to separate a face image I into a global image I_g, which expresses the overall features of the image, and a local image I_l, which expresses the detailed face features.

$$I = I_g + I_l, \qquad (2)$$

While the local image adds the details of the face, global images comprise information required for distinguishing between individuals. A feature vector space of global images can be constructed by applying PCA to the face images in the database and finding the principal components with large eigenvalues. Formula (3) ex-presses a global image I in terms of the basis B of a global feature space, a coordinate value X and an average face image μ :

$$I = BX + \mu, \qquad (3)$$

3.2.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique for reducing dimensionality. Using an orthogonal transformation, it transforms a set of possibly correlated variables into a set of linearly uncorrelated variables that are called principal components.

Assuming there is an initial dataset X which contains n variables $x_i \in X$ $(i = 1, 2, 3 \dots n)$ with p dimensions,

Step1: Standardization

If there are large differences between the ranges of initial variables, variables with larger ranges will dominate over variables with small ranges. That will give rise to biased results. To avoid this problem, it is necessary to standardize the range of the initial variables. by subtracting the mean and dividing by the standard deviation for each variable, all the variables will be converted into the same scale.

$$x_i^{new} = \frac{x_i - \mu}{\sigma}, \qquad (4)$$

where μ is the mean and σ is the standard deviation.

Step2: Covariance matrix computation

To identify relationships between variables, covariances matrix is computed since sometimes variables are highly correlated in a redundant way. The covariance matrix is a $p \times p$ symmetrix matrix (where p is the number of dimensions) which summaries the relationships between all the possible pairs of variables. It is denoted as A_{cov} .

$$\begin{bmatrix} cov(x_0, x_0), cov(x_0, x_1), cov(x_0, x_2) \dots \dots, cov(x_0, x_{n-1}) \\ cov(x_1, x_0), cov(x_1, x_1), cov(x_1, x_2) \dots \dots, cov(x_1, x_{n-1}) \\ cov(x_2, x_0), cov(x_2, x_1), cov(x_2, x_2) \dots \dots, cov(x_2, x_{n-1}) \\ \dots \dots \\ cov(x_{n-1}, x_0), cov(x_{n-1}, x_1), cov(x_{n-1}, x_2) \dots \dots, cov(x_{n-1}, x_{n-1}) \end{bmatrix}, \qquad (5)$$

In the main diagonal, the values are the variances of each variable since the covariance of variable with itself is its variance $cov(a, a) = var(a)$. And the covariance matrix is symmetric with respect to the main diagonal because the variance is interchangeable ($cov(a, b) = cov(b, a)$).

Step3: Calculate the eigenvalues and eigenvectors for the covariance matrix

Principal components are new uncorrelated variables, which can be thought as new axes. The eigenvector is a nonzero vector that will be changed when the linear transformation is performed. The corresponding eigenvalue is the scaling factor of the eigenvector.

$$A_{cov}v = \lambda v, \qquad (6)$$

where v is a vector and λ, called eigenvalue associated with eigenvector v, is a scalar. Eigenvectors and eigenvalues are calculated from the covariance matrix based on the equation (6).

Step 4: Obtain k eigenvalues and a matrix of eigenvectors

After computing the eigenvectors and sorting eigenvalues by their eigenvalues, top k eigenvalues and corresponding matrix of top k eigenvectors are obtained. They are selected as the principal components and formed the feature vector. In this work, top 80 eigenvalues are picked. A matrix of 80-dimensional eigenvectors is obtained, which is the B in equation (3).

3.3 Training the optimum-path forest classifier based on relevance feedback

Relevance feedback, a process that shows synthesis results to the users and updates classifiers based on user feedback, is often used in image retrieval with specific themes, such as oceans, cats or sunsets. Several researchers have proposed methods that employ various classifier types and reuse past classification results to obtain good results based on relatively minimal amounts of feedback [44, 47, 48].

As depicted in Fig. 3, in the proposed method, the OPF is trained based on the users' relevance feedbacks in the following four steps:

Step1: The system presents the user with five male face images and five female face images of different ages and waits for the user to select one he or she thinks to be closest to the face in his or her mind. Since none of those 10 images is likely to resemble the target face, the user will select the image that is the most similar to what they are imagining according to sex and age, which acts as the initial classification boundary.

Step2: The four images closest to the user's selected face image in the feature spaces are returned to the user. The user evaluates and labels the images as relevant (○) or irrelevant (×), which serve as the prototypes for the OPF classifier. This evaluation phase ends if the users are satisfied with at least one of the four face images.

Step3: An OPF classifier is built based on the prototypes as illustrated in Fig. 4. Then, the sample images are divided in to two classes: relevant and irrelevant.

Step4: Four border nodes are selected, and the corresponding images are presented to the users. The user evaluates and labels the images as relevant (○) or irrelevant (×), and the new marked training images constitute and replace the former prototypes to build a new OPF classifier.

At every iteration before step 4, the best relevant nodes, which are nodes located the farthest to irrelevant prototype and the closest to relevant prototype are selected and interpolated to create the resulting face image being presented to the user. If the user is satisfied, the whole relevance feedback procedure ends.

OPF [33, 44, 45] is originally for classification. It represents each class of images by optimum-path trees rooted at the given representative samples, called prototypes [46], [47]. The OPF works by modelling the classification as a graph partition in a given feature space. It starts as a complete graph whose nodes represent the feature vectors of all training samples in the dataset (Fig. 4c). All pairs of nodes are linked by arcs that are weighted by the distances (referred to as cost hereafter) between the feature vectors of the corresponding nodes. At each iteration of the relevance feedback, a set of training nodes are obtained by the user labelling the samples as relevant or irrelevant (step 2 of the aforementioned framework), a minimum spanning tree is constructed for the labelled samples (Fig. 4d), and the adjacent pair of the relevant and irrelevant samples are chosen as the relevant and irrelevant prototypes, respectively (Fig. 4e). Then, the graph is repartitioned by the competition process among prototypes, which offer optimum paths (the path with the lowest cost) to the remaining nodes of the graph and classify all nodes into relevant or irrelevant depending on whether they are connected to a relevant or irrelevant prototype (Fig. 4f). The optimum paths from the prototypes to the other samples are computed by the image foresting transform algorithm, which is essentially Dijkstra's algorithm modified for multiple sources and with more general path-value functions. Finally, all of the non-prototypes are directly or indirectly connected with the prototype that has the minimum cost. With the prototypes as the roots and the non-prototypes as the intermediate and terminal nodes, the optimum trees are built, which constitutes the OPF. It is known to that OPF have the ability to handle a large dataset effectively and efficiently compared with other representative classification algorithms, such as the support vector machine and the k-nearest neighbors algorithm [49]. Because of this, OPF is very important in systems that are based on the relevance feedback approach and generate results in a dialogic fashion.

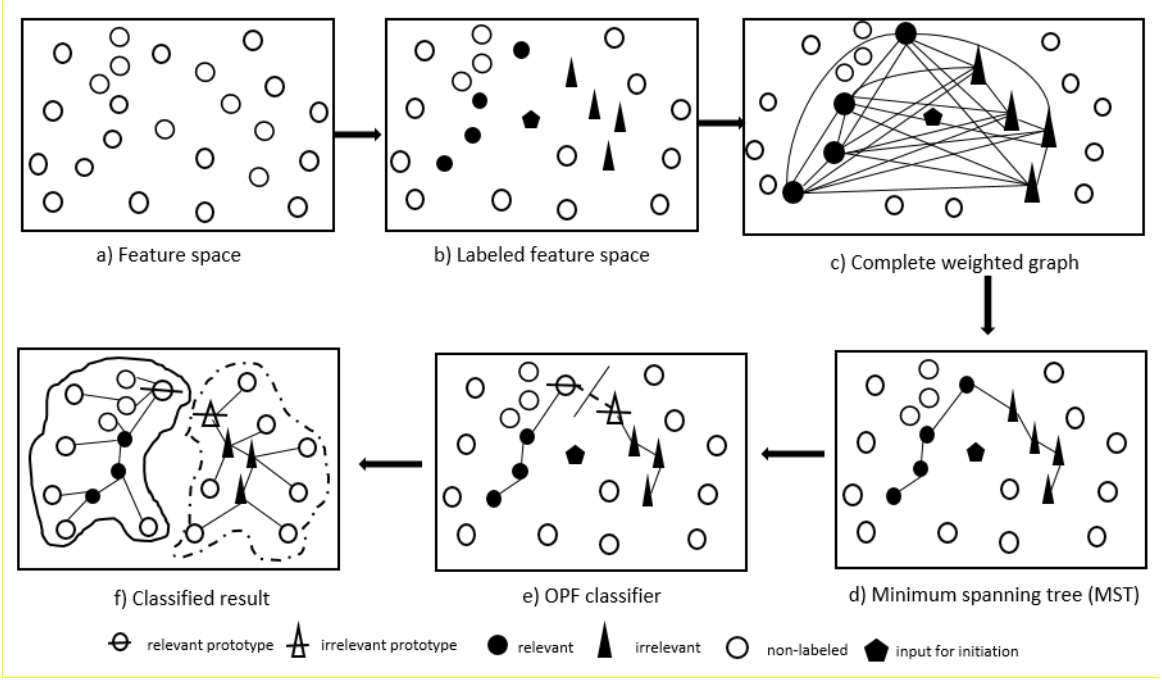


Figure 4. Training of OPF.

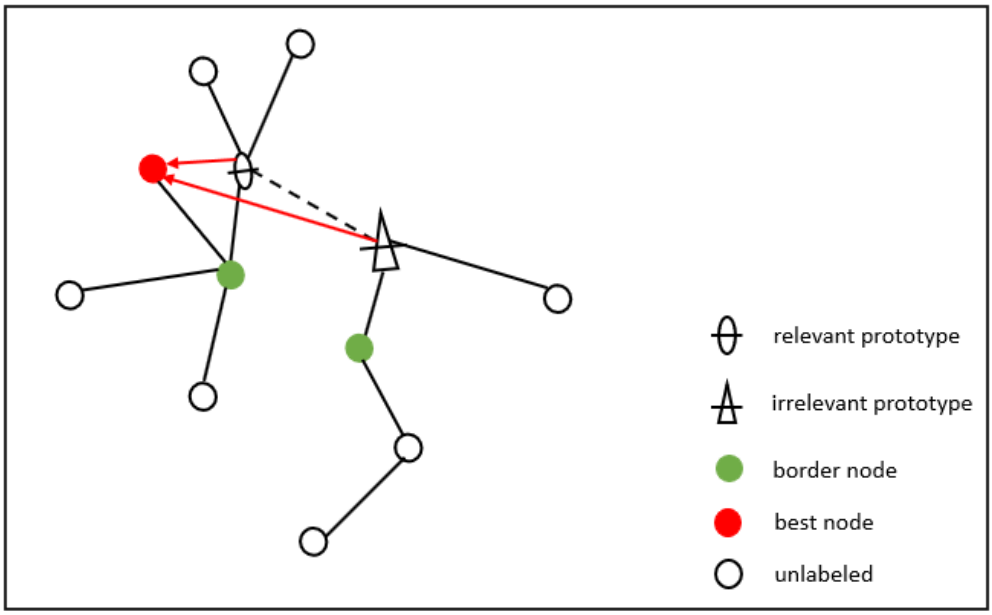


Figure 5. An example of classification with OPF

Fig. 5 shows an example of a classification with OPF. A minimum spanning tree is first constructed from all the samples. Then, the user labels some selected samples as relevant (○) or irrelevant (×). The paths that bridge relevant and irrelevant samples were thus focused on. The nodes bridged by the paths are prototypes,

which are represented as ⊖ and ⍋. All other unlabelled nodes whose parent is a relevant prototype are labelled as relevant, and the ones whose parent is an irrelevant prototype are labelled as irrelevant. The nodes next to the prototypes are called border nodes as indicated by the green dots. The node located the farthest from the irrelevant prototype and closest to the relevant prototype (depicted by the red node in the figure) is selected as the best relevant sample.

When selecting the border nodes that are returned to the user for labelling at each iteration and the best relevant nodes that are used to explore the candidate for synthesizing the final result, the costs of paths from all non-training nodes to all relevant and irrelevant prototypes were compared. To effectively update the classifier based on the user's feedback, the subsequent 4 samples to be labelled by the users are chosen from the nodes near the border of the classifier. The border nodes are defined as non-prototype nodes that belong to the relevant class and with the smallest ratio of the cost from these nodes to the relevant prototypes over the cost from these nodes to the irrelevant prototypes. The best relevant nodes should closer to the relevant prototypes and farther from the irrelevant. Therefore, best nodes are those that belong to the relevant class and with the largest ratio between the cost to the relevant prototypes and the costs to the irrelevant prototypes.

In the traditional relevance feedback-based image retrieval, the final result is the relevant case in the position farthest from the classification boundary. To establish the classification boundary correctly, the image shown to the user for feedback must lie near the classification boundary. OPF based retrieval thus requires an initial classification boundary that sits relatively close to the relevant case. This study satisfied this requirement by gathering age and sex input information at the beginning of the process.

In my implementation, the cost of any two adjacent nodes is assigned using L2 norm distance. It is assumed that there are s number of non-prototype samples $U_k(k = 1, 2, 3, \dots \dots, s)$ that belong to the relevant class, n number of relevant prototypes, and m number of irrelevant ones denoted as $p_i(i = 1, 2, 3, \dots \dots, n)$ and $q_j(j = 1, 2, 3, \dots \dots, m)$. The cost of the path from a non-prototype sample U_k to the relevant prototype p_i as $C_{U_K \to p_i}$ and the cost of the path from U_k to the irrelevant prototype q_j is denoted as $C_{U_K \to q_j}$. The ratio of $C_{U_K \to p_i}$ to $C_{U_K \to q_j}$, denoted as Relevance $U_k \to (p_i, q_j)$, can be computed as follows:

$$RelevanceU_k \to (p_i, q_j) = \frac{Avg(C_{U_K \to p_i})}{Avg(C_{U_K \to q_j})}. \qquad (7)$$

The 10 border nodes with the largest value of $RelevanceU_k \to (p_i, q_j)$ are chosen for the user to label. The best node chosen is the one with the smallest value of $RelevanceU_k \to (p_i, q_j)$.

3.4 Synthesising virtual face images using interpolation

The traditional relevance feedback approach is designed for searching actual images in a given database, making it impossible to synthesise non-existent face images. By synthesising images, however, it is possible to obtain the desired outcomes with a limited number of samples. This study thus proposes a process of synthesizing face images that do not exist in the database by interpolating multiple relevant images in positions far away from the classification boundary. In principle, any point near the best relevant node (i.e. the node that belongs to the relevant class and with the largest ratio between the cost to the relevant prototypes and the costs to the irrelevant prototypes) should be a desired face image.

As a practical solution, the top k ($k = 3$ in my implementation) best relevant nodes were selected, as shown in Figure 1, and calculate the result according to the following Formula (8):

$$x = \sum_i^k w(x_i) w_i / \sum_i^k w(x_i), \qquad (8)$$

Here, x and $x_i (i = 0, 1, 2)$ are the feature vectors of the resulting face images and the 3 best relevant images, respectively. The weight assigned to x_i is $w(x_i)$, which is based on the distance given by the classifier. In my implementation, $w(x_i)$ is assigned the average weight, which means all 3 images have equal weight.

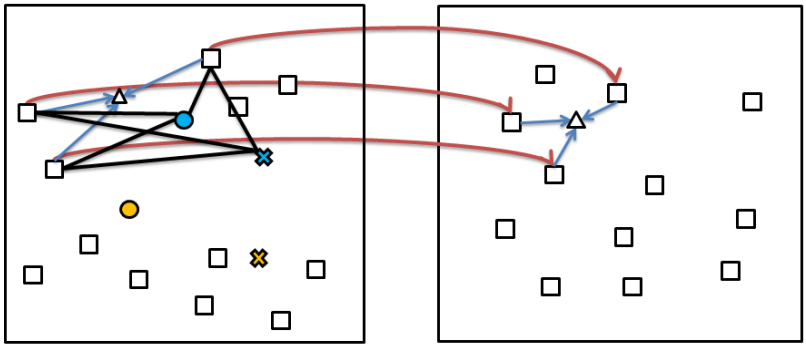
3.5 Registration by eyes and mouth

The sample images in the training database need to be aligned in order to create face images without blurring. In cases where the same images aligned only by one single registration point when synthesising new face images by interpolating several face images, the system was prone to blurring portions of the face away from the registration point due to the inherent individual variations among these different faces. Fig. 7 (a) and (b) show the results generated with the images were aligned by eye position and mouth position only, respectively. It can be observed the areas far from the registration positions are severely blurred.

To solve this problem, two image databases were built from the same source database: one composed of face images aligned by the eyes and the other composed of face images aligned by the mouth. In order to synthesise a clear face image, a group of images from the eye-aligned database and the corresponding images from the mouth-aligned data-base are used. More specifically, three procedures are carried out: first, a candidate image in the eye-aligned face feature space is synthesised; then, another face image in the mouth-aligned space is synthesised; by blending the two images, a clear composited face image is produced.

As the system makes it possible to obtain the same face from both databases, the relevance feedback process is needed to perform with one of the two databases to build the OPF for both databases. Fig. 6 illustrates the integration between the feature spaces of the two databases. When selecting the three highest-ranking

coordinates in the feature space defined by the images aligned by eyes, for example, the one-to-one correspondence between the two spaces means that the corresponding three highest-ranking coordinates can be obtained in the other feature space built from the examples aligned by the mouth. As the arrows in Fig. 6 show, the system thus enables coordinate matching across the two spaces. Thus, two face images can be synthesized by interpolating the three highest-ranking coordinates in the two spaces, respectively.



(a) (b)

Figure 6. Image correspondence between eye-aligned space and mouth-aligned space. In the eye-aligned feature space shown in (a), the yellow ○ and × represent user-labelled images, and the blue ○ and × represent the relevant and irrelevant prototypes. The □'s connected to the prototypes by black lines are the three best relevant images. The △ represents the generated virtual image interpolated using the three best relevant images. The correspondence between the best relevant images in the eye-aligned and mouth-aligned spaces are illustrated with red lines. The interpolated virtual image in the mouth-aligned space is shown with a △.

To fuse the two face images computed from the separately aligned spaces (i.e. the images represented by △ in Fig. 6[a] and [b]) and form a new image with clear face components, α blending is used, as given by Formula (9):

$$I = \alpha I_e + (1 - \alpha) I_m, \qquad (9)$$

I_e and I_m are the images from the eye-aligned and mouth-aligned spaces, respectively, and α is the blending weight. The value of α is set to 1 in the area above the eyes, value of α is set to 0 in the area below the mouth, and the α value is changed in the area between the eyes and the mouth for linear interpolation. The blended image is further filtered with a bilateral filter to decrease the degree of edge blur.

Fig. 7 (a), (b) and (c) show the resulting images synthesised in eye-aligned space, mouth-aligned space and by α blending the former two images, respectively. Fig. 7 (a) and (b) show that areas far from the registration areas are severely blurred, while in Fig. 7 (c), such flaws are alleviated.

(a)Image generated in eye-aligned space

(b) Image generated in mouth-aligned space

(c) Image generated via α blending

Figure 7. Comparison between single point registered faces and face obtained by α blending.

3.6 Experiment and discussion

3.6.1 Database

For sample image set, a total of 1,000 images of Asian faces from the CAS-PEAL database [50] and the Cartoon Face database [51] are used. All the images are made monochrome, and the resolution is set to 96×128. The database comprised only frontal face images, but the positions and sizes of the faces differed. The images are resized and cropped. Then two databases that were aligned by eye positions and mouth positions, respectively, were created. This study was concerned only with general face features, therefore, a low resolution of 96×128 for all the images are used. The images are set to monochrome to prevent colours not found in the original cases from appearing when the system interpolates multiple colour images.

Each dimension of the feature space corresponds to a pixel of the face images. There-fore, the feature vector has 12,288 (96×128) dimensions. Based on a primary component analysis, the 80 dimensions with the highest eigenvalues are employed as global face feature space of the proposed method, which provided a cumulative contribution ratio above 80%.

3.6.2 Experiments

To validate the effectiveness of the proposed method, a total of 12 subjects (all university students in their 20s, 11 of whom were male and 1 of whom was female) have attended experiments. During the experiments, the subjects were asked to ignore hairstyles when creating and evaluating the face images because the significant differences in hairstyles among the images in the database led to blurred hair in all the generated images.

The following three experiments were conducted to determine whether the subjects could create satisfactory face images using the system and how long (in terms of time and iteration count) this process would take.

Creating Imagined Face Images

In this experiment, each subject was asked to imagine a face and then asked to use the system to create a similar image. Fig. 8 shows the created images based on the subjects' imagined faces. In section 3.6.3, how similar these created face images are compared to the imagined faces will be evaluated.

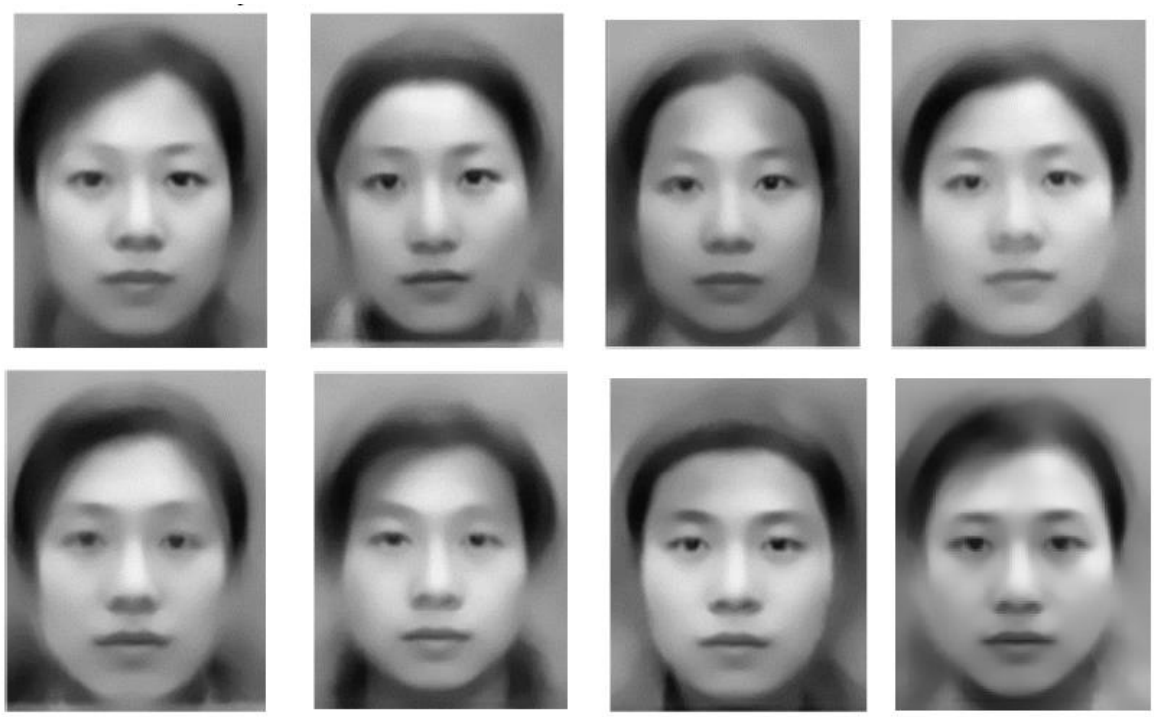
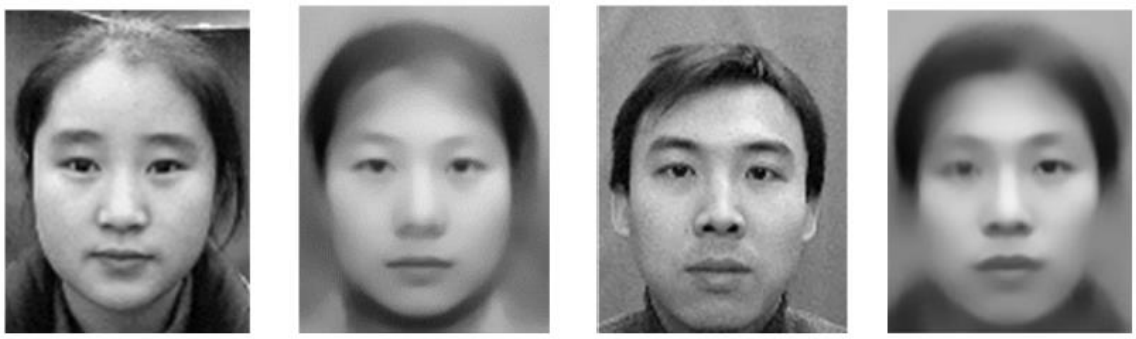


Figure 8. Images created based on the subjects' imagined faces

Creating Face Images Based on Briefly Presented Reference Images

In this experiment, a reference face image that did not exist in the database was presented to each subject for 3-4 seconds and the subject was asked to create a face image resembling the reference image to validate whether the system enabled the user to synthesize an image from his or her memory. Such a situation is similar to the case where an eyewitness has seen a criminal's face for a very short time and tries to reconstruct the face image based on his or her rough impression and memory. Fig. 9 shows the face images that the subjects saw for 3-4 seconds and the corresponding generated face images. As it can be seen from the figures, the resulting images capture some major features of the reference faces, such as the overall shape of faces and the relatively small sizes of the eyes.



Reference image A Created image of A Reference image B Created image of B

Figure 9. Reference images how n for 3-4 seconds and the corresponding created images

Creating Face Images Based on the Reference Images Presented During the Entire Process

For a more objective validation, a third experiment presented the subjects with a reference image for the entire duration of the process until they reached the result they found satisfactory. Fig. 10 shows two examples of the results. The resulting images maintain a basic consistency with their respective target images.

Reference image A Created image of A Reference image B Created image of B

Figure 10. Reference images shown during the entire experiment and the corresponding created images

Reference image A 1^{th} iteration 2^{nd} iteration 3^{rd} iteration 14^{rd} iteration

Figure 11. Changes of the synthesised result over the process

Fig. 11 illustrates how the resulting images actually changed over the process. The target image and the resulting image were noticeably different at first. As the subject went through iterations of the process, the face in the resulting image gradually came to more closely resemble the reference one.

3.6.3 Evaluation

Evaluation based on subjective scoring

In the three experiments mentioned above, the subjects were asked to score the results on a five-point scale (1: No resemblance; 2: Very weak resemblance; 3: Neither weak nor strong resemblance; 4: Somewhat strong resemblance; 5: Strong resemblance).

Fig .12 shows the average scores, in which the scores of three experiments are very similar. The average score for all three experiments came to 3.833. Many of the subjects declared that they were satisfied once the created images began to bear a somewhat re-semblance to the target faces. Fig. 13 shows the times (in seconds) that it took the subjects to arrive at satisfactory results.

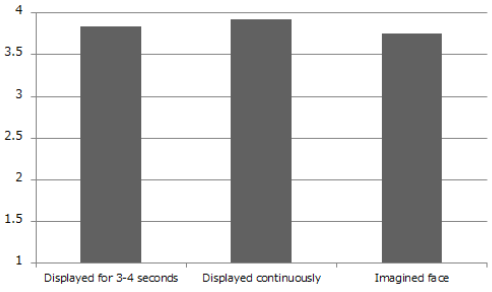


Figure 12. Average final scores

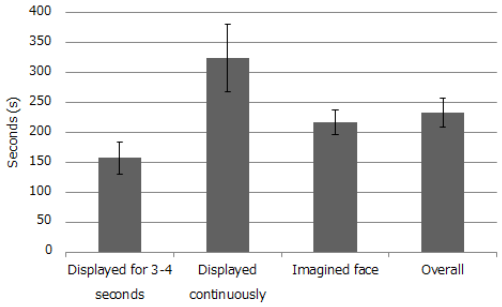


Figure 13. Average time to final results (in seconds)

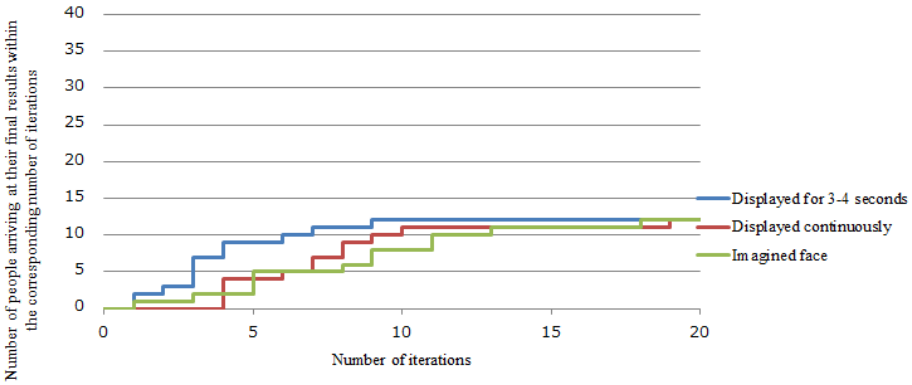
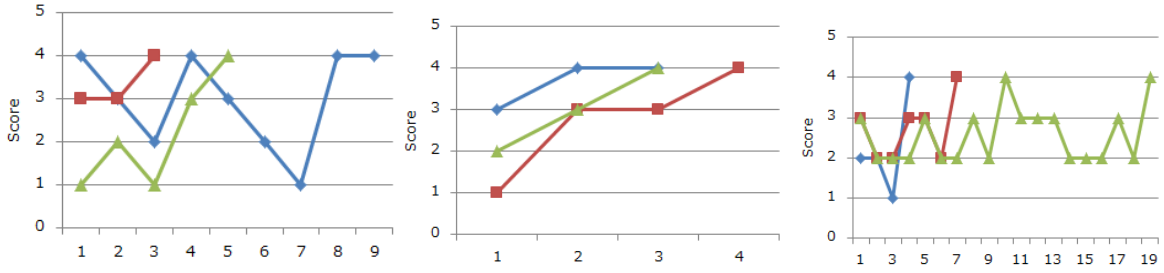


Figure 14. Iteration numbers of each experiment. The vertical axis of the graph represents the number of subjects who reached their final results with the number of iterations shown on the horizontal axis.

Fig. 14 shows the number of iterations that it took the subjects to arrive at satisfactory results. Fig. 15, meanwhile, illustrates the changes in scores for three subjects during the relevance feedback process. Each line represents a single iteration of the pro-cess by an individual subject. On average, it took 6.5 iterations for the subjects to arrive at the final results.



(a) (b) (c)

Figure 15. Changes in scores during the relevance feedback process: (a) Creating an imagined face image, (b) creating a face image based on briefly (3–4s) presented reference images, and (c) creating a face image based on the reference images presented during the entire process.

Evaluation based on matching test

To evaluate the effectiveness of proposed method, a matching test was conducted by letting a group of participants creating face images using the system from reference images, and then having another group of participants match the generated images with their reference images. In this test, 10 peoples of different age (5 in 20s, 1 in 30s, 3 in 40s and 1 in 50s) and gender (6 male and 4 female) were asked to synthesize face image, while another 13 peoples of different age (9 in 20s, 4 in 40s and 1 in 50s) and gender (10 male and 4 female) were asked to attend the matching test relating the synthesize image to the right reference image. The test was performed as the following 2 steps.

Face image generating step: 20 face images (12 female and 8 male) were chosen from the test image database as reference images. Each of the 10 subjects participated the face image generation was given 2 different reference images randomly selected from these 20 images and be asked to synthesize 1 face based on each reference image. Therefore, 20 synthesized face images generated from the 20 difference reference images were obtained.

Image matching step: For each of the 13 subjects participated the image matching test, the 20 pairs of synthesized image and references image were randomly divided into 10 groups. Each group contains 2 pairs of synthesized image and references image of the same gender. Thus, there were 6 female pairs and 4 male pairs. Then, the 10 groups were shown to the subject one by one, and for each group the subject was asked to match

between the generated image and the reference image. Since each of the 13 sub-jects performed the matching task for 10 groups, the total number of trials was 130. Out of which, 100 trials gave a correct matching result. A binominal test showed that the generated images were correctly matched to their corresponding reference image at a significance level above 99%. The result demonstrates that developed system can generate images resembling the reference images.

Fig. 16 shows the best matching image groups (100% matched) in the upper row, and worst matching groups in the lower row. For image C, 6 out of 13 participants correctly matched the reference image with the created image, while for image D, 7 out of 13 participants correctly matched the reference image with the created image.

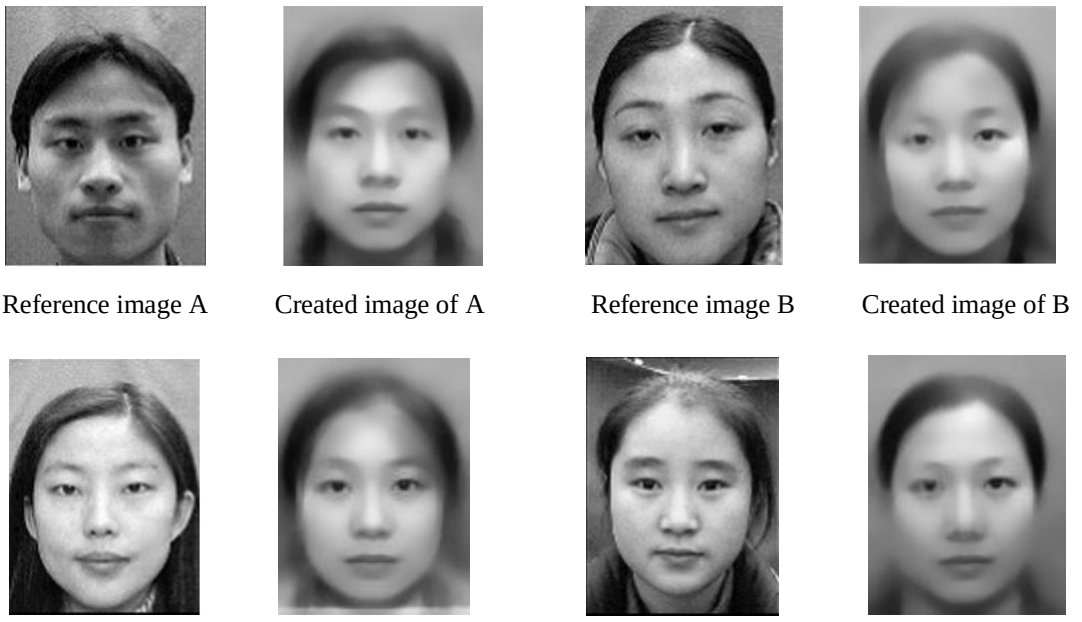


Figure 16**.** The examples used for Matching Test.

3.7 Discussion and Summary

3.7.1 Discussion

The results of the experiment reveal several findings. When the reference images were displayed for 3-4 seconds and then the subjects were asked to create their images without being able to see the original references, the subjects took fewer iterations and less time to arrive at their results than they did when the reference image was presented throughout the whole process. This is likely because the images were only visible to the subjects for a matter of seconds, which made it hard for the subjects to establish a clear, accurate mental picture of the

target face for comparison. Thus, the system-generated images probably created a slight recognition bias in the subjects' minds, leading them to con-verge on their final results relatively quickly. The subjects' evaluations of the resulting images showed several interesting trends, as well. In many cases, the evaluation scores remained relatively constant for several iterations before eventually increasing. This is because the experiment used the three highest-ranking results. Even if the weight of the three images have been shuffled, their relative mutual similarity and central position would have resulted in the same generated face and produced the same score.

Some users reported that sometimes they were satisfied with most parts of the generated face after a few iterations, but unsatisfied with one particular part. The users then continued the iteration process in anticipation of getting a better result for that particular part. But unfortunately, they obtained a globally worse image with other satisfying parts became less satisfied. Although allowing the users to evaluate and control the face as a whole is an advantage of proposed method over the component-based approaches like montage system, it is desirable to improve developed system by allowing the users to locally adjust individual parts.

Another drawback of this method lies in the proposed feature representation. A global feature space based on PCA was employed, which made it impossible to capture the personal detail well, causing the generated face quite similar to the average face. The second method, which will be introduced in the next chapter tries to improve image quality by using Convolutional Neural Network.

3.7.2 Summary

In this paper, a method for the semiautomatic synthesis of a face image from a user's imagination was proposed. By training an OPF based on the user's feedback, developed system successfully creates synthesised images that resembled the face images that users had in mind, but the image quality needs to be improved.

Chapter 4 Based on Conditional Generative Adversarial Network

4.1 Proposed method

As depicted in Fig. 17, the proposed method consists of two parts: (1) a relevance feedback framework for users to generate the landmarks of new candidate faces by evaluating the sample face images and (2) the face image generator using GP-GAN with the new landmarks resulting from the relevance feedback process. In the offline training phase, the algorithm proposed by Dlib [58] is applied to train the network for extracting the landmark features; these extracted landmarks are then used for training the GP-GAN.

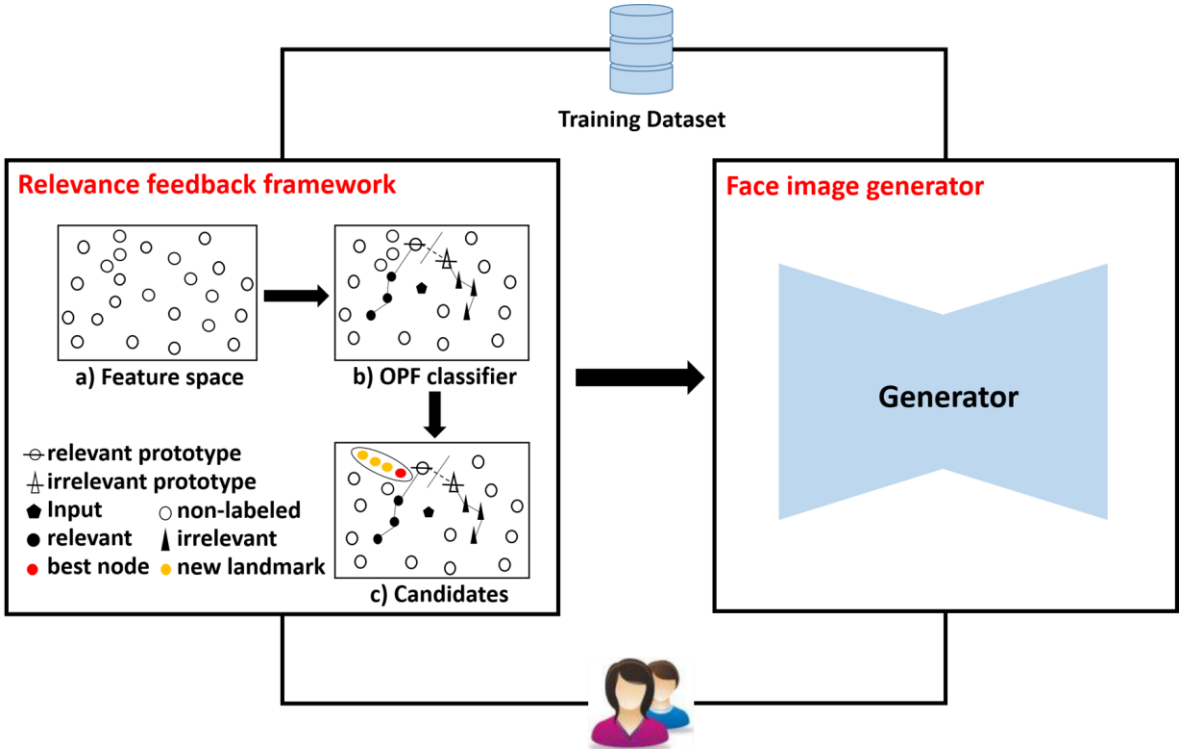


Figure 17. Overview of the proposed method.

The relevance feedback framework consists of three parts: constructing the feature space, training the OPF classifier, and exploring the candidate feature vectors.

The feature space is built based on the extracted landmark features. The OPF classifier is then trained for the feature space based on the relevance feedback, and the user explores the candidate feature vectors by

using the trained OPF classifier. The trained generator generates face images from the new landmarks obtained by interpolating the candidate landmarks. The network of GP-GAN is trained with the landmarks extracted from the training dataset. The final result is created by interpolating the landmarks of the candidate faces using the trained model of GP-GAN during the running stage. Fig.18 shows some results generated with the proposed system.

Figure 18. Results from proposed method of the second method. The images in the first row are the references, and those in the second row are the corresponding created face images.

The above relevance feedback framework relies on two core techniques: training an image classifier based on the user's feedback and exploring the candidates in the landmark feature space. The two following subsections describe the details of the two techniques, respectively.

4.2 Relevance feedback framework

As depicted in Fig. 19, the relevance feedback framework is realized in the following five steps:

Step 1: The system randomly selects 10 face images from the training dataset and shows these images to the user as the initial set. The user chooses the most similar image from the initial set, and then the system shows 10 images that are most similar to that selected by the user in order to initialize the relevance feedback process.

Step 2: The user labels each showed image as similar (○) or not similar (×).

Step 3: Based on the user's feedback, the system updates the classifier and explores some candidate landmarks in the feature space, generating face images from them using GP-GAN and then presenting the generated results to the user

Step 4: If the user feels that the generated results consist of images similar to the desired face to some extent, the user selects up to k most similar faces from these images. The selected images are then added to the candidate set. Otherwise, the user can choose to continue the iteration. The system shows 10 images near the border of the classifier to the user, and the system goes back to step 2. Otherwise, the system proceeds to step 5.

Step 5: The user can specify the degrees of similarity to the images in the candidate list. The landmarks of these images are interpolated using the degrees of similarity as the weight to produce the landmark to be sent to GP-GAN for producing the resulting face image.

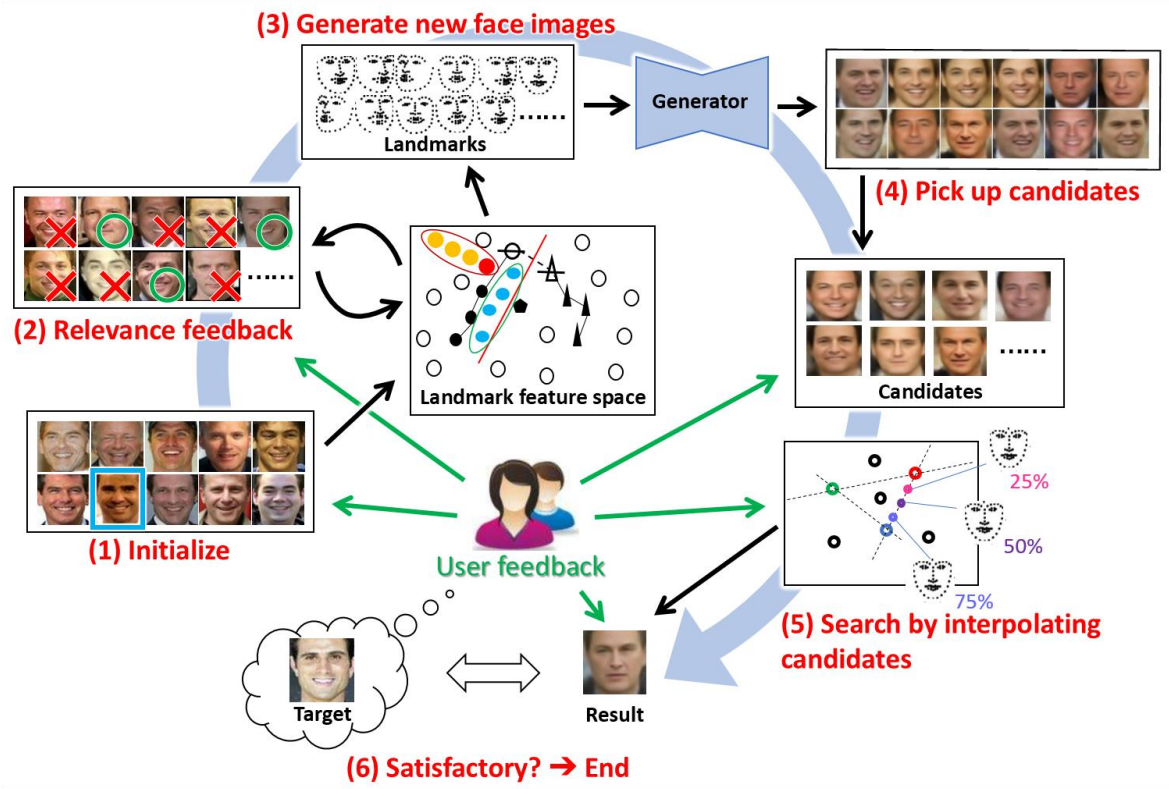


Figure 19. Relevance feedback framework.

4.2.1 Training the OPF classifier

The user's intention is reflected by first training a face image classifier based on the user's feedback. The OPF classifier is used for training the classifier. The procedure of training OPF classifier is the same as the first method, the details as described in 3.3. At each iteration of the relevance feedback, the user labels the 10 samples as relevant or irrelevant according to the similarity with the target face image. The border nodes, which are the face images to be returned to the user for labelling at each iteration, are also defined as non-prototype nodes that belong to the relevant class and with the smallest ratio of the cost from these nodes to the relevant prototypes over the cost from these nodes to the irrelevant prototypes; The best node is used to explore the candidates for synthesizing the final result, which is also defined as the relevant sample located farthest from the irrelevant prototype and closest to the relevant prototype. The implementation algorithms for border nodes and best node are also the same as the first method introduced in in 3.3.

4.2.2 Creating the candidate landmarks

In order to create the candidate landmarks of the desired face, the best node is moved along a certain direction vector $\vec{v}$ so that it becomes closer to the relevant prototypes and farther from the irrelevant ones. There are three core issues here: the direction in which the best node should be moved along, the step size of movement, and the valid range of distance to be moved.

The direction of movement

To obtain the direction vector $\vec{v}$ of movement, two composited vectors are computed: the composited vector $\overrightarrow{v_r}$, which is the summation of vectors from the best node to all relevant prototypes, and the vector $\overrightarrow{v_{ir}}$, which is the summation of vectors from all irrelevant prototypes to the best node. For example, Fig. 20 shows that given the best node depicted by the red circle ●, two relevant prototypes depicted by ⊖, and two irrelevant prototypes depicted by ⧋, first, composite vector $\overrightarrow{v_r}$ is calculated by computing the summation of vectors from the best node to the two relevant prototypes. Second, another composite vector $\overrightarrow{v_{ir}}$ is calculated by computing the summation of vectors from the two irrelevant prototypes to the best nodes. Finally, the direction vector $\vec{v}$ of movement is obtained as the composited vector of $\overrightarrow{v_r}$ and $\overrightarrow{v_{ir}}$.

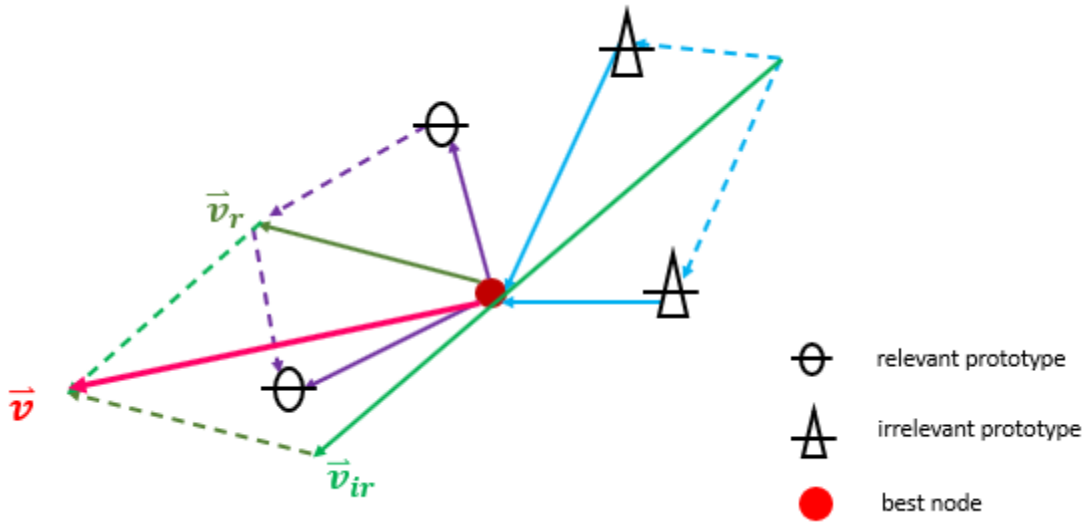


Figure 20. Direction of movement.

Assume there are n number of relevance prototypes and m number of irrelevance prototypes denoted as $r_i(i = 1, 2, 3, \ldots\ldots, n)$ and $ir_i(i = 1, 2, 3, \ldots\ldots, m)$, respectively. With the best node denoted as b, the direction vector $\vec{v}$ can be computed as follows:

$$\vec{v} = \vec{v_l} + \vec{v_{lr}} , \quad (10)$$

$$\vec{v_l} = \sum_{i=1}^{n} \overrightarrow{v_{b \to b}} , \quad (11)$$

$$\vec{v_{lr}} = \sum_{i=1}^{m} \overrightarrow{v_{tr_l \to b}} , \quad (12)$$

The step size of movement

The step size of movement is a critical parameter in this study, as it controls how much the result is changed each time. Choosing an appropriate distance of movement is difficult, as a value that is too small may make the change not obvious enough, whereas a value that is too large may result in skipping the optimal result. The sensitivity of the step size should depend on the extent of the dataset in the feature space. Therefore, the diagonal line of the bounding box of the training dataset in the feature space is computed at first. The length of the diagonal line of the bounding box is denoted as l and it can be computed as follows:

Assuming there is a training dataset P which contains n face images $p_i \in P(i = 1, 2, 3 \dots \dots n)$, each of which is represented with a m dimensional feature, denoted as $p_i = (p_{i,0}, p_{i,1}, p_{i,2} \dots \dots p_{i,m-1})$. In this study, $n = 13{,}233$ and $m = 136$. Let p_j^{max}, p_j^{min} $(j = 0, 2, 3 \dots \dots m - 1)$ be the maximum and minimum coordinate among all n faces in j dimension, then l can be computed as

$$l = \sqrt{\sum_{j=0}^{m-1} (p_j^{max} - p_j^{min})^2}, \quad (16)$$

When moving along $\vec{v}$ as shown in Fig. 21, the step size of movement is given as l_i which is calculated as follows:

$$l_i = \alpha \times l, \quad (17)$$

where α is a parameter controlled by the user.

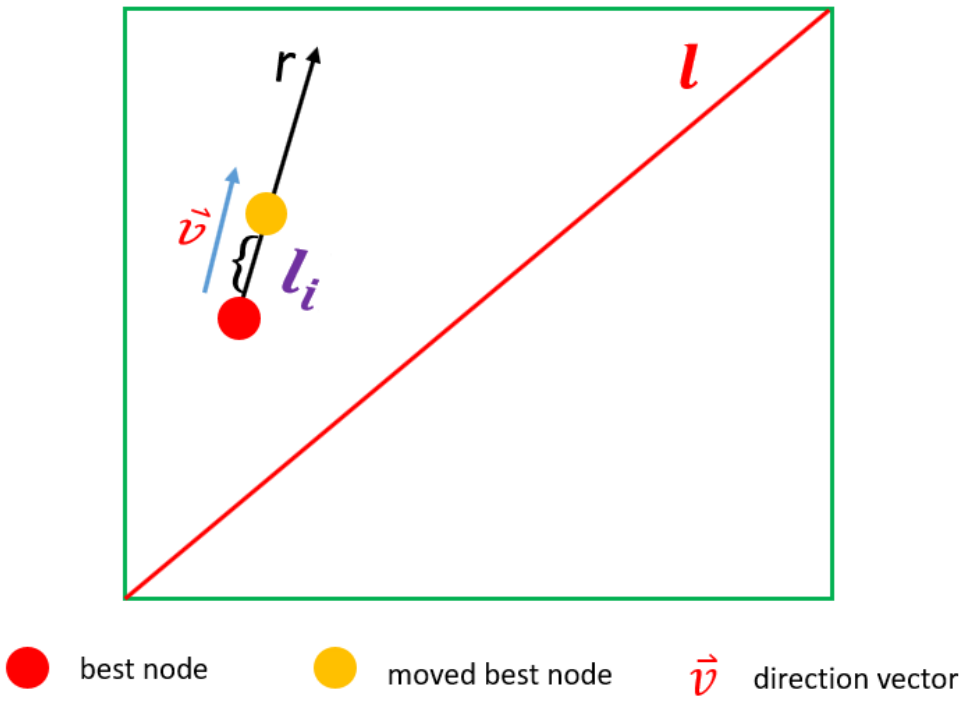


Figure 21. The step size of movement.

The distance of movement

For the distance of movement, the bounding box of the dataset is used for limiting the exploration in a safer area within which any point is likely represents the landmark of a face. A face image can be obtained by inputting the point to GP-GAN. The bounding box can be obtained by calculating the maximum and minimum coordinates of all n faces in the training dataset in each dimension. In this study, the bounding box is obtained from the training dataset P which contains 13,233 face images with 136 dimensions in the feature space. Empirically, it is found that a point located out of the bounding box of the training dataset has a high probability of not defining a face (as depicted in Fig. 22). Therefore, when moving along $\vec{v}$, it is validated that whether the new position is still within the bounding box. If it already exceeds the bounding box, the exploration is stopped. To ensure that the position after movement is still in the bounding box, the coordinate of each dimension of the moved position is considered. Specifically, if the coordinate in a certain dimension is larger than the maximum coordinate or smaller than the minimum coordinate of the corresponding dimension, the movement will stop. Assuming that $A\,(a_0\,, a_1, a_2 \ldots\ldots\, a_{135})$ is moved position, p_j^{max} , p_j^{min} $(j = 0, 2, 3 \ldots\ldots 135)$ are the maximum and minimum coordinate among all 13,233 faces of training dataset along j dimension. A must satisfy equation (18) so as it can be moved, otherwise, the moving procedure ends.

$$p_j^{min} \le A_j \le p_j^{max}, \qquad (18)$$

where $p_j^{max} = \text{maximum}(p_{1,j}, p_{2,j}, p_{3,j} \dots\dots p_{13233,j})$, $p_j^{min} = \text{minimum}(p_{1,j}, p_{2,j}, p_{3,j} \dots\dots p_{13233,j})$ $(j = 0, 2, 3 \dots\dots 135)$.

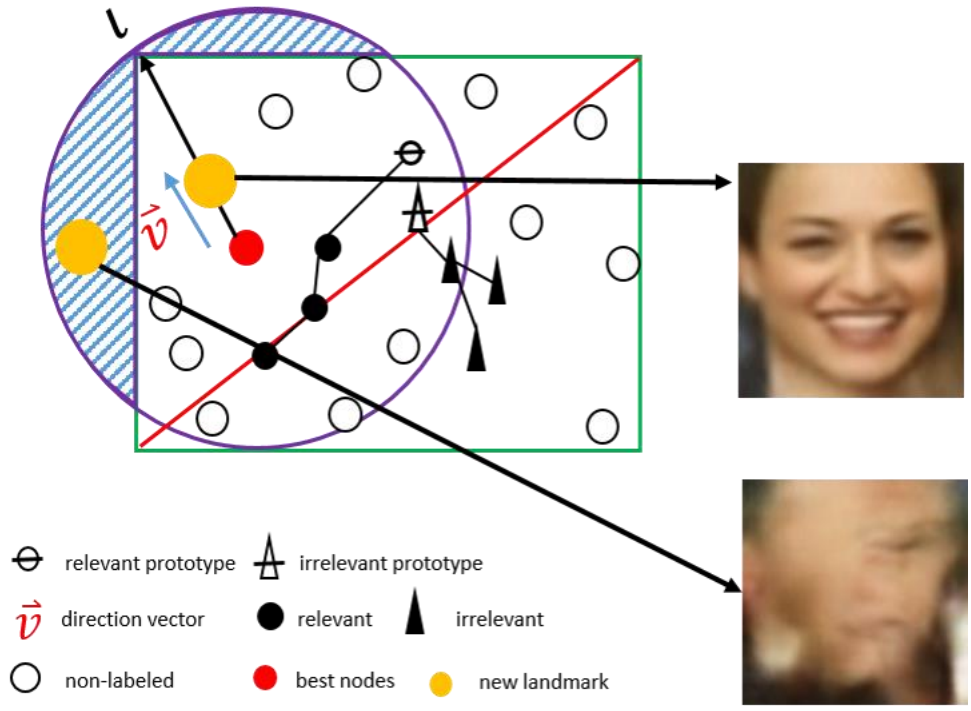


Figure 22. Unsafe area for creating new landmarks

4.3 Generative model for synthesizing face images (GP-GAN)

The purpose of GP-GAN [27] is to synthesize faces from their respective landmarks. As shown in Fig. 23, similar to the traditional GAN network, GP-GAN consists of two components: generator G and discriminator D in which G is designed based on the U-net [41] and DenseNet [43] architectures to leverage the advantages of these architectures and D is devised under the patch-based principle. Given a landmark, G does its best to generate the corresponding face images, whereas D tries its best to distinguish between real data and the generated images. Unlike other GANs, the network of GP-GAN adds a perceptual sub-network (based on VGG-16 architecture) and a gender-preserving one in addition to the discriminator. The model is learned by adversarial loss, perceptual loss, and a gender-preserving loss by minimizing the following objective function:

$$L = \mathcal{L}_{adv} + \lambda_P \mathcal{L}_p + \lambda_C \mathcal{L}_{cls} + \lambda_1 \text{L1} \ . \qquad (19)$$

Here, $\mathcal{L}_{adv}$, $\mathcal{L}_p$, and $\mathcal{L}_{cls}$ represent the adversarial loss, the perceptual loss, and the gender-preserving loss, respectively. The adversarial loss $\mathcal{L}_{adv}$ is based on the discriminator sub-network D, the perceptual loss $\mathcal{L}_p$ guides the generator using the L1 distance between the high-level features extracted from the VGG-16 [38] network, the gender-preserving loss $\mathcal{L}_{cls}$ measures the difference of the gender feature of the produced image and the real image, and L1 defines the error between the target and the generated image. The corresponding weights of the losses are λ_P, λ_C, and λ_1.

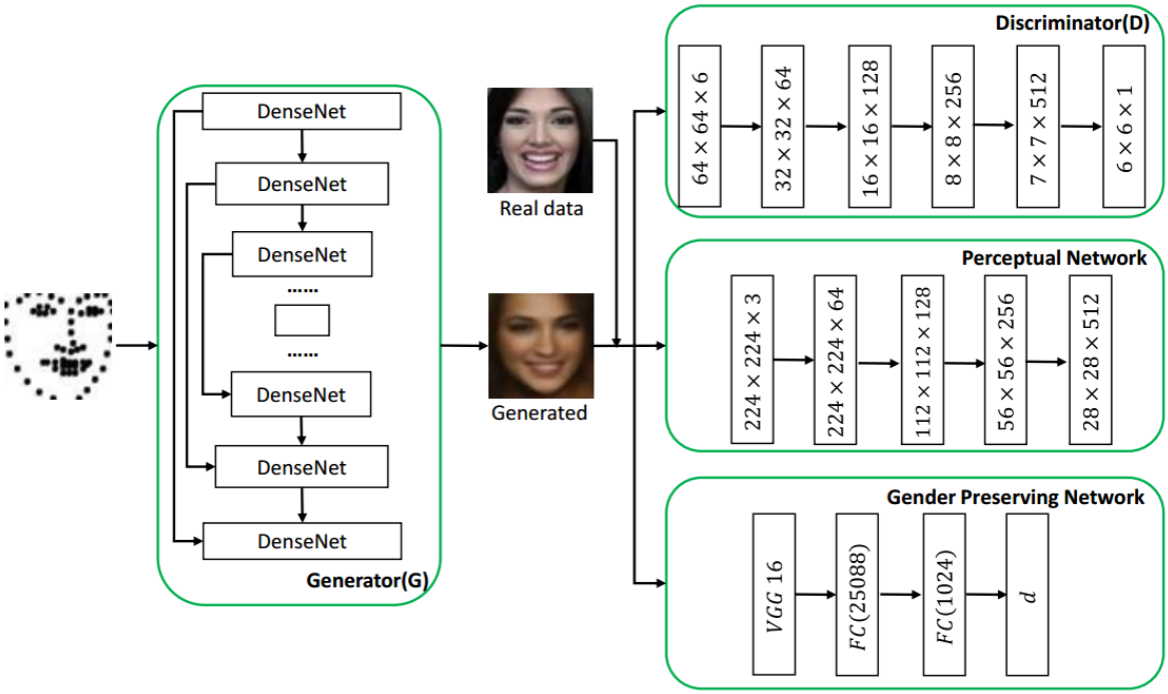


Fig 23. Structure of GP-GAN

4.4 Experiment and evaluation

4.4.1 Datasets and implementation details

For the generative model of GP-GAN, its parameters were learned based on the whole Labelled Faces in the Wild (LFW) dataset [53], [53], which contains 5,749 identities and 13,233 face images. There are now four different sets of LFW images—the original and the three different types of aligned images. The aligned sets include funnelled images (ICCV 2007), LFW-a, and deep funnelled images (NIPS 2012). The model of GP-GAN is trained based on official deep funnelling aligned [52], [53], and it uses the official training, validating, and testing View 1 in this experiment. The details of the training can be found in [27]. However, to obtain high-quality results, some parameters were adjusted for achieving the best learning rate and the number of epochs.

In the current implementation, the model of GP-GAN is trained on a single GTX 1070 GPU for approximately more than 40 hours (800 epochs). Landmark images are represented as black solid dots on a white background. During the training stage, both a landmark image and its corresponding real data are inputted into the network of GP-GAN. It was found that the appearance of the black dots can largely affect the performance of the trained GP-GAN model. Probably because of the aliasing, the resulting face image from the generator is not ideal when landmarks are presented with a binary image. To address this issue, a grayscale image is used, making the center of each dot the darkest and then changing it to white gradually toward the edge in order to achieve a good anti-aliasing effect. These kinds of grayscale landmark images are used for both training and testing.

The LFW dataset consists of the face images of different head poses and different expressions. As the geometric difference between two different poses is much larger than the geometric difference between the facial parts of two different persons, the pose feature is much more dominant than the shape feature of individual parts (e.g., the size and shape of the eyes, and nose). Therefore, if the faces of different head poses are included when training the OPF with relevance feedback, the detailed features of individual parts tend to be neglected. Similarly, there is a large geometric difference between an open mouth and a closed one. Basing on these observations, when training the OPF classifier, a total of 1,000 frontal face images are chosen from the LFW and divided these into a sub-dataset by gender and an open/closed mouth. For each subset, 80% of the face images are used for training, and the remaining 20% are used for testing. During the runtime relevance feedback process, the initially chosen 10 images consist of the faces of different genders and open/closed mouths. When the user selects the most similar image, the subset of training samples consisting of the selected face is automatically loaded for constructing the OPF.

In the current implementation, $k = 3$ is set for step 4 of the relevance feedback framework in Section III. Therefore, the user is allowed to select up to three most similar faces from the generated face images for addition into the candidate set at each iteration.

4.4.2 Experiments

Three types of experiments are conducted to validate the effectiveness of the proposed method. The first experiment invited participants to create face images, and the second experiment had another group of participants evaluate the generated results. The third experiment aimed to compare the proposed method with the first method. In all the experiments, the participants were asked to ignore the hairstyles in the face images.

Experiment for creating face images

This experiment includes three tasks—to generate face images based on the reference image, to generate face images according to the briefly presented reference image, and to synthesize the imagined face images. The second task is particularly designed by assuming forensic applications, such as assisting the police to create

face images of criminals for forensic purposes. The participants scored their generated results according to the image's similarity to the reference image on a six-point scale (0: not at all; 1: no resemblance; 2: very weak resemblance; 3: neither weak nor strong resemblance; 4: somewhat strong resemblance; 5: strong resemblance). The number of iterations taken before obtaining a satisfactory image was recorded to evaluate the performance of the proposed method. Ten participants (8 males and 2 females in their 20s–25s) joined all the three face images creating tasks.

Task 1, Creating face images based on the reference images

In this experiment, a reference face image was presented to each participant during the entire relevance feedback process and asked him/her to create a face image similar to the reference image. A total of 20 images that were randomly excluded from the dataset used for training GP-GAN and OPF were used in the experiment. The participants were asked to perform the task with all 20 images, and they were also required to score their created results. Some results are shown in Fig. 24. These will be evaluated in the second experiment.

Figure 24. Created face images based on the reference image. First row: reference images. Second row: the corresponding created face images.

Task 2, Creating face images based on the briefly presented reference images

This task is performed with the aim of validating whether the proposed method enables the synthesis of an image in the user's memory. A reference face image was presented to each participant for 3–4 seconds and had the participant create a face image resembling the reference image and then score the generated image. Twenty face images randomly excluded from the training set and are different from those used in task 1 were utilized as the reference images. Fig. 25 shows that the resulting images can capture the overall features of the reference face images, as well as some shape information of individual parts (e.g., the size and, shape of the eyes, noise, and mouth).

Figure 25. Created face images based on the briefly presented reference images. First row: the reference images. Second row: the corresponding created face images.

Task3, Creating the imagined face images

Unlike the two previous tasks, in this case, a participant is asked to create an image that he/she imagined without any reference image and then to score the result. With this third task, as no one knows about the face imagined by the participant, the scores of the results by the participant are the only measure for evaluation. Fig. 26 shows some results from this task.

Figure 26. Created face images that were imagined.

Fig. 27 shows that the average scores for all tasks are very similar and are all close to 4, which is much better than the scores of the existing method combining GAN with the genetic algorithm [32]. As proposed method uses a relevance feedback framework, the convergence speed is also an important measure for evaluating the performance of the system. Fig. 28 depicts that task 1, in which the reference images are presented during the whole process, took more iterations than the two other tasks. It is observed that this is mainly because the participants can check all the feature information better and more carefully during the entire process. The iteration number of task 3 is larger than that of task 2; this is likely because the mental image of the face may be influenced by the results during the relevance feedback, and it takes some time for the user to make the conclusion. Even so, the user can arrive at satisfactory results within an average of five iterations. This is much faster than that obtained with existing evolutionary algorithm-based approaches [8], [15].

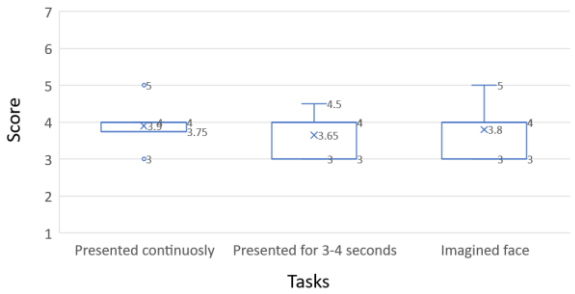


Figure 27. Final scores.

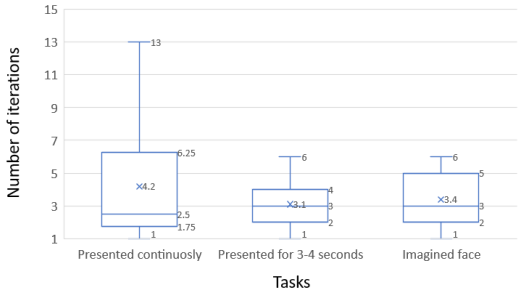


Figure 28. Number of iterations.

Fig. 29 illustrates the number of iterations of each participant in different tasks. The horizontal axis represents the participant, and the vertical axis represents the number of iterations. It took less than six iterations for 11 participants to arrive at the final results. In task 1, however, one participant took more than 12 iterations before she reached satisfactory results. The interview after the experiment reveals that although the participants obtained satisfactory results at the fourth or fifth iterations, they still attempted to perform further iterations because of their curiosity about what happens next if more iterations are conducted.

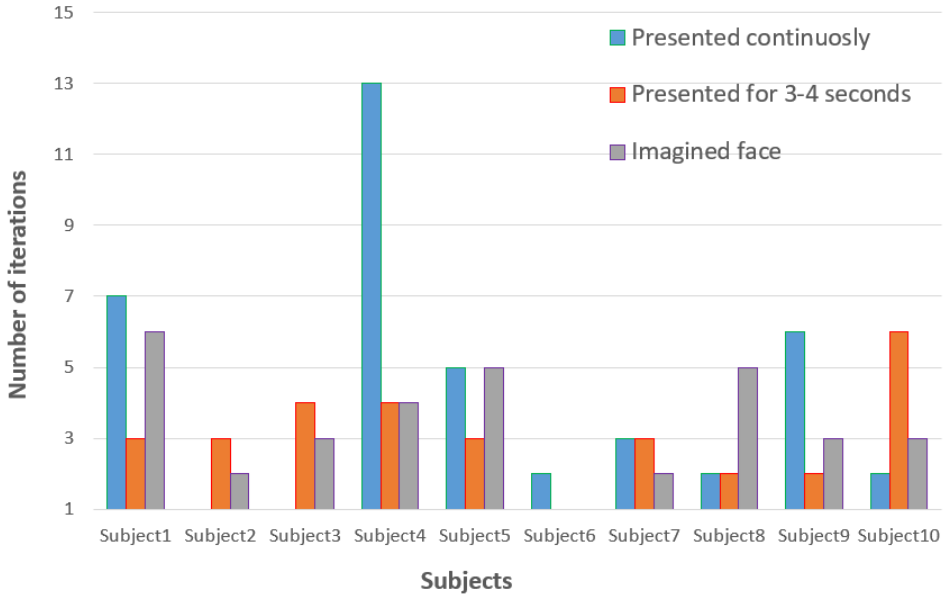


Figure 29. Number of iterations required for achieving satisfactory results.

Experiment for evaluating the generated face images

In this experiment, another group is asked to evaluate the face images generated in experiment 1 by finding the corresponding reference image of the generated image from a set of candidate images. Ten participants (7 males and 3 females, in their 20s–25s) who were different from those who joined experiment 1 were invited to perform the present experiment. As shown in Fig. 30, for each face image generated with experiment 1, three face images are presented to the users. Among the three images, one is the reference image used for creating the result, and the other two are selected from the training dataset, which are the faces closest to the reference image in the landmark feature space (based on L2 distance). The display positions of the three images are randomly shuffled to eliminate bias caused by the layout. The participant was asked to find the reference image from the three images. The circled face image in answer paper is selected by the user as the reference image. Ten images randomly selected from the results (as shown in Fig. 32) in tasks 1 and 2 of experiment 1 were used for evaluating. The average success rates of all 10 participants are shown in Fig. 30, with the horizontal axis representing the 10 generated images.

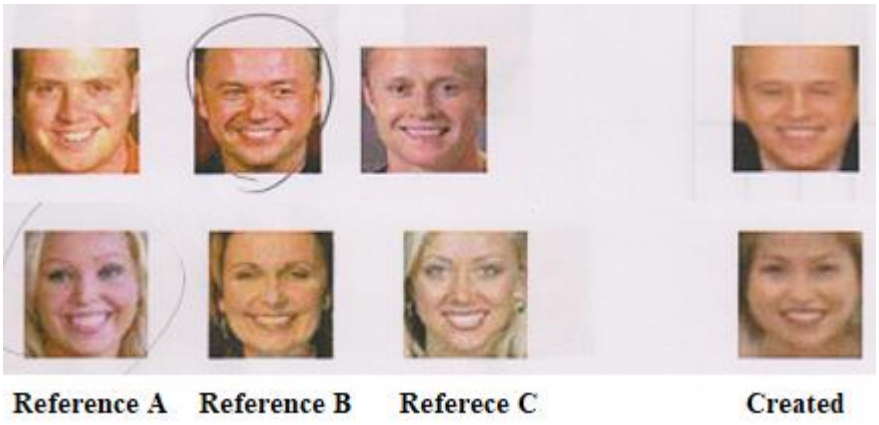


Figure 30. An example of experiment 2.

Fig. 31 shows that among the 10 generated images, three have a 100% matching success rate, and the lowest success rate is 60%. The phenomenon that the matching success rate of females is lower than that of males was also observed. This result may be related to the hairstyle. Although during the entire procedure of the experiment, all subjects were asked to ignore the hairstyle, they might have still been involuntarily affected. The subjects commented that the eyes were the most important feature that drew their attention.

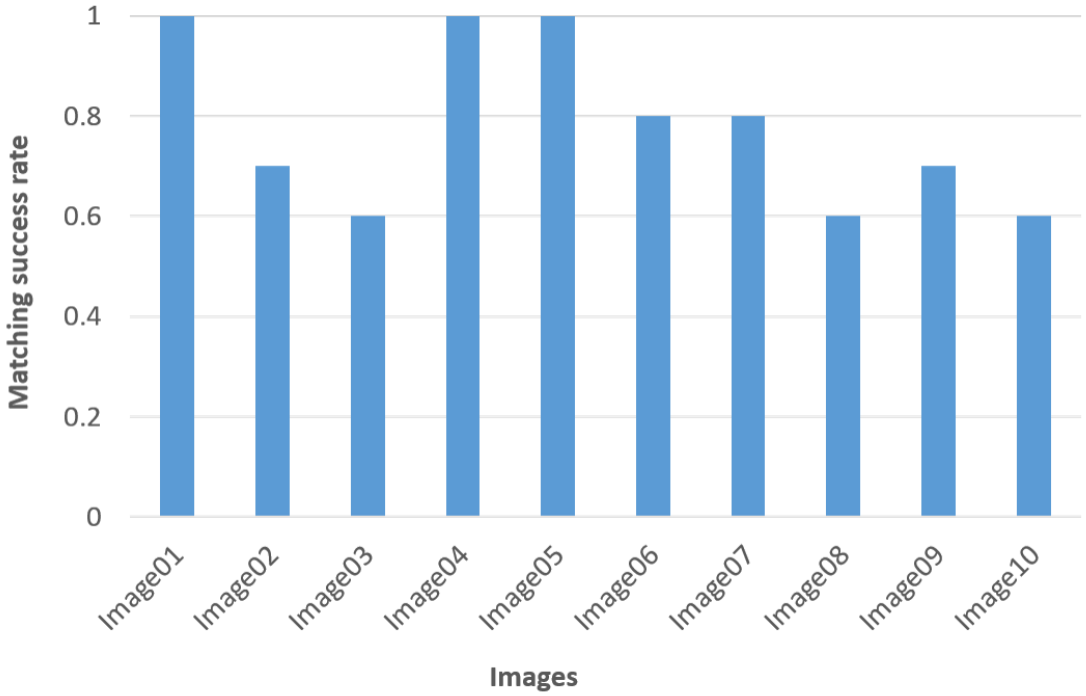


Figure 31. Matching success rate.

Figure 32. Results from experiment 1 are used for evaluation. The upper row is the reference face image and the lower row is the corresponding generated images with the proposed method.

The experiment for comparing with the first method

This experiment compares the proposed method with the first method based on PCA in terms of the number of iterations, image quality, and similarity to the target face. The same dataset is used to ensure a fair and objective comparison. As first method can only generate grayscale images, all images in the training dataset used for OPF were converted into grayscale ones at first and then resized to be the same size as the images used in first method. Next, facial features were extracted from the pixel level, and the PCA algorithm [54] was applied to reduce the facial features to 80 dimensions and then for training the OPF classifier, same as the first method.

To compare iterations and image quality, the participants who joined experiment 1 were invited to perform all the three tasks as those in experiment 1 with first method. Fig. 33 shows some results of the two methods; the first row presents the reference images, whereas the second and third rows present the images generated using second method and the first method, respectively.

Figure 33. Results using different methods. First row: the reference images. Second row: the generated face images using second method. Third row: the generated face images using the first method.

Fig. 34 shows the comparison results on the final scores of the three tasks, in which the horizontal axis represents the task types and the vertical axis represents the score. It can be found that the average score of the proposed method is higher than that of the first method for all tasks.

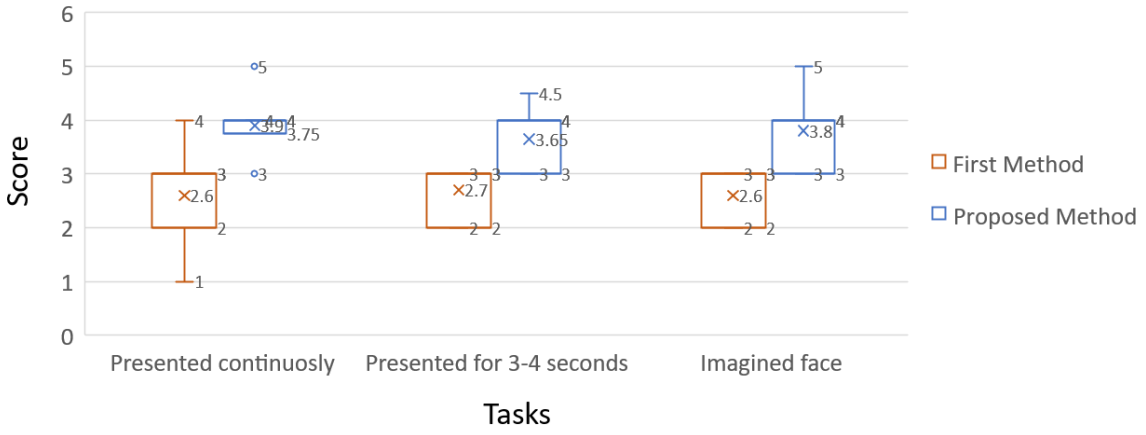


Figure 34. Comparison of the final scores.

Fig. 35 shows the number of iterations. The first method took more iterations than proposed method in all tasks. The proposed method outperforms the first method particularly for task 1, in which the participants can always compare the results with the reference image during the entire process. It also can be found that the average number of iterations of the proposed method is less than that of the first method for all tasks.

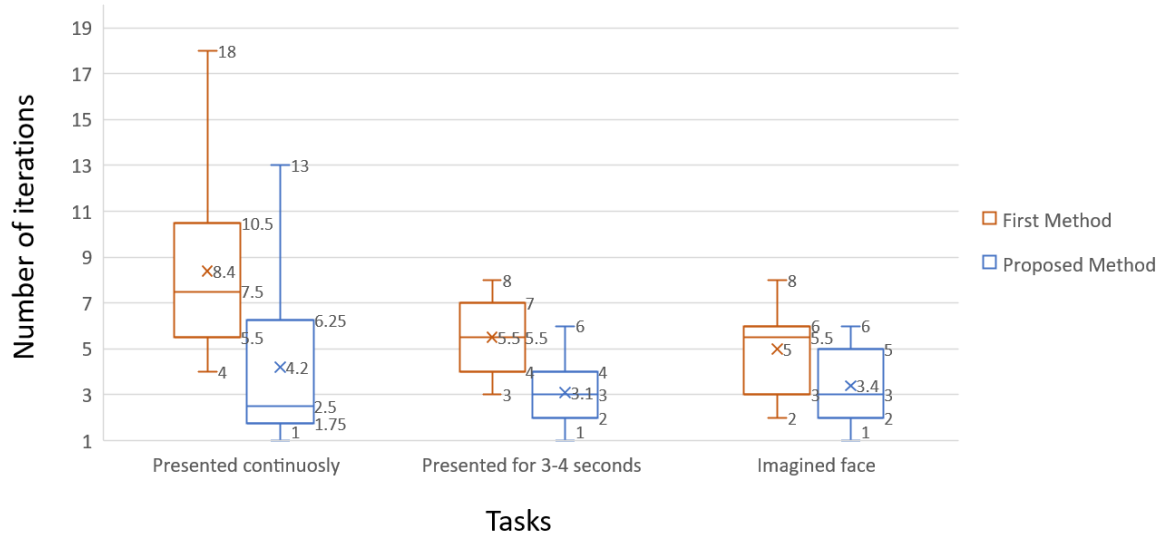


Figure 35. Comparison of the number of iterations.

For the similarity comparison, a new group of participants is invited who did not join any face image generation or matching tests to evaluate the face images created by the two methods; similar to experiment 1, a five-point scale was used based on the images' similarity to the reference images. Twenty subjects were randomly divided into two groups. The first group, consisting of 10 participants (10 males in their 20s–25s),

evaluated the results with the first method, whereas the second group, consisting of 10 participants (10 males in their 20s–25s) evaluated the results with second method. The participants were not asked to directly compare the results of the two methods, and non-overlapping groups of participants is invited to evaluate the results of the two methods separately because we wanted to focus on the evaluation of similarity and avoid any adverse evaluations of the first method caused by the low image quality. As the first method could not synthesize colour image, the face images were converted by our method into grayscale ones and then presented these images to the participants to eliminate the effect of colour. The participants were also asked to ignore the blur artifacts. Fig. 36 presents the result of the similarity comparison for all the reference images. The results generated by second method are evaluated to be more similar to the reference images. The experimental data for similarity score comparison can be found in Fig. 37.

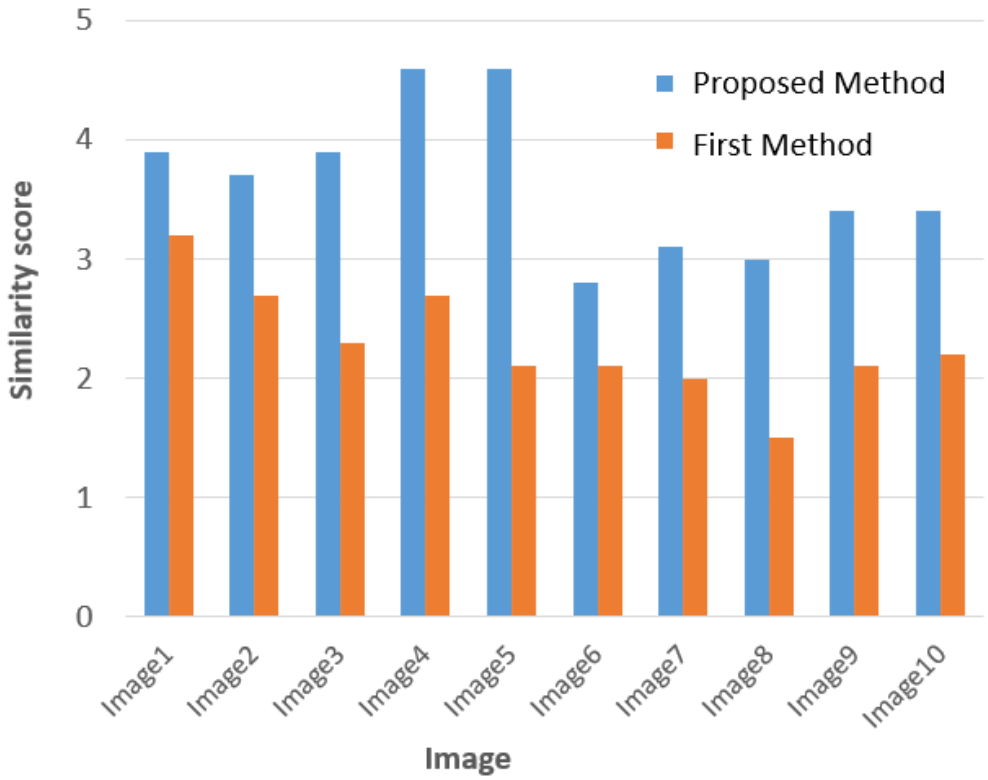


Figure 36. The similarity.

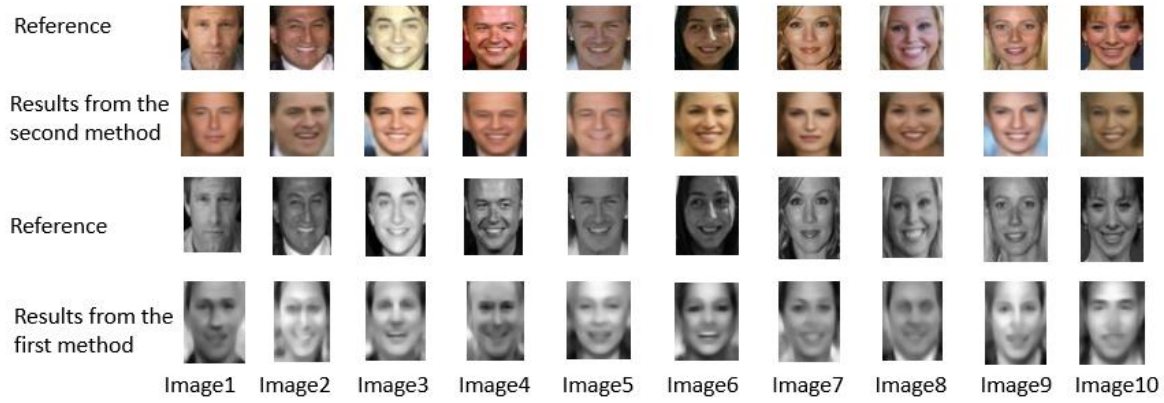


Figure 37. Experimental data for similarity score comparison. The upper row is the reference face image and the lower row is the corresponding generated images from the first method and the second method.

4.5 Discussion and Summary

4.5.1 Discussion

As shown in Fig. 18, Fig. 24, Fig. 25, and Fig. 26, most of the created images can well resemble the geometric features of the reference images, but they failed in capturing the details of texture features. For example, in Fig. 18, the wrinkles on the faces of Laura Bush and George W. Bush were not reproduced in the resulting image. This is because the GP-GAN model is controlled with a landmark that consists of the geometry information of facial parts only. To address this issue, a new GAN model is proposed in the next chapter trying to generate face images with high frequency textures. The participants of the experiments reported that when they focus on some particular parts, quickly converging to an image with that part resembling the reference image is possible. However, that particular part may become less similar to the reference image again after trying to improve the other parts. It is important to allow users to control each part independently and integrate the best results of all facial parts. The overall results are still blurred, although they are much better than those of the first method that used the PCA feature. The image quality may be further improved by carefully tuning the training parameters of GP-GAN. Nevertheless, contribution of proposed method is the approach of combining GAN with an effective relevance feedback framework; substituting GP-GAN with any state-of-the-art GAN model for a better image quality is not difficult. In the current implementation, the bounding box of all the training images in the landmark feature space is treated as the safe area for exploring new landmarks, but this is an approximated approach. A more accurate scope needs to be defined. Currently, the face images generated from the newly created landmarks are not included in the dataset for training the OPF. By adding the created face images to the training dataset at every step of the iteration, it is possible to expand the range of face images.

4.5.2 Summary

This study proposed a novel method to gain user control over face features by combining newest GAN model with the relevance feedback framework based on the OPF algorithm.

The experiment results demonstrated that the proposed method can be used to generate not only a face image resembling the target face but also a face image in the user's memory or imagination. The proposed method makes up for both the lack of user intervention in GAN and the low image quality in traditional methods.

Chapter 5 Generative Adversarial Network for synthesizing faces with the user's desired texture features from high-frequency features

5.1 Proposed framework

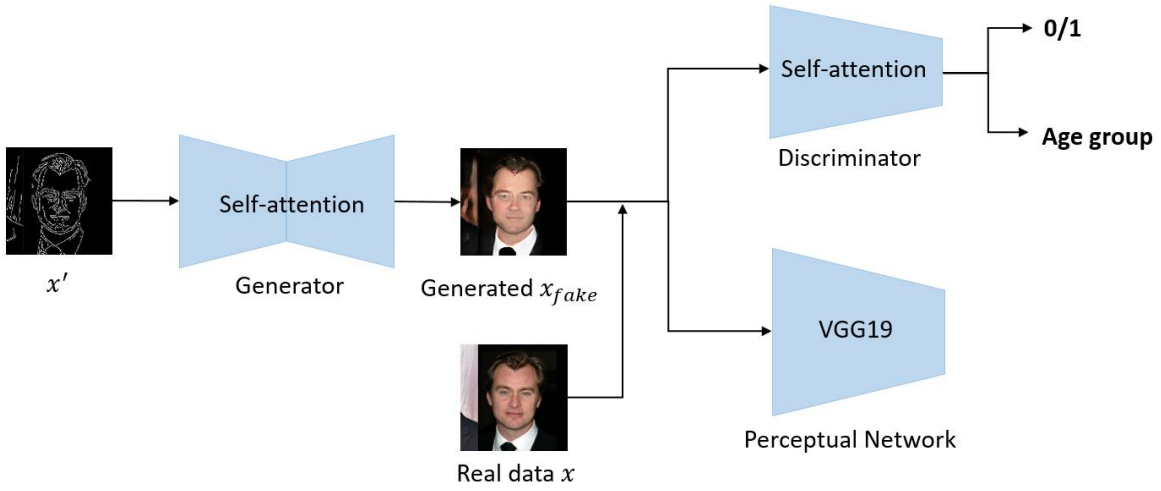


Figure 38. Overview of HF-GAN

As shown in Fig. 38, the proposed Generative Adversarial Network for synthesizing faces with user's desired texture features from high-frequency features (HF-GAN) is a framework to generate corresponding face image from its high-frequency features. It consists of a generator, a discriminator, and a perceptual network. The generator attempts to generate corresponding face images from the input image with given high-frequency features. In fact, the generator can not only generate face images but also learn the corresponding age group of high-frequency features while generating face images. The discriminator tries to distinguish between the real image and the generated image, and also classify the input images to its corresponding age group. The generator and the discriminator are trained simultaneously. The former is guided by adversarial loss, classification loss, reconstruction loss and perceptual loss. The latter is guided by adversarial loss and classification loss. The perceptual network is used to extract high-level features, which equals to the 5th layers of a pre-trained VGG-19 [38] network. The L1 distance between the extracted features of real data and generated data is used to guide the generator.

In particular, to learn a generator which can generate corresponding face images with the given high-frequency features extracted from the input image, the training steps are as follows:

Step1: High-frequency features are extracted using technology of edge detection from a real image x and drawn with white on a 128×128 image black background, resulting high-frequency feature image denoted as x'.

Step 2: Given a high-frequency image x', the generator tries to generate a corresponding face image x_{fake}.

Step 3: Both generated image x_{fake} and real image x are inputted into the discriminator. The discriminator tries to distinguish them and classify them to their corresponding age group c.

In addition, both the generator and the discriminator contain self-attention mechanism which will be introduced in Section 5.1.3.

5.1.1 Generator

U-Net [41] adopts longer skip connections to preserve low-level features. It can overcome vanishing gradient problem to some extent. Therefore, U-Net is employed for the generator. As shown in Fig. 39, a high-frequency feature image goes through three convolution layers, followed by six residual block layers, three transposed deconvolution layers, and finally, a generated image is obtained. The kernel of the first convolutional layer is 7×7 with stride 1. The second and third convolution layers have a kernel of 4×4 with stride 2. The first and second deconvolution layers have a kernel 4×4 with stride 2 . The kernel of the third deconvolution layer is 7×7 with stride 1. In addition, an attention layer is added in the middle of the six residual blocks. We also adopt instance normalization for all layers except the output layer.

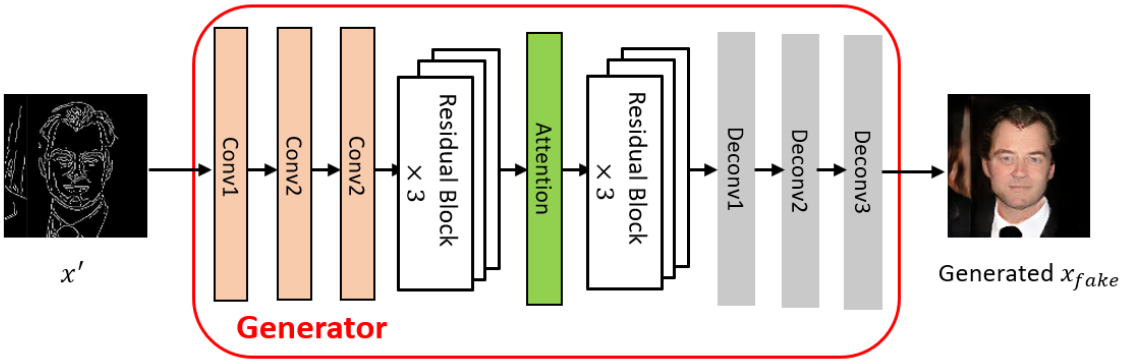


Figure 39. Architecture of generator

5.1.2 Discriminator

Discriminator is based on AC-GAN [55] and patch-based discriminator [23]. The attention mechanism is also added to the residual blocks same as the generator. Thus, the ability of discriminator is enhanced and the ability of generator is also strengthened indirectly. That improves the quality of generated image. As shown in Fig. 40, both generated image and real image goes through a convolutional layer, followed by six residual block layers, two transposed deconvolution layers at the same time. Finally, the probability is obtained, which x_{fake} came from the real image rather than generator as well as the age group of x_{fake}. The kernel of the first convolutional layer is 4×4 with stride 2. The kernel of the first and second deconvolution layer are 4×4 with stride 2 too. Same as generator, an attention layer is added in the middle of the six residual blocks and instance normalization is adopted for all layers except the output layer.

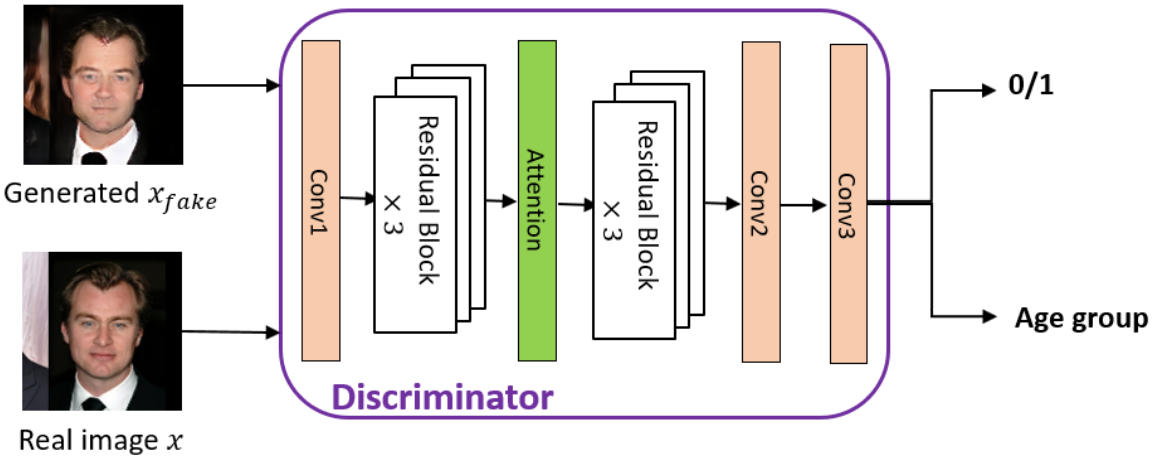


Figure 40. Architecture of discriminator

5.1.3 Variants of residual block

As shown in Fig. 39 and Fig. 40, in this work, the self-attention mechanism is incorporated into the GAN framework. In this way, the generator can generate images which contain more detailed features of face area, and the discriminator can check highly detailed features on the face area.

As shown in Fig. 40, the feature maps $x \in \mathbb{R}^{C \times N}$ from the previous residual block are inputted into the attention layer, where C is the channel number and N is the number of the feature maps. x is mapped to feature spaces f and g to calculate the attention weights, where $f(x) = W_f x$ and $g(x) = W_g x$, W_f and W_g are learned weight matrices by 1×1 convolutions.

$$\beta_{j,i} = \frac{\exp(S_{ij})}{\sum_{i=1}^{N} \exp(S_{ij})}, \tag{20}$$

where $S_{ij} = f(x_i)^T g(x_j)$, $\beta_{j,i}$ represents the extent to which the model attends to the i^{th} location when synthesizing the j^{th} area. The attention layer outputs $O_i(i = 1,2,3 \dots \dots n) \in \mathbb{R}^{C \times N}$

$$O_j = v\left(\sum_{i=1}^{N} \beta_{j,i} h(x_i)\right), \; h(x_i) = W_h x_i, \; v(x_i) = W_v x_i, \tag{21}$$

where W_g and W_h are learned weight matrices by 1×1 convolutions. The output of the self-attention layer is multiplied by a scale parameter γ and added to the output of residual layer $f_0(x; w_{f_0})$. The final output is:

$$y = \gamma\alpha + f_0(x; w_{f_0}), \tag{22}$$

where γ takes values from 0 to 1. In this work, it is initialized to 0 and increased during training. This makes the model focus on the local area at first, and then gradually consider long-range dependencies.

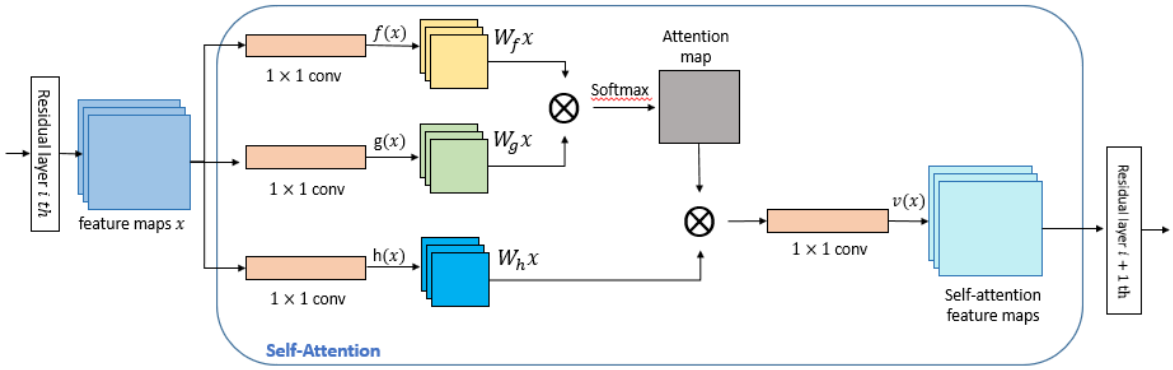


Figure 40. Architecture of self-attention. The self-attention is added between the residual layers. The ⊗ indicates matrix multiplication. The softmax operation is performed on each row.

5.1.4 Training objectives

Our goal is to train a generator G that can generate face image with desired texture information from the input high-frequency features. In order to achieve this, generator G is trained to translate an input image x' into an output x_{fake}, sampled from the generator, controlled by high-frequency features. We train a single discriminator D to output the probability distributions of both the real image x and fake one x_{fake}, and their corresponding age group c.

Overall objective: the model is trained for learning parameters by minimizing two objective functions, the equation (23) and equation (24), which are used to optimize the discriminator D and generator G respectively:

$$\mathcal{L}_D = \lambda_1 \mathcal{L}_{adv}^D + \lambda_2 \mathcal{L}_{cls}^D, \quad (23)$$

$$\mathcal{L}_G = \lambda_3 \mathcal{L}_{adv}^G + \lambda_4 \mathcal{L}_{cls}^G + \lambda_5 \mathcal{L}_{rec} + \lambda_6 \mathcal{L}_p, \quad (24)$$

where $\mathcal{L}_{adv}^D$ and $\mathcal{L}_{cls}^D$ are adversarial loss and classification loss of the discriminator respectively. $\mathcal{L}_{adv}^G$, $\mathcal{L}_{cls}^G$, $\mathcal{L}_{rec}$, and $\mathcal{L}_p$ are adversarial loss, classification loss, reconstruction loss and perceptual loss of the generator respectively. λ_1 and λ_2, are weights for adversarial loss and classification loss of the discriminator, respectively. λ_3, λ_4 , λ_5, and λ_6 are weights for adversarial loss, classification loss, reconstruction loss, and perceptual loss of the generator, respectively.

Adversarial Loss: To generate images that are realistic, an adversarial loss similar to that used in WGAN [18] is adopted for training HF-GAN network. In practice, to distinguish the generated image from the real image, the generator G tries to minimize the adversarial loss, while the discriminator D aims to maximize it during training stage. The adversarial loss of discriminator and generator, denoted by $\mathcal{L}_{adv}^D$ and $\mathcal{L}_{adv}^G$, are defined as:

$$\mathcal{L}_{adv} = E_x[D_{src}(x')] - E_{x'}[D_{src}(G(x'))], \quad (25)$$

where the generator G generates an image from high-frequency feature image x' that extracted from real image x. $D_{src}(x')$ is the probability of the real data estimated by discriminator D. $D_{src}(G(x'))$ is probability of generated image estimated by discriminator D. E_x is the expected value over all real data instances and $E_{x'}$ is the expected value over all generated fake instances $G(x')$.

Classification Loss: As indicated by the name, classification loss is used for evaluating the classification error of the generated image and real data. The discriminator D should have ability to distinguish the age groups of input images. For this purpose, classification losses, denoted as $\mathcal{L}_{cls}^D$ and $\mathcal{L}_{cls}^G$, is added to constrain the discriminator D and generator G in addition to adversarial loss, respectively, in this work. They are defined as equation (26) and equation (27):

$$\mathcal{L}_{cls}^D = -\mathbb{E}_{x,c}[-\log D_{cls}(c|x)] - \mathbb{E}_{x',c}[\log D_{cls}(c|G(x')), \quad (26)$$

$$\mathcal{L}_{cls}^G = \mathbb{E}_{x',c}[-\log D_{cls}(c|G(x'))], \quad (27)$$

where $D_{cls}(c|x)$ represents the probability of the real image x with age group c that computed by D. $D_{cls}(c|G(x')$ means the probability of the generated image $G(x')$ with age group c that computed by D. Discriminator tries to classify real image x and generated image $G(x')$ into their corresponding age groups

by minimizing the equation (26). Generator tries to generate image that belongs to corresponding age group by minimize the equation (27).

Reconstruction Loss: To guarantee that generated image preserves the content of real image, we applied cycle consistency loss [24, 56] to the generator, denoted by $\mathcal{L}_{rec}$. It is defined as the equation (28):

$$\mathcal{L}_{rec} = \mathbb{E}_{x'}[\|x - G(x')\|_1], \tag{28}$$

where L1 norm between the generated face image and the corresponding real image is used to measure reconstruction error.

Perceptual Loss: By minimizing the adversarial loss, generator is trained to generate realistic image. By minimizing the classification loss, generator is trained to generate image that is classified into corresponding age group. Although reconstruction loss can preserve the content of the real image but cannot preserve semantic feature, they fail to guarantee that generated image is perceptually plausible. Johnson *et al.* [56] introduced the perceptual loss function for style transfer and super-resolution. The perceptual loss is defined to measure high-level perceptual and semantic differences between images. Network using perceptual loss is trained based on errors between high-level image features that are extracted from pre-trained network. Same as their work, the proposed method employs perceptual loss for training the HF-GAN model. The VGG19[38] is trained with ImageNet dataset [39]. High-level features are extracted from pre-trained VGG19 model. L1 distance between these features of the real image and generated image is computed. The generator is learned under the guidance of this L1 distance, denoted by $\mathcal{L}_p$. The perceptual loss is defined as equation (29):

$$\mathcal{L}_p = \|V(x) - V(G(x'))\|_1, \tag{29}$$

where V is a particular layer of the VGG19 network, x represents the real image and $G(x')$ represents the generated image. The high-level features are extracted from Conv 5 of the VGG19 network in this work.

5.2 Experiments

5.2.1 Database

The HF-GAN is trained with the Cross-Age Celebrity Dataset (CACD) [57] that contains more than 160,000 images of 2,000 celebrities with age ranging from 16-62. All the images are annotated with age. After face detection, aligning and cropping, 163,104 images whose resolution is 250×250 pixels were obtained. Then these images were randomly divided into two subsets: 90% as training set and the remaining 10% as test set. For training images, there are five age groups: 11-20, 21-30, 31-40, 41-50, and over 50.

5.2.2 Training strategy

First, edge detection is employed for extracting high-frequency features. Then high-frequency images are represented as white lines on a black background with 128×128 pixels. Then high-frequency image and its corresponding real image are inputted into the HF-GAN for training. In the current implementation, the model of HF-GAN is trained on a single GTX 2080TI GPU for approximately more than 30 hours. Learning rate is 1×10^{-4}. The weights λ_1, λ_2, λ_3, λ_4, and λ_5 are set to 1 but the weight of the perceptual loss λ_6 is set to 0.001.

Since high-frequency features are usually high-dimensional, I tried to use HF-GAN model to generate face image from landmarks as well. Hence, HF-GAN is also trained with landmark images. First, we employ the algorithm of Dlib [58] for extracting landmark features. Then landmark images are represented as black solid dots and link these dots using black line on a white background. Both a landmark image and its corresponding real data are inputted into the network of HF-GAN for training. Parameter settings and training details are the same as training HF-GAN with high-frequency images.

5.2.3 Experiments

To evaluate the effectiveness of the proposed method, three types of experiments are conducted. The first experiment directly tested HF-GAN with CACD dataset using high-frequency features. The second experiment tested the proposed model with edited features to further validate the ability of HF-GAN in generating face image with user's desired texture feature. The second experiment is particularly designed by assuming specialized application, such as to create face images with user's desired texture feature instead of age attribute. The third experiment tested HF-GAN with CACD using landmarks instead.

1) Experiment for creating face images from the original high-frequency features

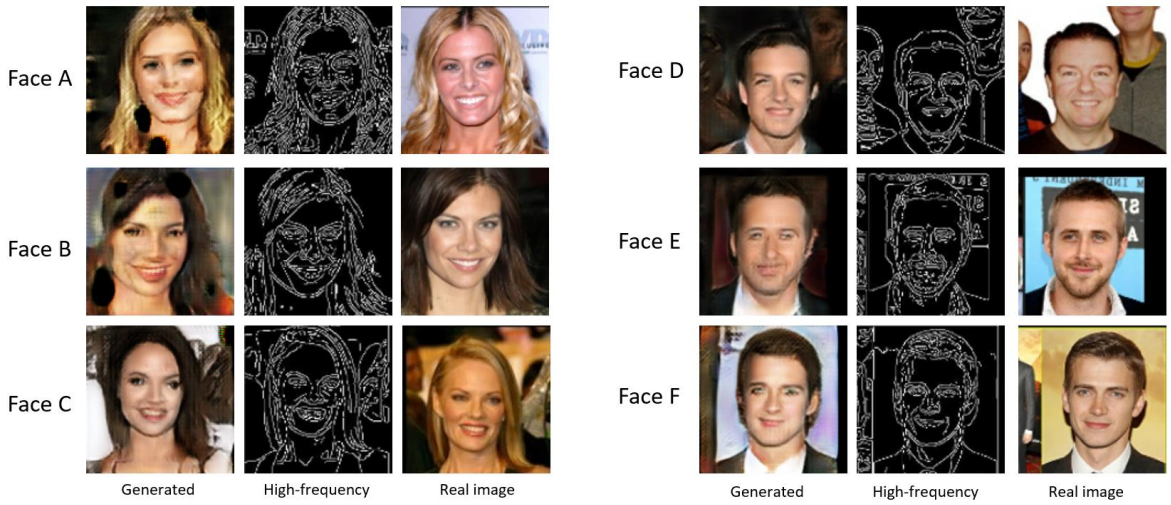


Figure 42. Results with CADA dataset using high-frequency features. For both image groups on the left and right, the first columns are created face images, the second columns are high-frequency images, and the third columns are real images.

Fig. 42 shows the results using proposed method with CADA dataset. We can clearly observe that the generated images can capture the overall texture features of the real image.

2) Experiment for creating face images from the edited high-frequency features

In this experiment, 70 images of persons that under the age of 30 years old were chosen from the dataset. First, high-frequency features were extracted from these images and drawn in white on a 128×128 image with black background. Then, the desired texture information is drawn in the form of few white lines on the original high-frequency feature image using MS Paint. Finally, the face images from these edited high-frequency feature images were generated with HF-GAN.

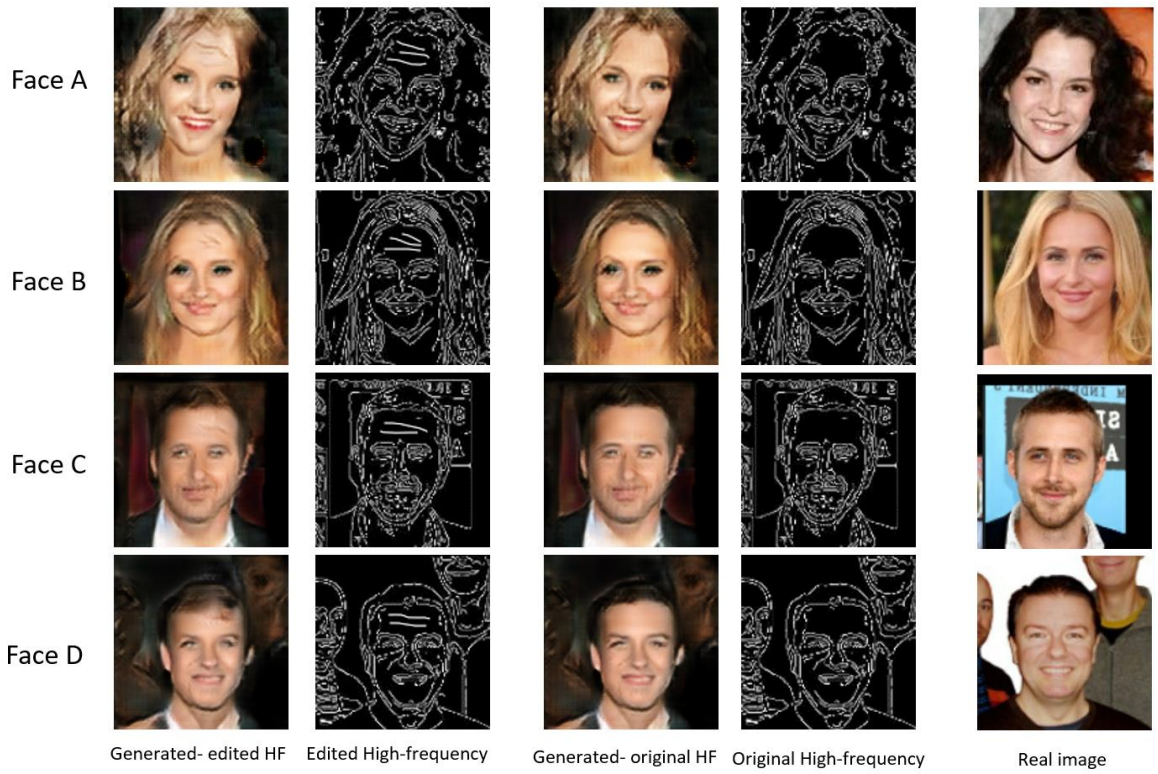


Figure 43. Results using edited high-frequency features. First column: the created face image from the edited high-frequency feature images. Second column: edited high-frequency feature images. Third column: face image generated with original high-frequency feature images. Fourth column: the original high-frequency feature images. Fifth column: the real images

Fig. 43 comparisons the images generated from original high-frequency images and edited high-frequency images. It can be learnt from the results that the inputted high-frequency features have an impact on the generated results. It is possible to synthesize face images with desired texture features by editing the high-frequency features appropriately.

3) Creating face images from the landmarks

This experiment is conducted to validate whether HF-GAN can generate face images from landmarks. First, landmarks are extracted from real images and drawn in white on an black foreground to create landmark images.

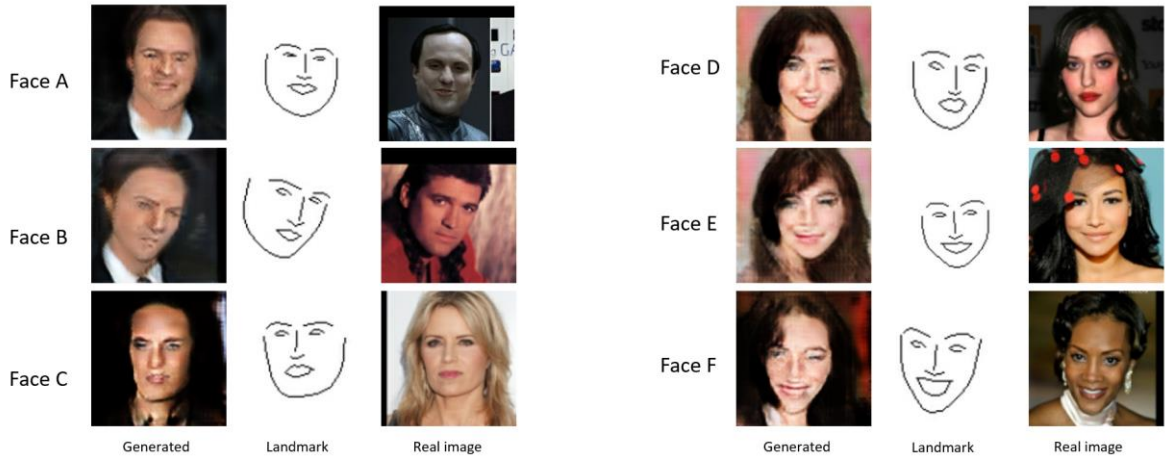


Figure 44. Results generated using landmarks. For both image groups on the left and right, the first columns are created face images, the second columns are landmarks, and the third columns are real images.

Some results are shown in Fig. 44. These results show that the proposed model can also sample face images from the landmarks. The synthesized images can capture overall feature of faces to some extent.

5.3 Discussion and Summary

5.3.1 Discussion

As shown in Fig. 42, Fig. 43, and Fig. 44, most of the generated images can capture the geometric features and texture features of the reference images, but the quality is not satisfying yet. For example, in Fig. 42, the cheek of face B and the eyes of face E are blurry. The results in Fig. 43 show that the face image with desired texture features can be generated from edited high-frequency feature image, but the wrinkles are unnatural. This is probably caused by the width of the drawn lines. It is necessary to develop a method to allow users to easily edit high-frequency features interactively. For generating face image from landmarks, as shown in Fig. 44, the results show that the quality of resulting image from landmarks is not as high as that from high-frequency features. To improve image quality, it is necessary to further carefully tune the training parameters of network and adjust architecture of the network. For feature editing, in the current implementation, MS Paint was used. Some tool for editing the high-frequency feature is required to further improve the image quality.

5.3.2 Summary

This study proposed a novel GAN model HF-GAN to generate face image with desired texture features. Furthermore, the attention mechanism is added to HF-GAN for capture detailed features on face area. The experiment results demonstrated that the proposed model achieves significant improvements in terms of

preserving texture features for generated face images. Additionally, HF-GAN allows users to control the generation result, to obtain the user's desired texture features on the resulting images, to some extent.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

In this thesis, three methods were proposed for semi-automatic generation of users' desired face image. The main contributions of this thesis are summarized as follows:

1. A user-friendly system is developed that can generate not only a face image resembling the target face but also a face image in the user's memory or imagination. Through an iterative approach based on relevance feedback strategy to translate facial features into input, the user only needs to look at several face images and judge whether each image resembles the face that he or she is imagining.

2. Two kinds of face feature representation, which could not only discriminate faces but could also be used to generate a face image, and is also compact enough to allow for the interactive relevance feedback process, have been proposed.

3. OPF classifier is employed for quickly retrieving the best nodes that reflect the user's feedback. New method is proposed for synthesizing the target face image, which is not existing in the training database, by interpolating or extrapolating the best candidate nodes.

5. For the generative model, a traditional algorithm and two deep learning-based methods are proposed. The traditional algorithm is designed based face hallucination method with PCA face representation. The second model is an improved GP-GAN model. By combining GAN with the relevance feedback framework based on the OPF algorithm in the second method, the proposed method succeeded in enable users to interact with GAN so as to reflect their intention in the results of GAN at runtime. HF-GAN is designed for generating face images that not only preserve geometrical features but also texture features.

6.2 Future work

Although the proposed methods of the thesis succeeded in creating face images that resemble the images in users' mind to some extent, there are still some problems to be solved.

1. Although when the user focuses on some particular parts, quickly converging to an image with that part resembling the reference image is possible. However, that particular part may become less similar to the reference image again after trying to improve the other parts. It is necessary for us to develop a technique to help users adjust facial parts independently and achieve the best overall impression.

2. For exploring new candidate landmarks, in the current implementation, the bounding box of all the training images in landmarks feature space is treated as the safe area. However, this is an approximated approach. A more accurate scope should be defined for exploring new candidate landmarks and replace the current approximated approach.

3. For the results produced by the HF-GAN model, there is still much room for improvement in quality. The HF-GAN model can be improved by optimizing architecture, training objectives, and combing some remarkable techniques to model.

4. Expression are also an important factor impacting the perception of face. It is an important future work to take into consideration of expression in generating face images.

Acknowledgements

This work was supported in part by the JAPAN JSPS KAKENHI under Grant 17H00737 and Grant 17H00738, and in part by the Natural Science Foundation of Zhejiang Province, China, under Grant LGF18F020015.

I am very grateful to those who have offered me encouragement and support during the course of my study.

I would like to express my special gratitude to Professor Mao, Supervisor, who patiently instructed and encouraged me during past several years. Professor Mao shared her professional knowledge and vision and offered me valuable suggestions which contributed to many of the papers I wrote. With the help of Professor Mao, I obtain experience in face representation, image processing, face generation, and deep learning frameworks, including Generative Adversarial Network (GAN) techniques.

I also like to express my gratitude to Professor Toyoura who supported my research and papers. I am appreciated for what he did for me. His support is of great help to me. Actually, Professor Toyoura taught me patience and attention to technical details and served as a mentor to encourage my professional growth.

Many thanks to Fushimi, who contributed greatly to the first method of this dissertation, including the implementation of the OPF algorithm; MA Tang who assisted me on the second method of this thesis and Dr. Li and Dr. Xu both of whom were supportive and provided thoughtful and meaningful comments to many of my research methods.

Finally, I wish to thank my family. Especially many thanks to my husband who believed in me, encouraged me and always supported me to pursue my dreams. Whenever I wanted to give up, he gave me the strength to move forward. A special thank you also to my son, who accompanied me on the academic journey and inspired me to achieve my potential.

In the future, I will continue my research and conquer all challenges. I will always remember those who helped me during my most difficult time and I will strive to serve others in the same way.

Reference

1. FACES, IQ Biometrix, [Online] Available: http://www.iqbiometrix.com.
2. Identikit, Identi-Kit Solutions, [Online] Available: https://identikit.net/.
3. E-FIT, [Online] Available: http://www.visionmetric.com/.
4. Frowd, C., Hancock, P., Carson, D.: EvoFIT: A Holistic, Evolutionary Facial Imaging Technique for Creating Composites. ACM Transactions on applied perception, pp. 19-39 (2004).
5. Chiang, C.C., and Chen, Z. Wei., Yang, C. N.: A Component-based Face Synthesizing Method. Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference, pp. 24-30 (2009).
6. FACES: FACES 4.0. [Online] Available: http://www.iqbiometrix.com/products_faces_40.html.
7. Owens, C.: Identikit enters its second decade—Ever growing at home and abroad. Finger Print and Identification Magazine, pp. 3–8, 11–17 (1970).
8. Christie, D., Davies, G., Shepherd, J., Ellis, H.: Evaluating a new computer-based system for face recall. Law and Human Behavior, vol. 5, pp. 209–218 (1981).
9. Christie, D., Ellis. H.D.: Photofit constructions versus verbal descriptions of faces. Journal of Applied Psychology, vol. 66, pp. 358–363 (1981).
10. Wu, D., Dai. Q.H.: Sketch realizing: lifelike portrait synthesis from sketch. Computer Graphics International Conference, pp.13-20 (2009).
11. Wang, X.G., Tang, X.O.: Face photo-sketch synthesis and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, pp. 1955-1967 (2008).
12. Simon, D.: Evolutionary Optimization Algorithms: Biologically Inspired and Population-Based Approaches to Computer Intelligence. Hoboken: Wiley. (2013).
13. Gibson, S. J., Solomon, C. J., Bejarano, A. P.: Synthesis of photographic quality facial composites using evolutionary algorithms. British Machine Vision Conference, British Machine Vision Association, pp. 221-230 (2003).
14. EvoFIT - Evolving Facial Composite Imaging, [Online] Available: evofit.co.uk.
15. George, B., Gibson, S.J., Maylin, M., Solomon, C.J.: EFIT-V- interactive evolutionary strategy for the construction of photo-realistic facial composites. Proceedings of the 10th annual conference on Genetic and evolutionary computation, pp. 1485-1490 (2008).
16. Halim., Abdul, M.F., Fiadh. A.l., Hamdi, H.: Facial composite system using genetic algorithm. International Conference on Computer Graphics, Imaging and Visualisation (2006).
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems, pp. 2672-2680 (2014).

18. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. Proceedings of the 34th International Conference on Machine Learning, pp. 214-223 (2017).
19. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 (2015).
20. Mirza, M., Osindero. S.: Conditional generative adversarial nets. arXiv:1411.1784 (2014).
21. Gauthier, Jon.: Conditional generative adversarial nets for convolutional face generation. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, vol. 5 (2014).
22. Wang, Z. W., Tang, X., Luo, W. X., Gao, S. H.: Face aging with identity-preserved conditional generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7939-7947 (2018).
23. Isola, P., Zhu, J. Y., Zhou, T. H., Efros, A. A.: Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp.1125-1134 (2017).
24. Zhu, J. Y., Park, T., Isola, P., Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, pp. 2223-2232 (2017).
25. Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8789-8797 (2018).
26. Choi, Y., Uh, Y., Yoo, J., Ha, J. W.: Stargan v2: Diverse image synthesis for multiple domains. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.8188-8197 (2020).
27. Di, X., Sindagi. V. A., Patel, V. M.: GP-GAN: Gender preserving GAN for synthesizing faces from landmarks. 24th International Conference on Pattern Recognition, pp. 1079-1084 (2018).
28. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4401-4410 (2019).
29. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2729-2736 (2010).
30. Perarnau, G., Van De Wejier, J., Raducanu, B., Álvarez, J. M.: Invertible conditional gans for image editing. arXiv preprint arXiv:1611.06355 (2016).
31. Antipov, G., Baccouche, M., Dugelay, J. L.: Face aging with conditional generative adversarial networks. IEEE international conference on image processing, pp. 2089-2093 (2017).
32. Bontrager, P., Lin, W., Togelius, J., Risi. S.: Deep interactive evolution. International Conference on Computational Intelligence in Music, Sound, Art and Design, pp. 267-282 (2018).

33. Papa, J., Falcao, A., Suzuki, C.: Supervised Pattern Classification Based on Optimum-Path Forest. Imaging Systems and Technology, vol. 19, pp. 120-131 (2009).
34. Huang, R., Zhang, S., Li. T., He, R.: Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. Proceedings of the IEEE International Conference on Computer, pp. 2439-2448 (2017).
35. Fu, Y., Guo, G., Huang, T. S.: Age synthesis and estimation via faces: A survey. IEEE transactions on pattern analysis and machine intelligence, vol. 32, pp. 1955-1976 (2010).
36. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733 (2016).
37. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018).
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, vol.115 pp. 221-252 (2015).
40. Kim, Yoon.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention, pp. 234-241 (2015).
42. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016).
43. Huang, G., Liu, Z., Van, D. M. L., Weinberger, K. Q.: Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708 (2017).
44. Da Silva, A. T., Falcao, A. X., Magalhaes, L.P.: A new CBIR approach based on relevance feedback and optimum-path forest classification. Journal of WSCG, vol. 18, pp. 73-80 (2010).
45. Papa, J., Falca, A.: Optimum-Path Forest: A Novel and Powerful Framework for Supervised Graph-Based Pattern Recognition Techniques. Institute of Computing University of Campinas, pp. 41-48 (2010).
46. Liu, C. Shum, H. Freeman, W.: Face Hallucination: Theory and Practice. International Journal of Computer Vision, vol. 75, pp. 115-134, (2007).
47. Ruthven, I., Lalmas, M.: A Survey on The Use of Relevance Feedback for Information Access Systems. The Knowledge Engineering Review, vol. 18, no. 2, pp. 95–145 (2003).
48. Li, H., Toyoura, M., Shimizu, K., Yang, W., Mao, X.: Retrieval of Clothing Images Based on Relevance Feedback with Focus on Collar Designs. Visual Computer, vol. 32, no. 10, pp. 1351-1363 (2016).

49. Silva, D., Tavares, A., Falcão, A. X., Magalhães, L. P.: Active learning paradigms for CBIR systems based on optimum-path forest classification. Pattern Recognition, vol. 44, pp. 2971-2978 (2011).
50. Li, H., Liu, G., Ngan, K.: Guided Face Cartoon Synthesis. IEEE Transactions on Multimedia, vol. 13, no. 6, pp. 1230-1239 (2001).
51. Gao, W., Gao, B., Shan, S., Chen. X., Zhou, D., Zhang, X., Zhao, D.: The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 38, no. 1, pp. 149-161 (2008).
52. Huang, G., Mattar, M., Lee, H., Learned-Miller, E. G.: Learning to align from scratch. Advances in neural information processing systems, pp. 764-772 (2012)
53. Huang, G., Mattar, M., Lee, H., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
54. Calder, A. J., Burton, A. M., Miller, P., Young, A. W., Akamatsu, S.: A principal component analysis of facial expressions. vol. 41, pp. 1179-1208 (2001)
55. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. Proceedings of the 34th International Conference on Machine Learning, vol. 70 (2017).
56. Johnson, J., Alahi, A., Fei-Fei. L.: Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision, pp. 694-711 (2016).
57. Chen, B. C., Chen, C. S., Hsu, W. H.: Cross-age reference coding for age-invariant face recognition and retrieval. In Proceedings of the European Conference on Computer Vision (2014).
58. King, D. E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, vol.10, pp.1755-1758 (2009).