Biometric Information Analyses Using Computer Vision Techniques

Bingzhang Hu

School of Computing
University of East Anglia

This dissertation is submitted for the degree of
Doctor of Philosophy

March 2019

I would like to dedicate this thesis to my grand father.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. Parts of this thesis have been taken from published/under review academic conferences. All of these papers were primarily written by me, Bingzhang Hu, during and as a result of my Ph.D. study. Involved papers are listed as follows.

- **Bingzhang Hu**, Feng Zheng and Ling Shao. Dual-reference Face Retrieval, published in AAAI 2018 (Chapter 5).
- **Bingzhang Hu**, Yuan Zhou and Ling Shao. What Will You Look Like at Age 'x'?, submitted to IJCAI 2019 (Chapter 6).
- **Bingzhang Hu**, Yan Gao, Yu Guan, Yang Long, Nicholas Lane and Thomas Ploetz. Robust Cross-View Gait Identification with Evidence: A Discriminant Gait GAN (DiGGAN) Approach on 10000 People, submitted to IJCAI 2019 (Chapter 7).

Bingzhang Hu
March 2019

Acknowledgements

I would like to extend my sincerest thanks to the following, who have all helped in the completion of this thesis.

First, I would like to express my sincerest thanks to my supervisor, Prof. Ling Shao, who has led me into the computer vision field, shown me inspiring directions and left me enough spaces to explore the mysteries of this field.

To my family, especially my parents, who have provided every possible material and spiritual support for my study, I would like to express my thanks. I love you.

I would also like to thank the help and inspirations from my dear colleagues. Feng Zheng, Yang Long, Daniel Organisciak have provided many helpful supports in both my research and living. Your encouragements in my dark time were and will always be the most precious memories in my life. I will extend my sincere thanks to Li Liu, Lining Zhang, Mengyang Yu, Yi Zhou, Yuming Shen, Jiaojiao Zhao, Shidong Wang, Yang Liu, Jing Li, Haofeng Zhang, Xiaoming Liu, Shengbing Liao, Heng Liu, Yuan Zhou and all of whom that may not be fully listed here.

Finally, I would like to thank Handan Zhang. You make me complete.

Abstract

Biometric information analysis is derived from the analysis of a series of physical and biological characteristics of a person. It is widely regarded as the most fundamental task in the realms of computer vision and machine learning. With the overwhelming power of computer vision techniques, biometric information analysis have received increasing attention in the past decades. Biometric information can be analyzed from many sources including iris, retina, voice, fingerprint, facial image or even the way one walks with. Facial image and gait, because of their easy availability, are two preferable sources of biometric information analysis.

In this thesis, we investigated the development of most recent computer vision techniques and proposed various state-of-the-art models to solve the four principle problems in biometric information analysis including the age estimation, age progression, face retrieval and gait recognition.

For age estimation, the modeling has always been a challenge. Existing works model the age estimation problem as either a classification or a regression problem. However, these two types of models are not able to reveal the intrinsic nature of human age. To this end, we proposed a novel hierarchical framework and a ordinal metric learning based method. In the hierarchical framework, a random forest based clustering method is introduced to find an optimal age grouping protocol. In the ordinal metric learning approach, the age estimation is solved by learning an subspace where the ordinal structure of the data is preserved. Both of them have achieved state-of-the-art performance.

For face retrieval, specifically under a cross-age setting, we first proposed a novel task, that is given two images, finding the target image which is supposed to have the same identity with the first input and the same age with the second input. To tackle this task, we proposed a joint manifold learning method that can disentangle the identity with the age information. Accompanied with two independent similarity measurements, the retrieval can be easily performed.

For aging progression, we also proposed a novel task that has never been considered. We devoted to fuse the identity of one image with the age of another image. By proposing a

novel framework based on generative adversarial networks, our model is able to generate close-to-realistic images.

Lastly, although gait recognition is an ideal long-distance biometric information task that makes up the shortfall of facial image, existing works are not able to handle large scale data with various view angles. We proposed a generative model to solve this term and achieved promising results. Moreover, our model is able to generate evidences for forensic usage.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The word "computer" was first recorded as being used in 1613 to refer to a person who performed calculations or computations [1]. In the end of the 19th century, "computer" started to be used to refer to a machine that performed calculations. The concept of modern computer was invented by Alan Turing in 1936 to describe a machine that manipulate symbols on a strip of tape according to a table of rules. Since then the architecture of modern computer has experienced many evolutions. In 1945, Von Neumann *et.al.* proposed a computer architecture which consists of a processing unit that contains an arithmetic logic unit and processor registers, a control unit that contains an instruction register and program counter, memory that stores data and instructions, external mass storage and input/output mechanisms. This architecture is still used today in most computers. Entering the 21st century, with the development across disciplines of semiconductor physics, boolean algebra, integrated circuit, artificial intelligence *etc.*, computer enters our daily life and plays an extremely important role in many areas including education, business, health care, entertainments *etc.*

Vision, the faculty or state of being able to see, is a natural ability of our human beings. For computer, the vision is the ability of automatic extraction, analysis and understanding high-level information from digital images or videos . One can imagine that there will be many potential applications when computers have, if not better than, competitive vision ability with human beings. Specifically, computer vision techniques have been used for Recognition, Detection, Tracking, Classification *etc.* For instances, computer vision have helped in X-Ray images reading [138], cancer detection, pedestrian detection, *etc.*

Among the broad research area and industrial applications involving computer vision techniques, the **biometric information analyses**, including face recognition and verification, age estimation, aging progression, gait recognition *etc.* have received much attention in recent decades as they are human-centric and closely related to our daily life. For example, face recognition technique has been widely used in many access control systems in various

zones, including train station, airports, *etc.* Its variation, facial image retrieval, can be used for missing person search and rescue. Biometric information, as a technical term for body measurements and calculations, includes age, gait, voice, finger print, retina, identity *etc*. It can come from different modalities of data such as images, videos, audios and raw data from different sensors. Images and videos, thanks to the development of mobile devices, are very easy to be obtained in daily life. There are thousands and millions of images and videos uploaded to the Internet every day [2]. Considering the broad application scenarios and the overwhelming emerging of the data, analyzing the biometric information within these prolific data source is very necessary.

1.1 Motivations

In this thesis, we explore the development of the current computer vision techniques on biometric information analysis, and propose different approaches to improve the performance on many tasks including age estimation, face retrieval, facial image generation and gait recognition. We address the problems of existing works as following:

a) Although many efforts have been spend on the age estimation problem, the machine's performance is still far away from human. Existing works [44, 90] show that a hierarchical framework is effective in solving age estimation problem, thereby the age can be predicted in a coarse-to-fine favor. A typical hierarchical age estimation framework first assigns age into different groups, however in existing works, such age groups are usually defined heuristically. Is there anyway to define or find an optimal age grouping protocol?

b) Driven from the above problem, we revisit a question that has been discussed in the community over the decades: what is an optimal way to formulate age estimation problem, Multi-classification or regression? Existing methods formulate age estimation as either one of them or a hybrid. In multi-classification frameworks, the labels are assumed independent with each other, which is contrary to the fact that human age is an ordinal set. The regression approaches partially preserve the natural ordinal structure of the human aging progress by treating each label as a numerical value. Nevertheless, the regression frameworks typically learn a linear kernel while the human aging is non-linear in the feature space. Are the multi-classification or regression framework the only ways to solve the age estimation problem?

c) If we can figure out which information is important to age estimation, can we retrieve face images across age? Existing works have obtained promising results in face

retrieval. However, they cannot hold the same performance under a cross-age setting. It becomes extremely difficult when one wants to retrieve the images across ages, or even retrieve the images at a specific age.

d) As an inverse problem of age estimation, human aging progression has always been very challenging. Existing aging progression methods can only generate facial images within a coarse age range rather than a specific age. Is it feasible to generate images that mimic an age in a reference image?

e) Biometric information analyses on facial images requires high quality images with fair resolutions. Gait recognition, as a long-distance identification method, is a good aid. However, existing works on gait recognition can not generate well on large scale dataset. Moreover, existing works suffer from the changes of the views between the probe and gallery.

1.2 Contributions and Thesis Outline

The contributions of this thesis include handling the problems addressed above and proposing solutions and models that achieve state-of-the-art performances.

The rest of the thesis are structured as below:

Chapter 2: Literature Review. An overview of the state-of-the-art methods including the age estimation, face retrieval, aging progression and gait recognition.

Chapter 3: A Coarse to Fine Age Estimation Approach. In this chapter, we focus on problem *a*. To find the optimal age grouping protocol in coarse to fine age estimation framework. We propose a clustering based grouping method for age estimation. First, we train a neural network to extract features from the images by optimizing a classification task. We then introduce a random forest based clustering method to find clusters in the learned feature space and each cluster is defined as an age group. Subsequently, for each group, a specific convolutional neural network based classifier is trained for the final fine age estimation. Systematic experiments are conducted on MORPH, FGNet and LaP dataset and the experimental shows the proposed method outperforms state-of-the-art. Also, an insight experiment shows, compared with existing heuristic age group protocols, our grouping protocol reveals the intrinsic structure of age space.

Chapter 4: Metric Learning for Age Estimation. Though a coarse to fine age estimation framework can yield promising results, the conclusion has never been drawn on what is the optimal formulation for age estimation. In this chapter, we address the problem *b*. We give up the obsolete multi-classification or regression framework, and propose an ordinal

metric learning approach for age estimation. In the proposed method, an ordinal age structure preserving metric is learned to exploit the prolific ordinal information among the age labels. We evaluate our proposed method on MORPH, FGNet and LaP dataset. The experimental results show that our metric learning based age estimation outperforms state-of-the-art.

Chapter 5: Dual Reference Face Retrieval. In this chapter, problem *c* is studied. We first propose a novel task, that is retrieving a person's face image at a specific age, especially when the specific 'age' is not given as a numeral as it used to be, i.e. 'retrieving someone's image at the similar age period shown by another person's image'. To tackle this problem, we propose a dual reference face retrieval framework, where the system takes two inputs: an identity reference image which indicates the target identity and an age reference image which reflects the target age. In our framework, the raw images are first projected on a joint manifold, which preserves both the age and identity locality. Then two similarity metrics of age and identity are exploited and optimized by utilizing our proposed quartet-based model. The experiments show promising results, outperforming hierarchical methods.

Chapter 6: Aging Progression. Age progression has received much attention in recent years. Here we revisit the aging progression problem and ask: is a numeral capable enough to describe the human perception of the age? To solve the problem *d*, we propose a new framework Dual-reference Age Synthesis (DRAS) that takes two images as inputs to generate an image which shares the same personality of the first image and has the similar age with the second image. In the proposed framework, we employ a joint manifold feature which consists of disentangled age and personality information. The final images are generated by training a generative adversarial network which competes against an age agent and an identity agent. Experimental results demonstrate the appealing performance and flexibility of the proposed framework by comparing with the state-of-the-art and ground truth.

Chapter 7: Gait Recognition Based on Generative Adversarial Networks In this chapter, problem *e* is considered. Gait is an important biometric trait for surveillance and forensic applications, which can be used to identify individuals at a large distance through CCTV cameras. However, it is very difficult to develop robust automated gait recognition systems, since gait may be affected by many covariate factors such as clothing, walking surface, walking speed, camera view angle, etc. Out of them, large view angle was deemed as the most challenging factor since it may alter the overall gait appearance substantially. Recently, some deep learning approaches (such as CNNs) have been employed to extract view-invariant features, and achieved encouraging results on small datasets. However, they do not scale well to large dataset, and the performance decreases significantly w.r.t. number of subjects, which is impractical to large-scale surveillance applications. To address this issue, in this work we propose a Discriminant Gait Generative Adversarial Network

(DiGGAN) framework, which not only can learn view-invariant gait features for cross-view gait recognition tasks, but also can be used to reconstruct the gait templates in all views — serving as important evidences for forensic applications. We evaluated our DiGGAN framework on the world's largest multi-view OU-MVLP dataset (which includes more than 10000 subjects), and our method outperforms state-of-the-art algorithms significantly on various cross-view gait identification scenarios (e.g., cooperative/uncooperative mode). Our DiGGAN framework also has the best results on the popular CASIA-B dataset, and it shows great generalization capability across different datasets.

Chapter 8: Conclusion and Future Work In this chapter, a brief summary of the contributions of this thesis is given as well as an outlook of the future work.

Chapter 2

Literature Review

To the best of my knowledge, most of the researches in computer vision or, even in a higher level, machine learning area, are devoted to improving three essential problems: the representation learning; problem modeling; and optimization. The contributions of existing works either lie on learning an optimal representation/feature that contains enough and useful information for solving the task; proposing different models and various handily objective functions; seeking a tricky and efficient way to solve the objective functions; or a combination of them. According to this fact, and to make the structure clear to read, we review the literature along these essential problems. As the thesis involves two major branches of biometric information: face and gait, we review the literature of them respectively. Additionally, many works have employed deep neural networks as the backbone of their proposed framework. To give the readers essential background of this technique, we give an overview on deep neural networks at the end of this chapter.

2.1 Biometric Information Analysis on Facial Image

Facial images contain much important information including identity, age, gender, expressions *etc.* Although there are various tasks in biometric information analysis on facial image, such as age estimation, face recognition and so on, a common challenge of them is to learn a representation that the computer can easily understand instead of letting the computer work directly on the raw image. In this section, we review the face verification and recognition, age estimation and aging progression respectively.

2.1.1 Face Verification and Recognition

Hand-crafted representations for face verification and recognition

Face verification and recognition are two similar tasks that are both concerning the subject's identity. Face verification, also called authentication, is to validate if the identities are the same between two subjects; while face recognition is to validate the subject's identity with many registered subjects. Feature learning in biometric information analysis on facial image can be broadly divided into two branches. One is the hand-crafted feature, where the feature is designed based on the specific domain knowledge or statistical information. The earliest feature extraction for facial verification dates back to 1964 [15], where they measure the distances between landmarks, *e.g.* nose, eyes, as the feature to verify whether two images belong to a same identity. In 1991, [135] proposed a face recognition system that projects face images onto a feature space by principal components analysis (PCA), where the significant variations of facial images, also known as eigenfaces are spanned. After the projection, each individual's facial image is then represented as a weighted summation of these eigenfaces. Their proposed system can achieve near-real-time face recognition and apply on new faces. Later on in 1997, [12] improved the performance by replacing the principal components analysis with fisher's linear discriminant projection on original images. Compare with eigenfaces based method, Fisher face based method are more robust to the lighting changes. To better describe a human face and more reliably detect features against the variations in image intensity and feature shape, Active Shape Model (ASM) [29] is proposed in 1995. ASM constrain the shapes by the Point Distribution Model (PDM), because the human face images is a structured data, for example, the position of eyes are statistically higher than that of nose. The PDM constraint allows the face shape change only within a range learned from a training set. ASMs are not only used in face recognition, but also work well in facial landmarks detection. Based on the ASM, a flexible appearance model (FAM) [78] as well as active appearance model (AAM) [27] are proposed to capture the texture and appearance information on facial images. Combined with the shape information extracted by ASM and the appearance information learned by FAM and AAM, the face recognition have achieved a higher performance. More recently, various heuristic features are widely used in face verification. Local Binary Patterns (LBP) [4] first compare the pixel value between the central pixel with its surroundings to encode the results as an 8-digit binary number. Then the histogram of these 8-digit binary numbers' values are calculated as the representation. LBP is successful in capturing texture information thus yield good performance on face verification. Histogram of Oriented Gradient (HOG) [30], describe the images with the distribution of intensity gradients and edge directions. Other popular

features for object classification such as SIFT [93] and SURF [11] are also employed in face verification area and yield soundable results.

Data-driven representations for face verification and recognition

Another branch is the learnt feature, also known as data-driven features. With the emergence of neural network [14], the feature learning can be solved simultaneously with the objective optimisation. In other words, the neural network accepts images as the input at one end and outputs the results at the other end, for example, the identity in the input. The neural network is generally an end-to-end system thus the value of the loss function can back propagate through each layer and the weights that are used for extracting features can also be updated. In this kind of end-to-end structure, to extract which information as the feature is automatically determined by the task. Compared with the hand-crafted features, the learned feature are task and data specific, which is generally more helpful to achieve a considerable performance. In the past decades, the computation capability meets its huge development due to the evolution in semi-conductor as well as other relative area. which makes it feasible to design and train a deep neural network. The first work employing deep neural network was published in 2015 [106], Parkhi et al. trained a neural network with the same architecture as AlexNet [70], which is previously used in ImageNet large-scale visual recognition competition. The performance of [106] on Labeled Faces in the Wild [60] LFW has achieved 97%, which is very close to human performance. Followed by faceNet [116] in 2015, Microsoft trained their face verification network based on vgg-16 and further improved the machine performance towards 99%. Recently, many other networks are proposed for face verification, such as [106, 132]. Although deep learning techniques have significantly improved the machine performance on face verification, the drawbacks of them are also non-ignorable. First, training such deep model generally requires large-scale external data, which is not aways feasible. Secondly, training such deep model is computational expensive, it is very unlikely to implement a real-time system on mobile devices, also there are some works introduced model compressing methods, *e.g.* Mobile Net [57]. Thirdly, reaching nearly 100% accuracy rate on LFW does not mean the model can generalise well on realistic data, especially when there are large illumination, pose, expression variations in realistic data. Lastly but the most importantly, the LFW dataset does not consider the age change between training and test. Whilst in the real world, people will get old, which will result in significant appearance change, thus the face verification systems are supposed to be robust to the age change.

To tackle the drawbacks and challenges discussed above, existing works have paid many efforts in problem modelling for face verification. There are several bunches of

face verifications works focusing on different challenges such as expression-invariant face verification, occlusion-robust face verification, face verification for make-ups, illumination-invariant face verification, *etc.* To achieve various objectives, different models are designed. For example, [151] proposed a discriminative marginal metric learning method for makeup face verification, where they use the interclass marginal faces to depict the discriminative information. [92] proposed a spatial and statistical pooling method for face verification with small occlussions. More generally, to learn a unified embedding that robust to the variations including illumination, pose, *etc.*, FaceNet [116] proposed a deep neural network that utilizes a triplet loss to enlarge the interclass distances while reducing the intraclass distances.

It is interesting to note that, although many works have been done on coping with the above mentioned variations in human facial image. There are very few works paying attention to the age change, which happens every minutes and seconds. One possible reason is that age is a very difficult biometric information to predict, even for human themselves. So the literature of age estimation is reviewed in the next section.

2.1.2 Age Estimation

Age is an extremely important information that can be drawn from the facial image. In many social scenarios, during face-to-face communications and interactions, people will behave and react differently according to the ages of their audiences. For example, one may use different language styles to a teenager and a senior. However, compared with face verification, the development of age estimation are far away from being used in industrial-level situations. There can be many reasons, the most important of which is the human aging process is determined by both the intrinsic factor, *i.e.* gene, and the extrinsic factors, *e.g.* life style, mental status, *etc.* So exploring the variances in these factors are extremely challenging.

Feature learning for age estimation

As discussed above, to learn a proper representation that embeds both the intrinsic and extrinsic factors from the facial image for age estimation is an important step. Although the features employed in face verification can also work in age estimation, such as LBP, HOG, AAM, ASM, *etc.* One may find an obvious fact that in face verification, the optimal feature is supposed to be age-invariant which is absolutely opposite in age estimation problem. To this end, many heuristic as well as data-driven features for age estimation are proposed in past few decades. In 1994, [75] proposed an age estimation framework, in which they calculate the distances between different facial landmarks and then use the ratios as feature to classify the subjects into infants, young adults and senior adults. However, this early work faced a

problem in distinguishing the young and old adults because the growth of human facial bones are relatively slow, thus the ratios may remain the same in their protocol. Researchers found the texture information, such as wrinkles on the face, is helpful in age estimation. Many works [4, 27] employed the LBP, AAM, Gabor features to encode the texture information and yielded considerable progress on this task. [47] proposed a method that utilizes biologically inspired features (BIF). The architecture to extract BIF is actually very similar to a two or four-layer neural network. In their work, they use 64 Gabor filters to extract different information along four directions in the first layer, namely $S1$. Followed by layer $C1$, the maximum response within each local spatial neighborhood in $S1$ is selected and flattened into a vector. They then employed the PCA to reduce the dimension of the final representation. On the other side, deep learning based works [84, 147] also achieved remarkable improvements. However, there are only small modifications on the network architecture. Additionally, the improvements on age estimation in [84] and [147] are somehow introduced by the multi-task settings, in which they predict the age as well as the gender or ethnicity. Besides these feature based frameworks, it is also worth noting that, a number of existing works are template based. [34] proposed an automatic age estimation method based on facial aging patterns. They defined an aging pattern as a sequence of personal face images sorted in order and construct the aging pattern subspace from the training images. During testing, they traverse all possible positions for the test image in the aging pattern subspace and find the one minimizes the reconstruction error. The position can be finally converted to the predicted age.

Problem modelling in age estimation

For the problem modelling, most of existing works either solve the age estimation as a classification problem or a regression problem. In the classification faction, each age is treated as an independent class [75]. Some works first divided the ages into small groups, for example, [84] classified the ages into groups of $(0-2,\ 4-6,\ 8-13,\ 15-20,\ 25-32,\ 38-43,\ 48-53,\ 60-)$. There are many drawbacks in the classification setting. Neither the individual class nor the grouped class setting considers the ordinal information, which is the natural of the human aging progress, between each class. In other words, when one classifies an 8 years old child as 10 or 20 years old by mistake, the classification framework measures these two mistakes as the same while in fact they are not. Another non-ignoble problem is, there can be many various protocols to group the ages, for example, fixed interval and flexible interval. However, there is no 'golden rule' to decide which protocol is optimal. Regression based methods try to regress images directly to a numeral. In regression settings, the ordinal information can be well preserved, and the model will receive a heavier penalty if the predicted result has a larger difference with the ground truth. Nevertheless, the regression

models are not able to approach the non-linear characteristics of aging progress. Moreover, the typical objectives in regression models cannot measure the variations happening in different aging periods. To address the issues in classification and regression frameworks, [35] proposed a label distribution learning method. Instead of simply outputting a numeral, they predict the age as a distribution. [21] formulates the age estimation as a ranking problem, where they learn several hyperplanes that can separate facial images as their ordinal information. Similar ideas are taken in [24], they propose a ranking convolutional neural network which consists of several basic CNNs. Each CNN predicts whether the age of the input image is higher than a fixed value.

In contrast with age estimation problem, an inverse task is aging progression. Aging progression shares many common challenges with age estimation. In the next section, we review the literature of aging synthesis.

2.1.3 Aging Progression

Aging Progression, as the mirror task of aging estimation, has received much attention in the literature because of its many potential applications including forensic art, entertainment, *etc*. For instances, looking for missing person for law enforcement authorities; face aging morph in social media applications.

The existing aging progression methods can be split into three folds, the model-based, the protocol-based and the generative adversarial networks based. As far back as 2002, Lanitis *et al.* described how the effects of aging on facial appearance can be explained using learned age transformations [79]. Then traditional age synthesis focus on facial muscle structure or skin's texture changes*etc*, or learn those features from the average faces of different age groups for age pattern transfer. Those models are usually very complex or neglects the differences between different persons [152]. Kaur *et al.* [68] proposed face texture transfer (FaceTex) framework augmented the prior work. FaceTex suppress facial texture comprising skin texture details around facial meso-structures (*e.g.* eyes, nose and mouth) and synthesize a facial image with different facial textures while maintaining the identity of the original one. Further, Makhzani *et al.* [94] proposed an Adversarial Autoencoders (AAE) using an adversarial training procedure to learn the latent vector, which inspired most of the state-of-the-art face aging methods. [5, 76, 91] investigate age synthesis based on GAN and AAE which can generate the personalized aging images at the tender age. The state-of-the-art GANs combined GAN with AAE, which can learn the intangible character through an encoder and generate images with photo realistic. Zhang *et al.* [152] proposed a conditional adversarial auto-encoder (CAAE) and described a synthesis protype based on GAN and AAE: personalized identities are indicated by map the original face image

to a latent vector via an encoder, then these identities and a corresponding numeral (age) are fed into the generator to synthesize face images. Antipov *et al.* [5] proposed an Age Conditional Generative Adversarial Network (Age-GAN) to generate identity-preserving synthetic images within required age categories, which use the Facenet to optimize latent vectors, and it can be considered as a part of CAAE. Recently, Wang *et al.* [139] proposed an identity-preserved conditional generative adversarial networks (IPCGANs) , which use an age classifier forces the generated face with the target age and use the multi-layer feature of age classifier as identity feature. Recently, Li *et al.* [87] proposed a Wavelet-domain Global and Local Consistent Age Generative Adversarial Network (WaveletGLCA-GAN) which adopt wavelet transform to depict the textual information in frequency-domain with given age labels, WaveletGLCA-GAN abstract age information from local patches of a given age face image and generate an image with the target age, but it needs forehead, eyes and mouth local patches of the target age images and consists five sub-networks which are complex. Despite of focusing on face aging synthesis, Expression Generative Adversarial Network (ExprGAN) [32] and StarGAN [26] can also be used for face aging synthesis. A style-based generator architecture [67] for generative adversarial networks, which leads to an automatically learned, unsupervised separation of high-level attributes (*e.g.* pose and identity when trained on human faces) and stochastic variation in the generated images (*e.g.* freckles, hair), and it enables intuitive, scale-specific control of the synthesis.

2.2 Biometric Information Analysis on Gait

Facial image analyses typically require high quality images, which usually need the subject to stand in front of a camera. Gait recognition, as a long-distance identification technique, is a very good aid for biometric information analyses on facial images. In this section, we review the development in gait recognition, especially cross-view gait recognition area.

Cross-view gait recognition methods can be roughly divided into three categories. The first category, for example, [6] is based on reconstructing 3D gait model from multiple calibrate cameras. These branch of methods have very obvious drawbacks — they rely on multiple cameras which are fully controlled and working cooperatively. Such requirements are very challenging to satisfied in real-world applications.

The second category typically achieve cross-view gait recognition by performing view normalization. For example, [37] first estimates the poses of lower limbs and then extracts the rectified angular measurements as well as trunk spatial displacements as features for gait recognition. However, such method is not always feasible especially when the lower limbs are not clearly visible hence the poses are difficult to be estimated. To tackle this problem,

[74] proposed a view normalization framework based on domain Transformation obtained through Invariant Low-rank Textures (TILT), where the gait images are normalised to the side view without knowing the prior pose of the gait. Nevertheless, the performances of such method is not promising when the gait images are captured in front and back view as there is a large view angle gap with the side view.

The third category is to learn a common space where the gait images from different view angles are mapped into a same feature space and then a metric is learnt to measure the similarities then perform the matching. For instance, [95] introduced the SVD-based View Transformation Model (VTM) to project gait features from one view into another. This method is improved by Kusakunniran et al. by using Truncated SVD (TSVD) [71] to avoid oversizing and overfitting of VTM. Instead of using the global features (e.g.,[95], local Region of Interest (ROI) was selected based on local motion relationship to build VTMs through Support Vector Regression (SVR).

There are also some variations in the third category. For example, Bashir et al. [10] used Canonical Correlation Analysis (CCA) to project gaits from two different views into two subspaces with maximal correlation. The correlation strength was employed as the similarity measure for identification. In [73], after claiming there may exist some weakly or non-correlated information on the global gaits across views [10], motion co-clustering was carried out to partition the global gaits into multiple groups of gait segments.

Most recently, deep learning approaches [120, 142, 148], [53] were applied for gait recognition, which can model the non-linear relationship between different views. In [120], the basic CNN framework, namely GEINet was applied on a large gait dataset, and the experimental results suggested its effectiveness when the view angle changes between probe and gallery are small. To combat large view changes, a number of CNN structures were studied in [142] on the CASIA-B dataset (with 11 views from 0 to 180), and Siamese-like structures were found to yield the highest accuracies. However, this dataset only includes 124 subjects, and the most recent work [128] found these CNN structures do not generalise well to a large number of subjects. In [148] and [53], GAN approaches are applied to generate gait features/images to a common view or a target view for matching. However, the generative nature of both GAN models limit the recognition accuracies, although they are more interpretable than the discriminant CNN-based approaches [53].

2.3 Neural Networks and Deep Learning

The concept of neural network is first proposed in 1943 in [98], where the terminology ‘connectionism’ was proposed and a connected circuits was used to simulate intelligent

behaviour. Later on in 1960, the neural network was first time applied to a real world problem at Stanford to approach an adaptive pattern classification machine [140]. The modern neural network architecture was proposed at 1998 by Yann LeCun in [81], where the multilayer neural network LeNet were employed to recognize the documents. On the other side, Hochreiter and Schmidhuber proposed the long short-term memory (LSTM) to deal with the gradient descent problem in analysing time-series data. Entering the 21^{st} century, the deep neural networks have received their flourishing development, among which the convolutional neural network [82] has yield very promising results in many computer vision tasks such as image recognition and object detection. Recently in 2014, Ian Goodfellow *et al.* proposed a generative adversarial network (GAN) [38], where two neural networks contest with each other in game theory.

Chapter 3

A Coarse to Fine Age Estimation

3.1 Introduction

The human face is a prolific information source during face-to-face communication. Many kinds of useful information, such as age, gender, race, *etc*., can be obtained from a single facial image. Among them, age is extremely important in many social occasions, because people may imperceptibly react or behave differently when facing others at various ages. For instance, when being asked road, one may prefer using more concise descriptions to young people but more concrete words to the seniors. Thereby automatic age estimation (AAE) from facial images, as an interesting task, has aroused increasing attentions in the literature over the past decades. There are many potential applications based on AAE such as age-specific human-computer interaction, commercial user management, demographics analysis, business intelligence, *etc.*

Although automatic age estimation has great potential values across various areas, it has always been a very challenging task. Compared with the machine's close-to-human performance on face verification task, the development of automatic age estimation is far away from being called good. There are mainly three reasons. First, different individual's aging progression varies a lot. In other words, aging progression is uncontrollable because it is not only affected by the intrinsic factor, *e.g.* gene, but also by the extrinsic factors, *e.g.* living environment, lifestyle, mental status and so on. It is very hard to build a model that can perfectly encode these variations. Moreover, the aging progression various in different age periods. Concretely, the aging of the children can be mostly illustrated by the changes of face shape; while the adults' aging can change the texture of the skin [126]; Another case is that the lower and upper halves of faces grow at different rates during formative years [134]. Secondly, the variations including illumination, pose and expression in the facial image make the age estimation extremely difficult. For example, when the subject in the

image is smiling, it is very likely that the eye pattern appears which may be considered as a wrinkle by machine. Lastly, it is difficult to choose a proper formulation to mimic the way of our human beings estimating others' ages. Existing works on age estimation can be broadly divided into two branches, the classification based and regression based. However, classification based methods ignore the ordinal information between each age class and treat them as independent classes. Although regression can take the ordinal information into account, [44] claimed that the regression is only able to give a rough range of the age as human aging progress is non-linear. Therefore they proposed a coarse-to-fine age estimation protocol, where the coarse age denotes a rough age range while the fine age indicates a specific numeric value. In their proposed method, the coarse age range is first obtained by a regressor, *i.e.* support vector regressor (SVR), then a fine age label will be locally adjusted by sliding the predicted value from the previous stage up and down to minimize the absolute error with the ground truth. [90] used a similar idea except for the coarse prediction stage is done in a classification favor. They first divided the age labels into small groups. However, a non-ignoble problem here is there is no principle to follow in grouping the ages. [84] groups the ages into $(0-2,\ 4-6,\ 8-13,\ 15-20,\ 25-32,\ 38-43,\ 48-53,\ 60-)$. [42] proposed a six groups protocol as: $10\pm5,\ 20\pm5,\ 30\pm5,\ 40\pm5,\ 50\pm5,\ 60\pm5$ years old. There is no convincing reason to say the 55 years old is supposed to belong 40 ± 5 group rather than 50 ± 5 years old.

To address the problem that both classification and regression have their drawbacks in formulating the age estimation task, in this chapter, we proposed a hybrid age estimation method with the combination of clustering and classification, which is shown in Fig. 3.1. We first train a deep neural network based on the strong labels, which stand for the annotated age labels obtained from the dataset. Then we use the output of the fully-connected layer in this neural network as the features of the images. Subsequently, we introduce a random forest based clustering method to explore the intrinsic structure of human aging thus find the natural groups of the data. Finally, we train independent CNN-based classifiers for each age group for the fine classification. Compared with other hierarchical frameworks [44, 90], the age groups found by our work is closer to their natural intrinsic. The random forest based clustering can find clusters without making any prior assumptions on the number of clusters, and is also robust for the outliers. Also, the features extracted by the convolutional neural network are more robust to the variations of illumination, pose and *etc.* compared to the hand-crafted features.

We evaluated our framework on three popular benchmarks and the experimental results show that our method outperform the state-of-the-art. Our contributions are three folds:

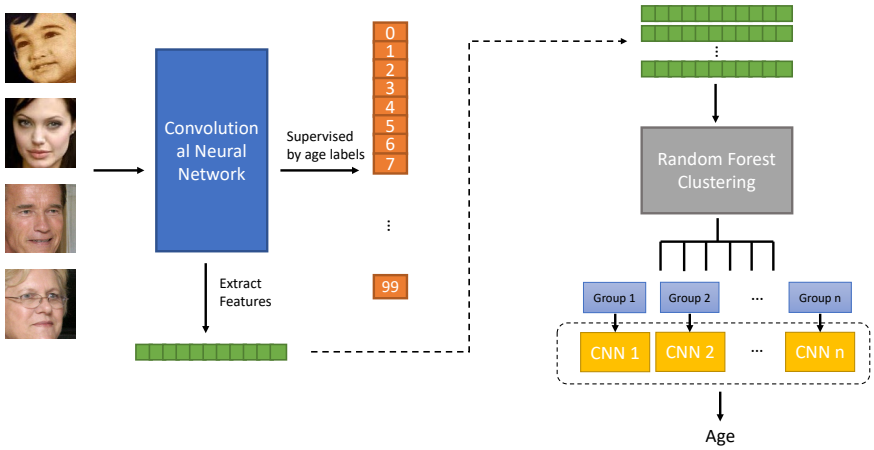


Fig. 3.1 An illustration of our framework. It contains two phases: i) feature learning phase and ii) clustering phase. Firstly, we train a CNN based feature extractor supervised by the age labels. Then we conduct a clustering on the learnt feature to group the ages. For each age group, an independent classifier is trained to predict precise age.

1. We proposed a novel coarse-to-fine age estimation framework. Different with previous works that heuristically assign the ages into predefined groups, our framework performs a clustering on the high-level representation of images that are drawn from a trained neural network to find a natural grouping protocol. Thus our framework can exploit the intrinsic structure of the human aging progression.

2. Compared with existing age estimation methods that employ popular neural network architecture such as AlexNet, Vgg-16, *etc.* In this work, we proposed a novel architecture especially for age estimation. From the experimental results we can find that our architecture is more robust to the variations in the facial images and has fewer parameters thus is able to avoid over-fitting.

3. The experimental results on three popular benchmarks show our framework outperforms the state-of-the-art.

The rest of the chapter is organized as follows: we review the related works in Section 2; our proposal is outlined in detail in Section 3; in Section 4, we discuss the experiments and results; we provide a short conclusion in Section 5.

3.2 Related Work

Age Related Features Human aging progression can be reflected in many aspects. From the view of psychophysics, [133] first studied the morphological changes associated with growth in biological forms. Drawing inspiration from Thompson's work, [119] sought to identify mathematical transformations which describes facial growth event. From the view of computer vision, [19] superimposed aging changes in shape and color on face images to simulate aging variations, and this work was extended by presenting a wavelet-based method for prototyping and transforming facial textures in [36]. In [105], aging variations is modeled by applying a standard facial caricaturing algorithm to the 3D face models. By applying a dense surface point distribution model, [62] used trajectories in the high-dimensional shape-space to express the shape changes associated with growth. These works reveal some of the important facts in the relationship between age and face.

The earliest age estimation work dates back to 1994, [75] uses geometric feature, which calculates the ratios between different measurements of facial landmarks (*e.g.* eyes, chin, nose, mouth, *etc.*), to classify individual into three age groups, namely, that of *infants*, *young adults* and *senior adults*. Geometric feature received good performance in discriminating infants and adult [134], however, it suffers in distinguishing young and senior as both shape and texture of face change during adult aging [126]. To overcome the drawbacks of the geometric feature, Active Appearance Model (AAM) [27] is proposed to model the shape and texture of face image. In AAM the statistic information of the shape and texture of face images in a dataset will be extracted during the training phase and used to extract features the unseen samples. [34] proposed an Aging Pattern Subspace based on AAM and achieved 6.22 years old of mean absolute error (MAE[1]). [27] employed AAM accompanied with a Quadratic Estimator and improved the performance to 4.63 years old. [34] proposed an automatic age estimation method based on facial aging patterns. They defined an aging pattern as a sequence of personal face images sorted in order and construct the aging pattern subspace from the training images. During testing, they traverse all possible positions for the test image in the aging pattern subspace and find the one minimizes the reconstruction error. The position can be finally converted to the predicted age. After 2007, local features became more popular in this field, such as Gabor [66], Local Binary Pattern (LBP) [42], Spatially Flexible Patch (SFP) [143] and Biologically Inspired Feature (BIF) [47]. The pipeline to extract BIF is actually very similar to the artificial neural network. In [47], they used a variation of BIF. First, they use 64 Gabor filters to extract different information along four directions in the first layer, namely $S1$. Followed by layer $C1$, the maximum response

[1] MAE is a measure of difference between two continuous variables X and Y, which is defined as $\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$

within each local spatial neighborhood in $S1$ is selected and flattened into a vector. They then employed the PCA to reduce the dimension of the final representation. The number of works utilizing deep neural networks for age estimation are relatively small, for instances [84, 147] However, the improvements of these works [84] and [147] are somehow introduced by the multi-task settings, in which the former predicts the age as well as the gender and the later predicts the age accompany with the gender and race.

Estimation method Given the aging feature representation, the age estimation can be viewed as a multi-classification problem or a regression problem or a hybrid of the two. For classification, SVM are the most widely employed methods. Using BIF + SVM, [48] achieved MAEs of 3.47 and 3.91 years old for male and female on Yamaha Gender and Age (YGA) datasets. In [20], Cao et al. formulated the age estimation as a ranking problem and proposed a novel method based on Rank-SVM and achieved a result of 5.12 years old MAE on MAE.

For regression, linear regression and SVR are most popular methods in literature. Besides these traditional methods, [46] employed Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA) for age estimation. In [46], they reported that the PLS and CCA yields the best result with a 3.98 years old MAE on MORPH dataset. It is worthy to note that, besides classification and regression settings, [54] proposed a label distribution learning method. Instead of simply outputting a numeral, they predict the age as a distribution. [21] formulates the age estimation as a ranking problem, where they learn several hyperplanes that can separate facial images as their ordinal information. Similar ideas are taken in [24], they propose a ranking convolutional neural network which consists of several basic CNNs. Each CNN predicts whether the age of the input image is higher than a fixed value.

3.3 Method

The pipeline of our proposed system is shown in Fig. 3.1. The training phase mainly consists of two steps. In the first step, a convolutional neural network, namely the feature extraction network, is trained using facial images with strong labels under a classification setting. The strong label here refers to exact age values, for example, from 1 to 100. After the feature extraction network converges, we use it to extract the high-level features of the training images, which are shown as green vectors in the figure. In the second step, we introduce a random forest based clustering method to find the natural age groups on the feature. Each cluster is treated as an age group. Because these age groups are formed in a clustering favor in the high-level feature space, the intrinsic structure of the aging progress is thus revealed. Subsequently, for each age group 1 to n, we train a specific classifier that is also based on

convolutional neural network for the final prediction. In the testing phase, the test images are first feed into feature extraction network to extract features. Then through the random forest clustering, we can get their age groups thereby use the corresponding classifier to estimate the age. In the following of this section, we will discuss each part of our proposed framework respectively.

3.3.1 Feature extraction network

Although many hand-crafted features such as LBP, HOG, *etc.* have been proved effective on age estimation. However, these hand-crafted features are designed based on domain knowledge thus are sensitive to the changes with respect to the conditions of the images. With the emergence of the neural network, the feature learning can be solved simultaneously with the objective optimizing. The advantage is the information can be extracted automatically according to the task without needing to know which specific information it is. Compared to heuristic features, deep learned features contains high-level information including the edge, shape, color, texture, *etc.* Figure. 3.2 shows the architecture of our proposed feature extraction network. The network takes the cropped $224 \times 224 \times 3$ images as the input, and forward propagate in two separate pathways as:

1. The upper pathway consists of three convolution layers with kernel sizes of 7×7, 5×5 and 3×3 repsectively. The kernal size is designed in decreasing manner to extract coarse-to-fine information.

2. Similarly in the lower pathway, there are only two convolution layers with kernal size 3×3 and 1×1. The 1×1 convolution kernel here can capture cross channel information, which is inspired by human visual cortex.

The motivation of employing two pathways is inspired by [122], where a two-stream ConvNet architecture is proposed to extract spatial and temporal information. In the upper pathway, the kernel sizes are set relatively large to extract more global information such as edges. In this pathway, more precise information, for example, the texture is extracted.

After convolution layers, the feature maps from two pathways are flattened and then concatenated into one single vector. Followed by a fully-connected layer and a soft-max layer, the probability of each class is obtained thus a cross-entropy loss can be optimized to train the weights.

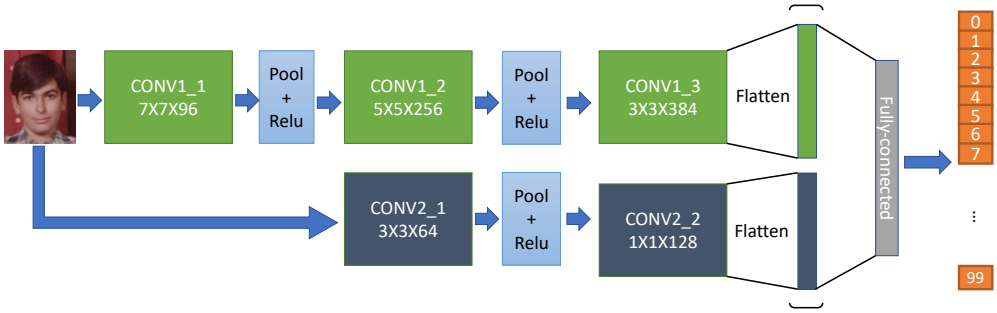


Fig. 3.2 The architecture of proposed feature extraction network. The image is fed into two pathways with different convolution layers. Then the feature maps in each pathway are flattened and concatenated to form the final representation. The feature extraction network is trained by a cross-entropy loss.

3.3.2 Random forest based clustering

As mentioned above, some existing works [43, 44] report that a coarse-to-fine classification for age estimation can enhance the performance. However, existing works either group the age heuristically or uniformly, for example, 10-year-old gap for each group [42]. A fact is that the growing speed in different age periods are not the same. For example, there exists a huge difference between a two-year-old baby and an eight-year-old child, whilst there is not much changes between a 32 and 38 years old adults. We believe that a good grouping protocol is supposed to reflect the intrinsic structure of the age itself, rather than defined empirically. To this end, we introduce a random forest based clustering method to explore an optimal age grouping protocol.

Clustering has been studied extensively in statistics, machine learning, data mining, *etc.* Traditional approaches on clustering can be divided into two categories, partitional clustering and hierarchical clustering. The former partition data points into k groups in which data are more similar to each other, while the latter keeps merging the nearest groups of data records to form clusters. Recently, a decision tree based clustering method was proposed in [88], which has shown its great efficiency and accuracy.

To build the clustering random forest, we first extract features from a feature extraction network. We use $\mathbb{I} = \{I_1, I_2, \cdots, I_N\} \in \mathbb{R}^{224\times224\times3\times N}$ to denote the images, where N is the

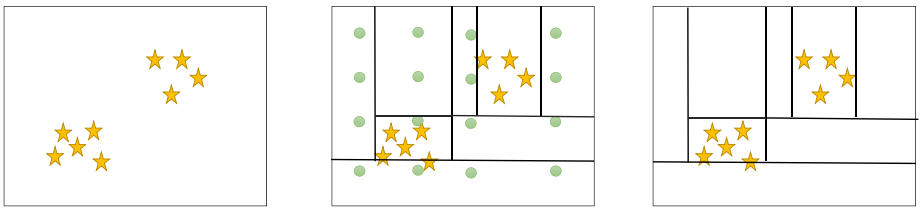


Fig. 3.3 Demonstration of random forest based clustering. (a) Distribution of 'real' data. The data from original dataset are shown as yellow stars. (b) Pseudo data added. The 'pseudo' data (shown as green filled circles) are constructed obeying the uniform distribution and then decision trees are employed to distinguish the 'real' and 'pseudo' data. (c) Removing pseudo data and the partitions of the space can be regarded as the results of clustering.

size of training set. And we use F to denote the feature extraction network, thus the feature can be obtained as $X_i = F(I_i) \in \mathbb{R}^{N\times d}$, where d is the dimension of the feature. Each point of X is labeled as real classes, which are shown as stars in Fig. 3.3 (a). We then build a set of 'non-existing' points $\hat{X}$ with uniform distribution $\hat{X} \sim U(X|\min(X), \max(X))$ in space $\mathbb{R}^d$, which are illustrated as circles in Fig. 3.3 (b). Finally, as shown in Fig. 3.3 (c), the partitions of the space can be obtained by build a random forest to distinguish the real class and pseudo class.

Compared with other clustering method, the random forest based clustering is able to find clusters without making any prior assumptions and non-parametric. Such a characteristic exactly matches our motivation of discovering the intrinsic data structure of the age space. Secondly, the random forest based clustering can find clusters not only in full dimension space but also subspaces, which means it is robust to the dimension of the feature. Moreover, different with other clustering methods, it is not trying to find a centrist of each cluster, which is the same with the age grouping problem, where we can hardly say which age is the center of the group. Most importantly, the random forest based clustering can deal with outliers

efficiently because outliers are typically appears in sparse area in $\mathbb{R}^d$ and the decision trees can easily partition the dense and the sparse area.

3.3.3 Fine age estimation within age groups

With the clustering results from random forest, several age groups can be organized. Suppose there are c groups, and we denote them as $G = \{g_1, g_2, \cdots, g_c\}$. For each group g_i, a specific CNN based classifier is trained for fine age estimation. We first analyze the age distribution within each age group. For group g_i, we have a set of unique labels $\left\{y_1^i, y_2^i, \cdots, y_{n_{g_i}}^i\right\}$ and the corresponding frequencies $\left\{f_1^i, f_2^i, \cdots, f_{n_{g_i}}^i\right\}$, where n_{g_i} is the number of unique labels in group g_i. Then a threshold ε is introduced to filter the outliers. The label set for classifier of g_i is defined as:

$$Y_{g_i} = \left\{y_k^i | f_k^i > \varepsilon, 1 \leqslant k \leqslant n_{g_i}\right\}. \tag{3.1}$$

It is worth noting that, in our proposed age grouping protocol, we found there may exist overlapping between two groups. For example, 35 years old may exist in a group consisting mostly 40+ ages and it may also appear in a group whose majority is 20+. This finding supports our assumption the age groups have their intrinsic structure. Some 35 may be 35 or younger and others may be 35 or elder. This characteristic has never been captured in other settings in the literature.

3.4 Experiments

We evaluated our method on three popular age estimation datasets: MORPH [110], FG-Net [77] and ChaLearn Looking at People dataset (LaP) [33]. The statistics of these three datasets are given in Tab. 4.1. In the followings of this section, we first introduce the details of our experimental settings and the evaluation metric. Then the analysis and discussions of our experimental results on each dataset are given respectively.

Table 3.1 Statistics of the datasets

Dataset	#Images	#Subjects	Age range
MORPH	55608	13000	[16,77]
FGNet	1002	82	[0,69]
LaP	4112	Not given	[0,89]

3.4.1 Experimental Settings

Preprocessing We first detected faces in each dataset by a cascade detector. Specifically, in the LaP dataset, there may exist more than one face in a single image. For images where more than one faces are detected, we only took the face with the highest score. Subsequently, a multi-task CNN was employed to detect the landmarks for each face. After that, we cropped the face image into size 64×64 and aligned them with the locations of eyes.

Data Augmentation Inspired by [84], where over-sampling along four corners of the original images improves the network's performance, in our experiments we augmented the images by random cropping and flipping the original images. We first scaled the input image into 256×256 and then cropped it into 224×224 with random shift along four directions.

3.4.2 Evaluation Metrics

We employed the mean absolute error (MAE) and the cumulative score (CS) to evaluate the age estimation. The MAE is computed as:

$$\varepsilon = \frac{\sum_{i=1}^{N} \|\hat{y}_i - y_i\|_2}{N}, \tag{3.2}$$

where $\hat{y}_i$ and y_i denote the prediction and the ground-truth of the i_{th} testing image, and N indicates the total number of testing images.

The cumulative score at error ε is formulated as:

$$CS(\theta) = \frac{N_{\varepsilon \leq \theta}}{N}, \tag{3.3}$$

where $N_{\varepsilon \leq \theta}$ is the number of the images whose absolute error is less than θ.

3.4.3 Experiments on the MORPH Dataset

The MORPH dataset consists of 55608 images of 13000 subjects, and the age labels in MORPH range from 16 to 77. To have a fair comparison with previous methods, we followed the experimental settings in [21], in which the whole dataset is randomly divided into training and testing set with the ratio of $4:1$ and ensure that no overlap exists between these two subsets. Besides, we performed 5-fold cross-validation.

Table 3.2 Comparison of MAE with different state-of-the-art age estimation methods on MORPH dataset.

Method \ Feature	AAM	BIF	HOG	LBP	SURF	Deep Learning Feature
OHRank	6.07	3.82				
Red-svm	6.49					
SVR	6.99	4.31				
SVM	7.55	4.91				
KNN	9.39					
BT	11.97					
AGES	8.83					
CCA		5.37	4.34	6.13	5.29	
rCCA		4.42				
KCCA		3.98				
PLS		4.56				
KPLS		4.04				
3-Step		4.45				
GEF [90]		4.08	4.05			
DEX						3.25
Yi						3.63
MRCNN						3.27
ORCNN						3.34
DeepRank						3.57
DeepRank+						3.49
DLA						4.77
Ours						3.05

Comparisons with the State-of-the-art Methods

We compared the proposed method with a branch of conventional methods, we choose bio-inspired feature (BIF) as it yields the best performance among hand-crafted features in the past, and the AAM feature as it can extract both the shape and texture information. For the estimation framework, we selected ordinal hyperplances ranker (OHRank), support vector machine (SVM), support vector regressor (SVR), k-nearest neighbor (KNN), binary tree (BT), aging pattern subspace (AGES), canonical correlation analysis (CCA), regularized CCA (rCCA), kernel CCA (KCCA), partial least squares (PLS), kernel partial least squares (kPLS) and 3 step method. We also compared our method with recently proposed deep learning methods, such as deep expectation (DEX), metric regression CNN (MRCNN), ordinal regression CNN (ORCNN), deep rank and DLA.

The experimental results of MAE are shown in Tab. 3.2, It can be seen that the proposed method dominates the state-of-the-art methods with a remarkable gap. Also from Tab. 3.2, we can find that the BIF can achieve best results among all the hand-crafted features when combined with SVM and SVR. It is a very interesting finding and we think this is because the BIF can be regarded as a 4-layer neural network. The Gabor filters in BIF actually play the same role as the convolution operations in CNNs. It is also worth noting that the LBP+CCA achieved better result than BIF+CCA, which reveals the fact that texture information is very important for age estimation task. Compared with [90], which also uses the coarse-to-fine age estimation structure, our method can obtain 1 year old less MAE than them. One may argue that the improvement can be benefit from deep learning techniques as the results in the table show that the deep learning based methods have obtained overwhelming advantages. However, the performance of our method also beats the other deep learning based works.

We also investigated the CS score of the proposed method. In this comparison, we choose some representative works to compare. The CS curves are shown in Fig. 3.4 and our method can be found outperforming the state-of-the-art. It is also worth noting that, the CS score reaches 87% when the age error tolerance is 5 years old. Additionally, we can observe that the rank of each method in Fig. 3.4 stays the same when age error tolerance increases, it is worthwhile to investigate the variance each method. However, since the results of other methods are directly borrowed from their original paper, we cannot obtain their variances, thus we leave it as a future work, which aims to a comprehensive review on age estimation methods.

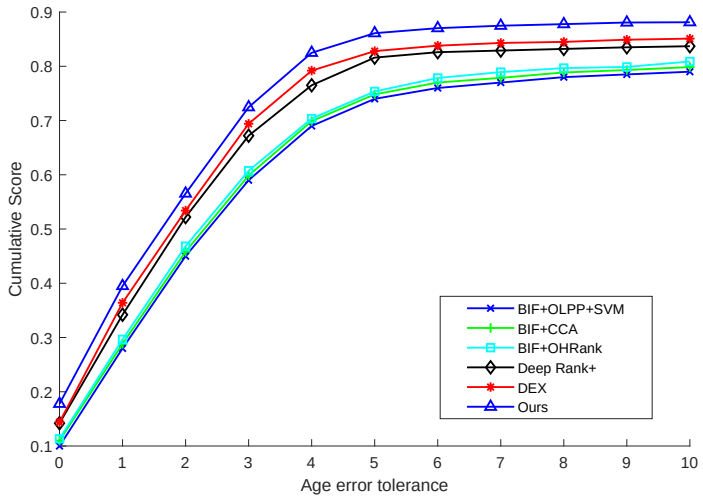


Fig. 3.4 Comparison on Cumulative Score with L in $[0, 10]$ on MORPH dataset

Table 3.3 Comparison of MAE with different state-of-the-art age estimation methods on FGNet dataset.

Method	MAE	Feature	Corse-to-fine
BT	10.35	AAM	✗
KNN	8.96	AAM	✗
SVM	7.25	AAM	✗
AGES	6.77	AAM	✗
SVR	5.91	AAM	✗
RankBoost	5.67	AAM	✗
GP	5.39	BIF	✗
RUN2	5.33	AAM	✗
Red-svm	5.24	AAM	✗
DEX	4.63	Deep	✗
OHRank	4.48	AAM	✗
DLA	4.26	Deep	✗
LARR	5.07	BIF	✓
GEF	3.55	BIF	✓
Ours	3.36	Deep	✓

3.4.4 Experiments on the FGNet Dataset

The FGNet dataset is relatively small and there are totally 1002 colour or grey facial images of 82 individuals with large variations in pose and expression in FGNet dataset. We follow the experimental settings in [21]. As the dataset is too small to train a deep neural network, existing deep learning based works such as DeepRank, MRCNN, ORCNN have not reported their performance on FGNet. In this subsection, we compared our method with many traditional methods, including classification based methods (SVM, KNN, BT, GP, Red-svm), regression based methods (SVR, AGES) and corse-to-fine based methods (GEF, LARR). DEX and DLA as the only two deep learning based methods that reported the performances in their papers are also compared. The experimental results are shown in Tab. 3.4.4. The feature of each method is listed in the third column.

Among the hand-crafted feature based methods, the grouping fusion (GEF) [90] achieved the best MAE. It is interesting to note that, GEF even outperformed the deep learning based methods. Not surprisingly, our method beat the others and yielded the best performance on FGNet although the FGNet is too small for a deep neural network. We conclude this is because there are only three layers in our proposed network and the parameters are relatively less. Moreover, the data augmentation technique helps the network to learn from the dataset.

The cumulative score is also studied for different methods. We can find that our method gained a significant advantage when the age error tolerance at 4 and 5.

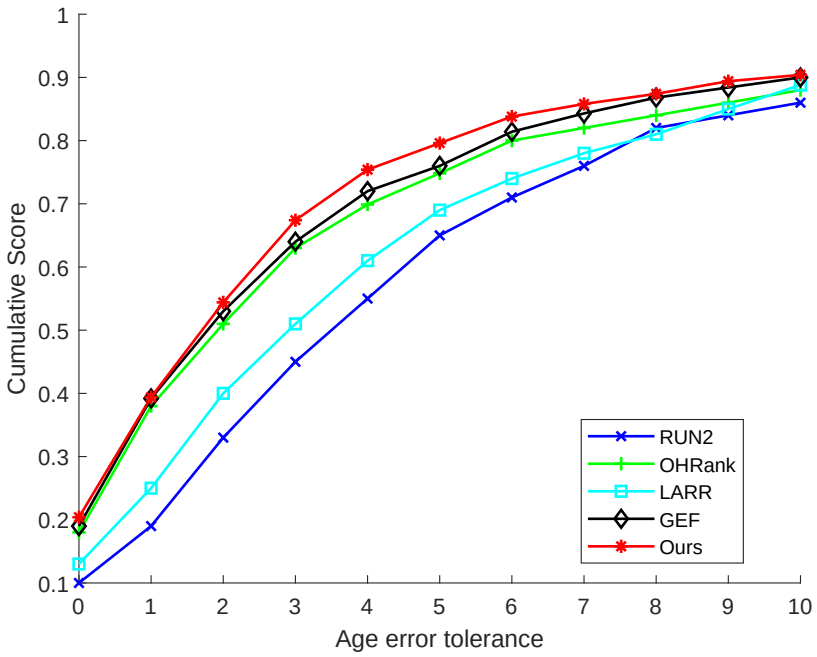


Fig. 3.5 Comparison on Cumulative Score with L in $[0, 10]$ on FGNet dataset

3.4.5 Experiments on the LaP Dataset

To further validate the robustness of our proposed framework, we conduct a comparison on Looking at People competition dataset (LaP). LaP dataset consists of 4112 images for training and 1500 images for validation with apparent age labels ranging from 0 to 89 years old. The images in LaP dataset are more realistic thus challenging, which have large pose, expression, scale and illumination variances. Fig. 3.6 shows some example images of LaP. The performances are evaluated by MAE and Gaussian errors and shown in Tab. 3.4. The Gaussian Error is defined as:

$$\varepsilon = 1 - e^{-\frac{(\hat{y}_i - \mu_i)^2}{2\sigma_i^2}}, \tag{3.4}$$

where $\hat{y}_i$ is the prediction and μ and σ are the mean and standard deviation of the humans annotations on the age of image I_i. Unfortunately, the performance of our method is not the best. We did not yield the lowest gaussian error when compared with the competition teams; however, this is because in the competition, the external data is allowed for the training as reported in [33]. This indirectly reflected that the When compared with ORCNN and DeepRank, which are also trained on the data only from LaP, our model still achieved the best performance, on both gaussian error and MAE.

Table 3.4 Comparison of MAE and Gaussian error with different methods on LaP dataset

		Gaussian error	MAE
Team	CVL_ETHZ	**0.265**	Not given
	ICT-VIPL	0.271	Not given
	WVU_CVL	0.295	Not given
Method	ORCNN	0.322	4.78
	DeepRank	0.301	4.53
	Ours	0.294	4.47

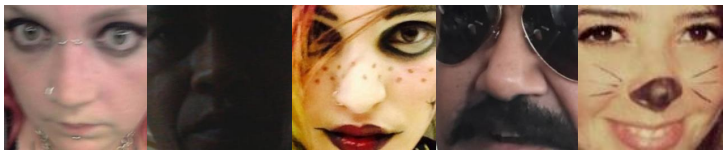


Fig. 3.6 Some 'scary' training sample in LaP dataset.

3.4.6 Experiments on different age group protocols

To evaluate our clustering based age group protocol, we conducted several experiments on different age group settings on MORPH. The original age group protocol in our framework is denoted as *AGP*0, and the age range in each group is shown in the first row in Tab. 3.5. Then we merge each adjacent groups and get the protocol AGP1 to AGP8. We also employ the same age group protocol as in [84] and [42], which are $(0-2,\ 4-6,\ 8-13,\ 15-20,\ 25-32,\ 38-43,\ 48-53,\ 60-)$ and $(10\pm5,\ 20\pm5,\ 30\pm5,\ 40\pm5,\ 50\pm5,\ 60\pm5)$. Because the age range in MORPH is 16-67, we abandoned the 0-2, 4-6 and 8-13 groups for [84]. Similarly, we only keep the age groups that exist in MORPH for [42] and the 66 as well as 67 years old are assigned into 60 ± 5 group. These two settings are denoted as *AGP*9 and *AGP*10. For *AGP*9 and *AGP*10, to get the coarse classes, we trained the classifiers using the same architecture with our feature extraction network. The experimental results are shown in Tab. 3.6. We can find that the self-explored group protocol AGP0 is optimal. It leads the board with a significant advantage. The AGP1,2,6 and 7 are very close to AGP0, while the AGP3,4,5 and 8 are not good. A possible reason is the age distribution is biased in $22-43$ years old, merging the groups in this range will lost the benefits that introduced by

Table 3.5 Age Group Protocols

AGP0	16-23	22-27	28-32	30-43	36-47	48-55	53-67
AGP1	16-23	22-32		30-47		48-67	
AGP2	16-27		28-32	30-47		48-67	
AGP3	16-27		28-43		36-47	48-67	
AGP4	16-27		28-43		36-55		54-67
AGP5	16-23	22-43			36-67		
AGP6	16-32			30-43	36-67		
AGP7	16-32			30-55			53-67
AGP8	16-32	22-55					53-67
AGP9	15-20	25-32	38-43	48-53	60-67		
AGP10	20 ± 5	30 ± 5	40 ± 5	50 ± 5	60 ± 5		

Table 3.6 Experimental results on different age group protocols

Protocol	AGP0	AGP1	AGP2	AGP3	AGP4	AGP5
MAE	3.05	3.12	3.11	3.18	3.18	3.22
Protocol	AGP6	AGP7	AGP8	AGP9	AGP10	
MAE	3.14	3.15	3.19	6.32	3.36	

the coarse-to-fine setting. For AGP9 and AGP10, performances are even worse. For AGP9, because the age groups are not perfectly matched with the age labels, many ages cannot be predict with AGP9, which results in a huge MAE at 6.32.

3.5 Summary

In this chapter, we proposed a novel coarse-to-fine framework for age estimation. We first trained a feature extraction network in a supervised classification favor. Then we introduced the random forest based clustering method to explore the intrinsic structure of age groups. After obtaining each age group, we trained separate fine-grained classifiers for a precise age estimation. A series of systematic experiments have been conducted to evaluate our proposed method. The experimental results show that the proposed method outperforms state-of-the-art.

Chapter 4

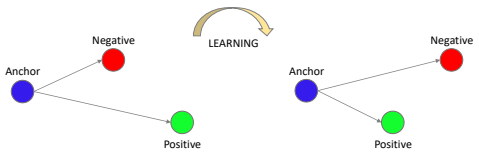
Metric Learning for Age Estimation

4.1 Introduction

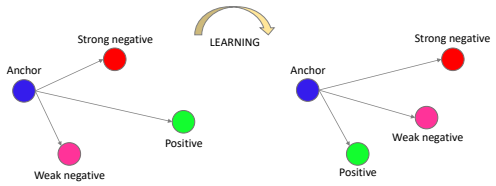
Age estimation has attracted much attention in the computer vision community since the past few decades as human age is one of the most important facial traits during face-to-face communications. Many efforts have been made in actual and appearance age estimation problems to tackle the variations of pose and expression, occlusions, *etc.* Remarkable progress has also been made with emerging deep learning mechanism recently. However, there still exists a non-ignorable gap between the human and machine performance.

Existing works on age estimation can be broadly divided into two branches by the ways they formulate the problem. One branch formulates the age estimation as a multi-classification problem, in which the class label is either a single age or an age group. For example, [84] divides the age labels into 8 age groups. However, in multi-classification frameworks, the labels are assumed independent with each other, which is contrary to the fact that the human age is an ordinal set. By contrast, the other branch that employs regression approaches partially preserve the natural ordinal structure of the human ageing progress by treating each label as a numerical value. Nevertheless, the regression frameworks typically learn a linear kernel while the human ageing is non-linear. For example, the ageing process between 5- and 15- years old is obviously not the same as that between 60- and 70-years old. Besides, since the human ageing is a random process thus non-stationary, learning such non-stationary kernel likely leads to over-fitting.

Recently, a small fraction of works [146, 24, 21, 86] is proposed which formulates the age estimation as an ordinal regression problem. In these approaches, a series of binary classifiers are employed to determine which age slot the input image is supposed to fall in. Some of these works did yield soundable improvements. However, in [24] and [146], k binary classifiers are required to be trained, which is redundant and memory consuming.



(a) The traditional triplet [117] loss minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity.



(b) Our proposed loss considers the distance on distances. Besides the constraints in traditional triplet loss, our proposed loss forces the distance difference, which is between the two distances in traditional triplet loss, to increase if we replace the negative sample with a stronger one.

Fig. 4.1 Comparison between the traditional triplet loss and our proposed quartet loss.

Moreover, training such binary classifiers with deep neural network expects enough data, which is not always feasible. For a small dataset, constructing training set with balanced positive and negative samples for each binary classifier is sometimes tough.

In this chapter, we propose an ordinal metric learning method for the age estimation, in which an ordinal age structure preserving metric is learned to exploit the prolific ordinal information among the age labels; thus the final estimation can be easily conducted in a retrieving manner. The proposed ordinal metric is learned by introducing a proposed quartet model. Concretely, as shown in Fig. 4.1 (b), a quartet consists of an anchor image I_a with the age label $y_a = i$, a positive image I_p which shares the same age label with I_a, a weak negative image $I_{n'}$ with the age label $y_{n'} = j$ and a strong negative image $I_{n''}$ with the age label $y_{n''} = k$ satisfying that $k > j > i$ or $k < j < i$. More concretely, here 'positive' means close enough and 'negative' stands for distant. The ordinal information among age labels can be perfectly preserved by imposing these three natural constraints of human ageing progress on the ordinal metric:

1. The distance between I_a and I_p is smaller than that between I_a and $I_{n'}$ as well as I_a and $I_{n''}$, as the first pair shares the same age label;

2. The distance between I_a and I_p is smaller than that between $I_{n'}$ and $I_{n''}$. This is to ensure a small intra-class variance;

3. The images with different age labels are pushed away from the anchor image by satisfying constraint i). Since the age labels are ordinal, the $I_{n''}$ is supposed to be pushed further away than $I_{n'}$ from I_a.

We evaluated our method on three popular benchmarks and the experimental results outperform the state-of-the-art techniques. Our Contributions mainly lie in the following two aspects:

1. To the best of our knowledge, the proposed method is the first work that formulates the age estimation as an ordinal metric learning problem. Regarding the age estimation as a metric learning problem can generally preserve the ordinal information among the age labels while exploring the non-linear human ageing progress. Moreover, we introduce a deep convolutional neural network guided by a quartet model to optimize the facial representation and the ordinal metric simultaneously.

2. As shown in Fig 4.1, compared with the traditional triplet models that constrains the distance between images, the proposed quartet model constrains a 'high-level' distance which is measured on the distances. For a dataset with n samples, the dimension of the number of the potential quartets is $\mathcal{O}(N^4)$ since given any triplet, one can always find another N samples to construct the quartet, thereby training a deep model on the small dataset becomes feasible. Additionally, the proposed quartet model is able to mine deeper structure information from ordinal data thus can be extended to other similar tasks, such as face retrieval with relative attributes, *etc.*

The rest of the chapter is organized as follows: we review the related works in Section 2; our proposal is outlined in detail in Section 3; in Section 4, we discuss the experiments and results; we provide a short conclusion in Section 5.

4.2 Related Work

Generally, the age estimation can be divided into two subroutines, feature representation and the estimation. In this section, we review the previous works on learning age-related facial

representation, the methods for age estimation as well as some relevant works on metric learning.

Feature Representation: Extracting feature representations for age estimation can be broadly divided into two branches: the hand-crafted feature and the deep learning feature. The earliest approach of age estimation based on facial images dates back to 1994, [75] uses geometric features, in which the ratios between different measurements of facial landmarks (*e.g.* eyes, chin, nose, mouth, *etc.*) are calculated to classify the individual into three age groups, namely *infants*, *young adults* and *senior adults*. Unfortunately, it suffers in distinguishing young and old adults as both the shape and texture of the face change during ageing [126]. To overcome the drawbacks of the geometric features, the Active Appearance Model (AAM) is proposed in [28]. AAM can simultaneously capture the shape and texture information of face images thus yields some improvements at that moment. Later on, the local binary patterns (LBP) [104] are introduced for the age estimation. The LBP feature encodes the facial image into a binary code by comparing the pixels with their neighbours and the texture information are preserved. Similar hand-crafted features such as HDLBP, Gabor, bio-inspired feature are proposed to tackle the different challenges in human age estimation in the past decades, however, tuning the hand-crafted feature is a trial and error process, which is less efficient. With the emerging deep neural network, the feature representation and the age estimator can be learned simultaneously by training an end-to-end neural network. [147] employ a convolutional neural network (CNN) to estimate age as well as classify the gender and race. Similarly, [85] collect a dataset and perform age estimation and gender classification using a CNN.

Age Estimation: Age estimation can be regarded as either a multi-classification problem or a regression problem. In the multi-classification settings, each class can be a single age or an age group. For example, the earliest work [75] simply classify the individual into infants, young adults and senior adults. In [36], each age label is regarded as an independent class while [85] divide the age labels into 8 classes. Classic classification techniques such as SVM, CCA [46], PLS [45] are employed to improve the result. However, the age itself is highly correlated thus the classification approaches which ignore the ordinal information among the age labels are suboptimal. It has been shown that the regression methods [43, 101] outperform the classification methods because in regression settings, the age labels are treated as numerical values. However, the human ageing progress is non-linear thus it is difficult for a regression model with a linear kernel to approximate. Recently [24] formulates the age estimation as an ordinal ranking problem. By training a series of basic CNNs, the ordinal ranking problem is converted to several binary classification problems, where each CNN is only required to determine whether the input image is older than a fixed age. [146] proposed a

scattering network to first extract features, and then employ the principal component analysis (PCA) to reduce the feature dimension, and finally predicts the age via category-wise rankers. These methods yield the-state-of-the art results, however, training several binary classifiers is time-consuming and training each classifier in a one-vs-all fashion may suffer from the unbalanced training set.

Metric Learning: Many machine learning algorithms depend critically on a good metric over the input space. Metric learning target on discovering the relationships between samples and samples. [49] induced a contrastive loss to ensure that the neighbours are pulled together while the non-neighbours are pushed apart on the learned metric. Different with the contrastive loss that only considers pairwise examples at a time, [136] and [117] proposed the triplet loss, which minimizes the L_2-distance between an anchor and a positive sample, both of which belong to the same instance, and maximizes the distance between the anchor and a negative sample. However, the traditional triplet-loss may lead to a large intra-class variation during testing. [25] added a fourth sample in the triplet to enlarge the inter-class variation thus reducing the intra-class variation. [58] proposed a quartet-based model to learn two metrics simultaneously that measure the similarities between images subject to the age and personality. Nevertheless, these works are limited by only considering the relationships between samples and samples. Our proposed quartet model compares the distance on the distances.

4.3 Method

Human ageing is an ordinal and non-linear progress. Conventional multi-classification methods ignore the ordinal information in this progress and the regression approaches try to treat the non-linear ageing progress evenly. To preserve the ordinal information in human ageing while learning the non-linear progress, we formulate the age estimation as an ordinal metric learning problem and introduce a quartet model to impose the learned metric. To keep the mathmatical notation consistent, we use $\mathbb{I} = \{I_1, I_2, \cdots, I_N\}$ to denote the images, where N is the size of training set; and $Y = \{y_1, y_2, \cdots, y_N\}$ are their corresponding labels. Given a possible quartet $(I_a, I_p, I_{n'}, I_{n''})$ which satisfies $y_a = y_p > y_{n'} > y_{n''}$ or $y_a = y_p < y_{n'} < y_{n''}$, our target is to learn a metric $\mathscr{D}(F(I_m), F(I_n))$, where the $F(I_m)$ denotes the feature representation of I_m, m and n are two arbitrary image indices and $\mathscr{D}(\cdot,\cdot)$ is a metric that measures the distance between the feature of image I_m and I_n. The larger the $\mathscr{D}(F(I_m), F(I_n))$ is, the more dissimilar the I_m and I_n are. In the following of this section, we discuss the expected properties of the metric $\mathscr{D}(\cdot,\cdot)$ and the details of our proposed framework.

4.3.1 Conventional Triplet Loss: Distance between Images

As shown in Fig. 4.1 (a), conventional triplet model consists of an anchor sample I_a, a positive sample I_p, which shares the same label with the anchor, and a negative sample I_n which is of a different label. A typical objective function of the triplet model is formulated as:

$$L_{\text{triplet}} = H(\delta, \mathscr{D}(F(I_a), F(I_n)), \mathscr{D}(F(I_a), F(I_p))), \tag{4.1}$$

where $H(\delta, \alpha, \beta) = \max[0, \delta - (\alpha - \beta)]$ is a hinge loss. By minimizing the L_{triplet} in Eq. 4.1, the dissimilar image x_n is pushed at least a margin δ further away than the x_p from the anchor x_a. Thus the retrieval or the other following tasks can be conducted on the learned metric.

It is natural to employ the triplet model to learn a metric for the age estimation as each age label can be treated as a class. In an age estimation problem, we have:

Constraint 1 *The image that shares the same age label with the anchor is supposed to be closer to the anchor on the metric than the one with a different label.*

Thereby for a quartet $(I_a, I_p, I_{n'}, I_{n''})$, two triplet loss can be obtained as:

$$\begin{aligned} L_1 &= H(\delta_1, \mathscr{D}(F(I_a), F(I_{n'})), \mathscr{D}(F(I_a), F(I_p))), \\ L_2 &= H(\delta_2, \mathscr{D}(F(I_a), F(I_{n''})), \mathscr{D}(F(I_a), F(I_p))). \end{aligned} \tag{4.2}$$

As reported in [25], the conventional triplet loss may result in a large intra-class variation as well as a small inter-class variation. Inspired by their work, we have:

Constraint 2 *The images sharing the same age label should be closer than those with diverse labels on the metric.*

To satisfy constraint 2, a loss function can be derived as:

$$L_3 = H(\delta_3, \mathscr{D}(F(I_{n'}), F(I_{n''})), \mathscr{D}(F(I_a), F(I_p))). \tag{4.3}$$

Regarding $H(\delta, \alpha, \beta)$ as a mechanism that pushes α a margin of δ away from β, it can be found that the H is applied on the distances between images in Eq. 4.2 and 4.3. Thus we name them the 'distance-between-images' loss. With exploring and leveraging the relationships between samples, the triplet based metric learning approaches outperform the classification and regression ones in some area; and an essential reason is that generally the classification and regression frameworks only consider the relationships between the samples and labels, which is with the dimension of $\mathscr{O}(N)$ in a dataset with scale N, while the triplet based metric learning utilises $\mathscr{O}(N^3)$s. However, a 'distance-between-images' loss is not capable of preserving the ordinal information in human ageing.

4.3.2 Quartet Loss: Distance on Distance

Equation 4.1 constrains the distance between I_a and the weak negative sample $I_{n'}$ to be at least a δ_1 larger than that between I_a and I_p. Similarly, a margin of δ_2 is enforced for the strong negative sample $I_{n''}$. As the age label is an ordinal set, it is intuitive that:

Constraint 3 *The strong negative images are supposed to be pushed further than the weak negative images on the metric.*

If having:

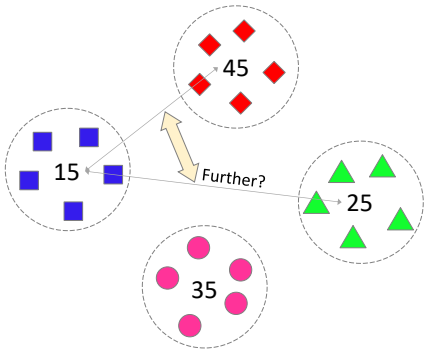
$$\begin{aligned} \Delta^{'} &= \mathscr{D}(F(I_a), F(I_{n'})) - \mathscr{D}(F(I_a), F(I_p)), \\ \Delta^{''} &= \mathscr{D}(F(I_a), F(I_{n''})) - \mathscr{D}(F(I_a), F(I_p)), \end{aligned} \tag{4.4}$$

where $\Delta^{'}$ is the difference of the distance between I_a and $I_{n'}$ and that between I_a and I_p; and $\Delta^{''}$ is defined in a similar fashion. It can be deduced that the $\Delta^{''}$ is supposed to be larger than $\Delta^{'}$. Therefore a hinge loss can be obtained as:

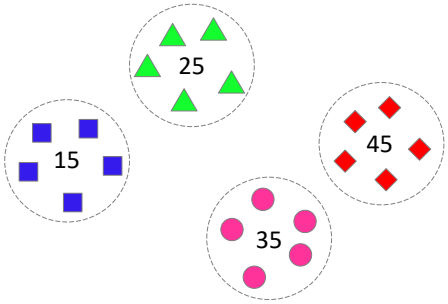
$$L_4 = H(\delta_4, \Delta^{''}, \Delta^{'}). \tag{4.5}$$

Noticing that in the Eq. 4.5, the mechanism H is applied to the difference of the distance, we refer to this quartet loss as the 'distance-on-distances' loss. With the 'distance on distances' loss, the ordinal information is preserved. Furthermore, the 'distance on distances' loss explores the relationships of the distances between samples, thus more information can be utilized for the training.

A comparison between the conventional triplet loss and the quartet loss is illustrated in Fig. 4.2. As the conventional triplet loss only constrains the distance between similar pairs and dissimilar pairs, the samples which share the same age label are formed into clusters as shown in Fig. 4.2 (a). Although the L_3 in Eq. 4.3 can be induced to reduce the intra-class variation, the ordinal information is ignored. Thereby it may result in a situation where the centroid of the age 25 is further away from that of the age 15 compared with the centroid of the age 45, which is not acceptable for the following estimation. This drawback is solved by introducing our proposed quartet loss. As illustrated in Fig. 4.2 (b), the samples are clustered in the order of their labels.



(a) Under the mechanism of the traditional triplet loss, the samples sharing the same age labels are formed into clusters, which is favoured by a typical classification or regression model. However, the centroid of the age 25 is further away from that of the age 15 compared with the centroid of the age 45.



(b) With the proposed quartet loss, the strong negative samples are pushed further away; therefore the ordinal information is preserved.

Fig. 4.2 A comparison between the conventional triplet loss and the proposed quartet loss on age estimation.

4.3.3 Optimization

The overall objective function is written as:

$$\mathscr{L} = \lambda_1(L_1 + L_2) + \lambda_2 L_3 + \lambda_3 L_4, \tag{4.6}$$

where λ_{1-3} are the weights of each loss. To minimize $\mathscr{L}$, we employ a deep neural network to simultaneously optimise the feature extraction f and the ordinal metric $\mathscr{D}(\cdot,\cdot)$. Our proposed network is shown in Fig. 5.4. The network takes quartet as input and the four images are first passed through four weight-shared convolutional layers to extract the latent

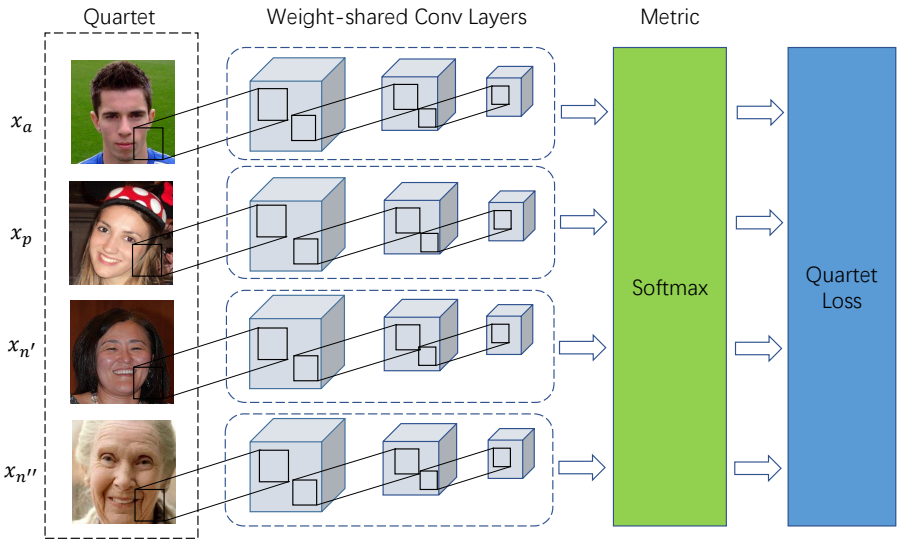


Fig. 4.3 The architecture of our proposed deep network.

representations. Subsequently, a softmax layer is introduced as the metric to ensure that the value of $\mathscr{D}$ lies in $[0, 1]$.

Although the quartet model is able to leverage more information in the dataset, the computation cost for the large-scale dataset can be a matter during training. Several techniques have been proposed for the triplet selection in [25, 117]. Generally, the selection protocol can be divided into online and offline selection branches. The offline selection scheme uses a pre-trained model to select hard positives and negatives in a subset. By contrast, in the fashion of the online selection, a mini-batch is forwarded through the current network and hard samples are selected according to their corresponding contributions to the loss. In this approach, we employ the online selection protocol and select the hard positive samples as well as randomly add negative samples.

4.4 Experiments

We evaluated our method on three popular age estimation datasets: MORPH [110], FG-Net [77] and ChaLearn Looking at People dataset (LaP) [33]. The statistics of these three datasets are given in Tab. 4.1. In the followings of this section, we first introduce the details of our experimental settings and the evaluation metric. Then the analysis and discussions of our experimental results on each dataset are given respectively.

Table 4.1 Statistics of the datasets

Dataset	#Images	#Subjects	Age range
MORPH	55608	13000	[16,77]
FGNet	1002	82	[0,69]
LaP	4112	Not given	[0,89]

4.4.1 Experimental Settings

Preprocessing We first detected faces in each dataset by a cascade detector. Specifically, in the LaP dataset, there may exist more than one face in a single image. For images where more than one faces are detected, we only took the face with the highest score. Subsequently, a multi-task CNN was employed to detect the landmarks for each face. After that, we cropped the face image into size 64×64 and aligned them with the locations of eyes.

Hyper-parameters For the hyper-parameters used in our method, we set $\delta_1 = 0.01$, $\delta_2 = 0.01$, $\delta_3 = 0.02$, $\delta_4 = 0.01$, $\lambda_1 = 0.5$, $\lambda_2 = 0.3$ and $\lambda_3 = 0.2$ empirically.

Testing method and evaluation metrics As we estimate the age in a retrieving manner on the learned ordinal metric, we randomly selected a subset from the training set as the reference set and take the nearest candidate's label as the output. An alternative subset selection protocol is discussed in the experiments on LaP dataset. We employed the mean absolute error (MAE) and the cumulative score (CS) to evaluate the proposed method.

4.4.2 Experiments on the MORPH Dataset

The MORPH dataset consists of 55608 images of 13000 subjects, and the age labels in MORPH range from 16 to 77. To have a fair comparison with previous methods, we followed the experimental settings in [21], in which the whole dataset is randomly divided into training and testing set with the ratio of 4 : 1 and ensure that no overlap exists between these two subsets. Besides, we performed 5-fold cross-validation.

Comparisons with the State-of-the-art Methods

To compare our approach with state-of-the-art methods, we divided the existing methods into two branches of which one utilises the hand-crafted features and the other employs the deep learning features. Concretely, we choose bio-inspired feature (BIF) as it yields the best performance among hand-crafted features in the past. We picked the Canonical Correspondence Analysis (CCA), Orthogonal Locality Preserving Projection (OLPP) + SVM and the Ordinal Hyperplanes Ranker (OHRanker) as the estimation mechanism for the hand-crafted feature. For the deep learning approaches, we compared with the ORCNN and

Table 4.2 Comparison of MAE with different state-of-the-art age estimation methods on MORPH

Method	MAE
BIF+CCA	3.98
BIF+OLPP+SVM	4.99
BIF+OHRanker [21]	4.95
MRCNN [102]	3.42
ORCNN [102]	3.27
DeepRank [146]	3.57
DeepRank+	3.48
Ours	**3.07**

DeepRank [146]. We directly borrowed the results reported in their papers. The experimental results of MAE are shown in Tab. 4.2 and the CS curves are illustrated in Fig. 4.4. It can be seen that the proposed method dominates the state-of-the-art methods with a remarkable gap and even outperforms the BIF+CCA, in which the extra gender information is taken into consideration.

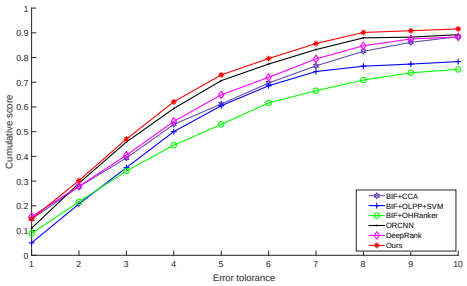


Fig. 4.4 Comparison on Cumulative Score with L in $[0, 10]$ on MORPH dataset

4.4.3 Experiments on the FGNet Dataset

The FGNet dataset is relatively small and there are totally 1002 colour or grey facial images of 82 individuals with large variations in pose and expression in FGNet dataset. We follow the experimental settings in [21]. In the first stage, we compared our methods with SVM, SVR, RankBoost and OHRanker by employing BIF feature as the representation. The MAEs are listed in the Tab. 4.3. It can be found that our method outperforms the conventional methods which using hand-crafted features. One may argue that the improvements are benefits from the power of the deep neural network. However, there are currently no results reported from those deep learning approaches as training a deep network with such small

Fig. 4.5 Illustration of the two selected reference sets on LaP dataset. The first row (set a) was selected by picking the images with the highest confidence score and the second row (set b) is selected randomly.

dataset is extremely difficult. To be fair, we further conducted an experiment to compare with those deep learning methods.

Table 4.3 Comparison of MAE with different state-of-the-art age estimation methods on FGNet dataset

Method	MAE
SVM	7.32
SVR	5.76
RankBoost	5.57
OHRanker	4.48
Ours	**3.96**

Comparisons with the Deep Learning Approaches

Training a deep neural network on FGNet is generally infeasible as insufficient data may lead to over-fitting. However, as our proposed quartet model can utilize $\mathcal{O}(n^4)$ information from the dataset, training our network becomes reliable. To avoid the influence of the number of the parameters, in this section, we trained our network, the ORCNN and the DeepRank with the same feature extraction layers and only altered the final objective function. The experimental results are shown in the Tab. 4.4. It is not surprising that the DeepRank and ORCNN suffer from the insufficient data although they trained k classifiers and k different subsets can be used. The experimental results on the FGNet dataset illustrate that our proposed method can handle small-scale dataset by leveraging more information.

4.4.4 Experiments on the LaP Dataset

To further validate the robustness of our proposed framework, we conduct a comparison on Looking at People competition dataset (LaP). LaP dataset consists of 4112 images for

Table 4.4 Comparison of MAE with different deep learning methods on FGNet dataset

Method	MAE
MRCNN	4.72
ORCNN	4.65
DeepRank	4.32
Ours	**3.96**

Table 4.5 Comparison of MAE and Gaussian error with different methods on LaP dataset

		Gaussian error	MAE
Team	CVL_ETHZ	**0.265**	Not given
	ICT-VIPL	0.271	Not given
	WVU_CVL	0.295	Not given
Method	ORCNN	0.322	4.78
	DeepRank	0.301	4.53
	Ours(b)	0.279	4.31
	Ours(a)	0.274	**4.27**

training and 1500 images for validation with apparent age labels ranging from 0 to 89 years old. The images in LaP dataset are more realistic thus challenging, which have large pose, expression, scale and illumination variances. The performances are evaluated by MAE and Gaussian errors and shown in Tab. 4.5. It can be seen from the table that compared with the state-of-the-art methods, the performance of our method is the best. We did not yield the lowest Gaussian error when compared with the competition teams; however, this is because in the competition, the external data is allowed for the training reported in [33]. Additionally, in this experiment, we constructed two reference sets to conduct the retrieval, namely the reference set *a* and *b*. In reference set *b*, we randomly choose images from each age label while in the reference set *b*, we selected the images with the highest confidence score. The results are denoted as Ours(a) and Ours(b) respectively in Tab. 4.5. One can draw the conclusion that the selection of the reference set does influence the estimation result but not much, which in other side shows the robustness of our framework.

4.5 Summary

In this chapter, we propose a metric learning approach for the age estimation problem. With a proposed quartet model, the natural constraints of human aging progress are imposed on the learned metric; thus the ordinal information is preserved. We conducted the experiments on three benchmarks, and the experimental results show that our method outperforms the

state-of-the-art methods. Additionally, the experiment on FGNet dataset shows that our method still works on small dataset as quartet model exploring deeper structural information.

Chapter 5

Dual Reference Face Retrieval

5.1 Introduction

Over the past few decades, face retrieval has received great interest in the research community for its potential applications such as finding missing persons [65] and matching criminals with CCTV footage for law enforcement [131]. Apart from a pinch of face retrieval works that are text based [13], most existing frameworks are based on the content, in which a target person's image is required as the query input, and the system retrieves all the images belong to the target person in the database. Though these works in some kind improved the benchmark in the past, they fail to catch up with the pace of the new demands of the face retrieval in the age of big data. For example, rather than retrieving all the query identity's images indiscriminately, we may prefer picking out the specific ones with some certain attribute, *e.g.* age. Huge volume of online images make this kind of fine-grained face retrieval both feasible and indispensable. It is feasible as such large scale dataset can contain many images taken from someone's different age periods, thereby it is necessary to select them out in some potential applications.

Considering such a task – *retrieving Emma Watson's image at 23*, although it is not absolutely impossible for conventional face retrieval frameworks to handle, as it can be solved by concatenating an age estimation system at the end to select the right images as illustrated in Fig. 5.1.(a), there are many drawbacks in such a hierarchical framework. One of them is that using a single numeral is not capable to describe the human perceptions of the age, because the human performance on age estimation is with a large mean absolute error(MAE) as well as a large variance [50], which means generally a human prefers to guess the age within a range rather than a certain numeral. Also, for humans, it is easier to estimate someones' age by comparing with age-known faces than directly assigning a facial image to a numeral [21].

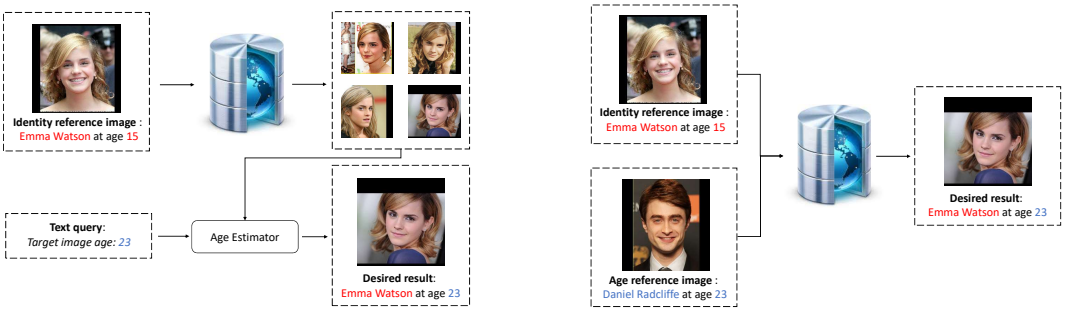


Fig. 5.1 Comparison between conventional face retrieval framework and our proposed dual-reference face retrieval framework

As the old saying goes, 'One look is worth a thousand words', the problem of *retrieving Emma Watson's images* is better solved by inputting one Emma Watson's picture and telling the machine: *retrieving someone's images, and the 'someone' is shown in the picture*. Since a numeral is not representative enough to describe a person's age, and in some scenarios we do not even care about the certain age but the similarities in term of the age, what if we use an image to represent the target age? In this chapter, we proposed a novel face retrieval framework as shown in Fig. 5.1.(b), in which an age reference image is inputted to reflect the target age besides the identity reference image. We refer to the proposed framework as dual-reference face retrieval (DRFR). With the DRFR, the problem of *retrieving Emma Watson's image at 23* is turned into *retrieving someone's images, which are in the similar age of the age reference image*.

In DRFR, the raw images are first projected onto a joint manifold, which preserves both the age and identity locality. Subsequently, as the age and identity are measured differently on the joint manifold, a similarity metric for each is exploited and optimized via our proposed quartet-based model shown in Fig. 5.3. The final retrieval is conducted on the learned metric.

The contributions of our proposed method mainly lie in the following three aspects:

1. The task: retrieving someone's image at some age is an emerging task as more and more precise retrieval is required due to the explosive web images.

2. The model: a joint manifold of identity and age is exploited in this chapter, it simultaneously preserves the localities of these two aspects. Besides, a novel quartet-based model coordinated with two Mahalanobis distances is proposed to measure the similarity between image pairs.

3. The framework: our proposed DRFR task can be abstracted to a high-level task — dual reference/query retrieval, which might lead to an emerging research direction. The existing retrieval methods generally take a single query or multiple queries indicating the same semantic information, while in our dual-reference framework, more than one semantic information can be taken into consideration.

The remainder of the chapter is organized as follows: we review related works in Section 2; our proposal is outlined in detail in Section 3; in Section 4, we discuss the experiments and results; we provide a short conclusion in Section 5.

5.2 Related Work

To the best of our knowledge, the task of the dual reference face retrieval has never been raised in the literature, and there are no similar existing works, thus we review related works in the areas of face retrieval and age estimation, focusing on those papers which explore facial feature representation, age variation capturing and similarity metric learning.

Facial Feature Representation: A broad array of research has been completed on facial feature representation. As facial features extracting is not the core part of our framework, we just give a rough review here. For a comprehensive review, we refer our readers to [9]. Early works mainly take heuristic features such as Gabor [89], HOG [31], LBP [4] or their extensions such as FPLBP [141], LTP [129]. However, designing hand-crafted features is a trial and error process which is less than adequate for our purpose. Another branch of research regarding facial features is based on utilizing deep learning. For example, [127] employed a nine-layer deep neural network to extract facial features for face verification and [125] proposed a carefully designed deep convolutional networks for joint face identification-Verification.

Age Variation Capturing: Age variation capturing is rarely considered in conventional face retrieval approaches because in most works to date, features are required to be age-invariant. In contrast, as we need to retrieve the image of a 'certain' age, we need a framework which embeds the age variation in our final facial representations. Approaches capturing age variation can primarily be found in age estimation literature. The earliest approach of age estimation based on facial images dates back to 1994, [75] uses geometric features, in which the ratios between different measurements of facial landmarks (*e.g.* eyes, chin, nose, mouth, *etc.*) are calculated to classify the individual into three age groups, namely *infants*, *young adults* and *senior adults*. Unfortunately, it suffers in distinguishing young and old adults as both the shape and texture of the face change during aging [126]. To overcome the drawbacks of geometric features, the Active Appearance Model (AAM) is proposed in [28].

AAM is able to simultaneously capture the shape and texture information of face images. Our proposal is inspired by Aging Pattern Subspace [34], in which a serial of a person's images is treated as an aging pattern. However, our proposed joint manifold is different because we also embed the identity information at the same time.

Similarity Metric Learning: Once the proper facial image representation is selected, the retrieval is conducted based on the similarity measurements. [49] induced a contrastive loss to ensure that the neighbors are pulled together while the non-neighbors are pushed apart on the learned metric. Different with the contrastive loss that only considers pairwise examples at a time, [136] and [117] proposed the triplet loss, which minimizes the L_2-distance between an anchor and a positive sample, both of which belong to the same instance, and maximizes the distance between the anchor and a negative sample. However, the traditional triplet-loss may lead to a large intra-class variation during testing. [25] added a fourth sample in the triplet to enlarge the inter-class variation thus reducing the intra-class variation.

5.3 Method

For convenience, we define I_i^m as an image of the individual with identity i at age m. Input an image pair (I_i^m, I_j^n), where i is the target identity and n is the objective age, thus our required output is I_i^n. As discussed, DRFR consists of two stages. Firstly, a mapping function is learned to project the raw images onto a joint manifold. Subsequently, to measure the similarity between each pair of images, the two metrics are learned on the low-dimensional space, based on a quartet model. We devote the rest of this section to outlining these two stages.

5.3.1 Joint Manifold

A face image with d-dimensional feature representation can be considered as a point in the d-dimensional space containing rich information such as age, gender, race, identity. Manifold learning is first proposed in [113], in which they believe that the high-dimensional data is sampled from a smooth low-dimensional manifold. Thus it is natural that information from a facial image can be represented within low-dimensional manifolds embedded in a high-dimensional image space [52]. Many applications already utilize low-dimensional manifolds to embed human face images, such as face recognition and age estimation. However, our proposed joint manifold as illustrated in Fig. 5.2 is very different; instead of treating the age and identity as two separate degrees of freedom in a single manifold, with the assumption that the age and identity are both manifolds sampled from a higher-dimensional manifold.

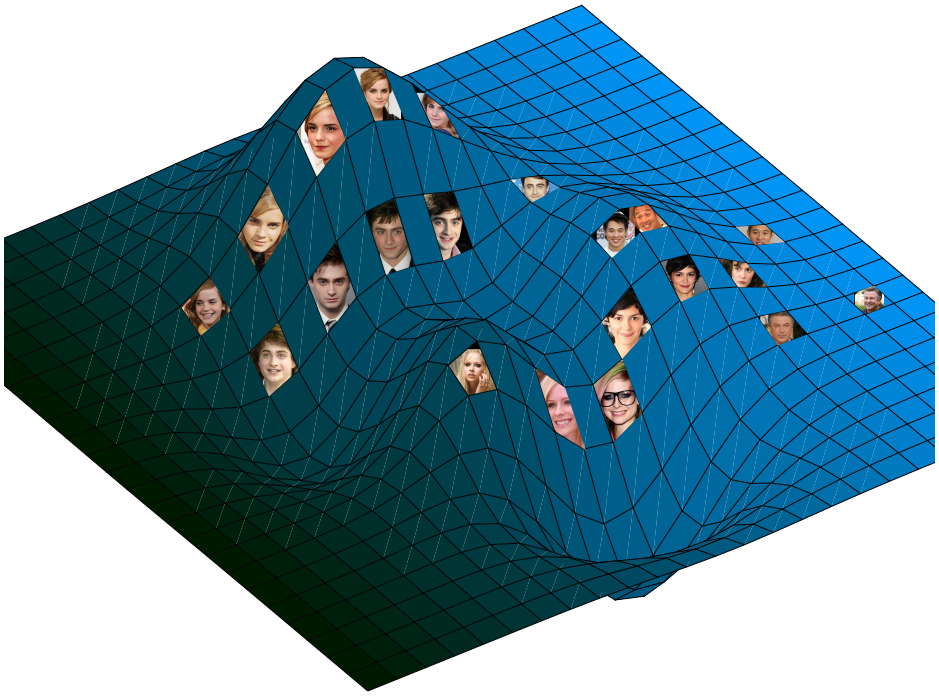


Fig. 5.2 An illustration of the joint manifold of age and identity.

Let X be the original representation of the raw images and Y be the low-dimensional joint manifold, define the mapping function of the joint manifold to be $F : X \to Y$. Since both the locality of the age and identity can be represented as matrices, let S denote the set of all such similarity matrices. Specifically, the matrix $S^n \in S$ reflects the similarity among all the individuals' images at age n; similarly, S_i denotes the similarity over those images belonging to an individual with identity i across all ages. The desired properties of f are discussed below.

Preserving locality of individual space

We first calculate the similarity matrix S^n. In detail, among all the images at age n, if two images are nearby in original feature space X, we mark the similarity[1] as $\exp\left(-\frac{\|x_i^n - x_j^n\|_2^2}{t}\right)$, where $x_i^n \in X$ is the original feature representation of image I_i^n and $\|\cdot\|_2^2$ is the $l2$-norm, otherwise their similarity is 0. Thus the similarity matrix S^n under age n is calculated as:

$$S^n(x_i^n, x_j^n) = \begin{cases} \exp\left(-\frac{\|x_i^n - x_j^n\|_2^2}{t}\right) & \text{if } x_j^n \in \mathcal{N}(x_i^n), \\ 0 & \text{otherwise,} \end{cases} \tag{5.1}$$

[1] t is set as 1 here.

where $\mathcal{N}(x_i^n)$ denotes the neighbors of x_i^n. To preserve the locality, we require the nearby points in X to remain close to each other after being embedded into $Y = F(X)$, thus we optimize the function:

$$\min_F \sum_n \sum_{i,j} \| F(x_i^n) - F(x_j^n) \|_2^2 S^n(x_i^n, x_j^n). \tag{5.2}$$

Preserving locality of age space

Similarly, to calculate the age similarity matrix S_i, we gather all the images of the individual i, and assign $\exp\left(-\frac{\|x_i^n - x_i^m\|_2^2}{t}\right)$ as the similarity if $m-n$ is below a threshold ε, otherwise the similarity is 0:

$$S_i(x_i^n, x_i^m) = \begin{cases} \exp\left(-\frac{\|x_i^n - x_i^m\|_2^2}{t}\right) & \text{if } |m-n| < \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

To preserve the local smoothness, we optimize the function:

$$\min_F \sum_i \sum_{m,n} \| F(x_i^n) - F(x_i^m) \|_2^2 S_i(x_i^n, x_i^m). \tag{5.4}$$

5.3.2 Similarity Metric Learning Based on a Quartet Model

After both the original age and identity spaces are mapped onto a joint manifold, different measurements should be taken to obtain the similarity of the two aspects. In the proposed model, two similarity metrics are learned based on a novel quartet model, which is a graph with 4 vertices as shown in Fig. 5.3. The vertices sets $V = \{F(x_i^m), F(x_i^n), F(x_j^m), F(x_j^n)\}$ are the embedded points of $\{(x_i^m), (x_i^n), (x_j^m), (x_j^n)\}$, and the edges are defined as the distance between each embedded point. We use $\Phi(\cdot,\cdot)$ to denote the difference measurement function whereby the smaller $\Phi(\cdot,\cdot)$ is, the more similar the two images are. In the following of this subsection, the properties of the desired metrics are introduced.

Individual metric

Considering two image pairs (x_i^m, x_i^n) and (x_j^m, x_i^n), which are shown in the quartet model in Fig. 5.3, it is very clear that on the individual metric, the distance between x_i^m and x_i^n is smaller than that between x_i^m and x_j^n, because these two pairs of images both have the age gap $m-n$ while the first image pair (x_i^m, x_i^n) belongs to the same individual i. Mathematically,

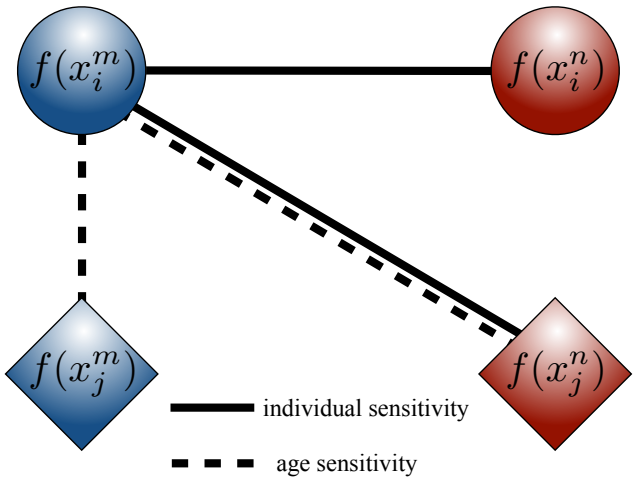


Fig. 5.3 An illustration of our proposed quartet model. The blue symbols indicate the embedded points of images at age m and the red ones stand for those at age n. The circle symbols represent the embedded points of images of individual i while the diamond ones stand for those of individual j. The lengths of the lines connecting any two symbols can be regarded as the distance between the corresponding embedded points. Thus in any triangle in the quartet sample, the length of its hypotenuse is larger than that of its leg.

there is:

$$\begin{aligned} &\Phi_{ind}(F(x_i^m),F(x_i^n)) < \Phi_{ind}(F(x_i^m),F(x_j^n)) \\ &\forall(i,j,m,n), \end{aligned} \tag{5.5}$$

where Φ_{ind} measures the individual difference between any pair of images.

Additionally, the distances between image pair (x_i^m,x_j^m) and (x_i^m,x_j^n) are supposed to be similar because the individual metric is uncorrelated with the age, which can be written as:

$$\begin{aligned} &\Phi_{ind}(F(x_i^m),F(x_j^m)) = \Phi_{ind}(F(x_i^m),F(x_j^n)) \\ &\forall(i,j,m,n), \end{aligned} \tag{5.6}$$

Age metric

Similarly on the age metric, the distance between image pair (x_i^m,x_j^m) is smaller than that between (x_i^m,x_j^n), and the distances are close if the age gap within each image pair is same.

Thus we have:

$$\begin{aligned}&\Phi_{age}(F(x_i^m),F(x_j^m)) < \Phi_{age}(F(x_i^m),F(x_j^n))\\&\forall(i,j,m,n),\end{aligned} \tag{5.7}$$

$$\begin{aligned}&\Phi_{age}(F(x_i^m),F(x_j^n)) = \Phi_{age}(F(x_i^m),F(x_i^n))\\&\forall(i,j,m,n),\end{aligned} \tag{5.8}$$

where Φ_{age} measures the age difference between any pair of images.

Quartet loss

To obtain the discussed characteristics of the individual and age metrics, a loss function which maximize the margin between the distances in Eq. 5.5 and Eq. 5.7, and meanwhile minimize the margin between the distances in Eq. 5.6 and Eq. 5.8 is designed. For convenience, we first define d as the distance of two images embedded in the joint manifold $\mathscr{Y}$: $d_{ij}^{mn} = F(x_i^m) - F(x_j^n)$ and take the Mahalanobis distance as the distance measurement. Thus the $\Phi(\cdot,\cdot)$ can be written as:

$$\begin{aligned}\Phi_{age}(F(x_i^m),F(x_j^n)) &= {d_{ij}^{mn}}^{\top}\mathbf{M}_{\text{age}}d_{ij}^{mn},\\ \Phi_{ind}(F(x_i^m),F(x_j^n)) &= {d_{ij}^{mn}}^{\top}\mathbf{M}_{\text{ind}}d_{ij}^{mn},\end{aligned} \tag{5.9}$$

where $\mathbf{M}_{\text{age}}$ and $\mathbf{M}_{\text{ind}}$ are the Mahalanobis matrices. To maximize the margin, the hinge loss function:

$$H(y) = \max(0, \delta - y) \tag{5.10}$$

is employed.

Thereby for a quartet sample indexed by (i,j,m,n), the loss L_{ij}^{mn} can be defined as:

$$\begin{aligned}L_{ij}^{mn} =& H({d_{ij}^{mn}}^{\top}\mathbf{M}_{\text{age}}d_{ij}^{mn} - {d_{ij}^{mm}}^{\top}\mathbf{M}_{\text{age}}d_{ij}^{mm})\\ &+H({d_{ij}^{mn}}^{\top}\mathbf{M}_{\text{ind}}d_{ij}^{mn} - {d_{ii}^{mn}}^{\top}\mathbf{M}_{\text{ind}}d_{ii}^{mn})\\ &+||{d_{ij}^{mn}}^{\top}\mathbf{M}_{\text{age}}d_{ij}^{mn} - {d_{ii}^{mn}}^{\top}\mathbf{M}_{\text{age}}d_{ii}^{mn}||_2^2\\ &+||{d_{ij}^{mn}}^{\top}\mathbf{M}_{\text{ind}}d_{ij}^{mn} - {d_{ij}^{mm}}^{\top}\mathbf{M}_{\text{ind}}d_{ij}^{mm}||_2^2.\end{aligned}$$

And the loss over the whole training set is

$$L = \sum_{i,j,m,n} L_{ij}^{mn}. \tag{5.11}$$

5.3.3 Optimization

Considering the loss function $\mathscr{L}$ and the joint manifold as the regularization term, the overall objective function is:

$$\begin{aligned} L_{all} = L &+ \sum_{n}\sum_{i,j} \parallel d_{ij}^{nn} \parallel_2^2 S^n(x_i^n, x_j^n) \\ &+ \sum_{i}\sum_{m,n} \parallel d_{ii}^{mn} \parallel_2^2 S_i(x_i^m, x_i^n), \\ &\text{s.t. } \mathbf{M}_{\text{ind}} \succeq 0, \mathbf{M}_{\text{age}} \succeq 0, \end{aligned} \tag{5.12}$$

where $\mathbf{M} \succeq 0$ implies that $\mathbf{M}$ is a semi-definite positive matrix, thus pseudometrics are allowed.

As both the Mahalanobis matrices $\mathbf{M}_{\text{age}}$ and $\mathbf{M}_{\text{ind}}$ as well as the embedding function F need to be learned in Eq. 5.12, we employ a deep network to optimize them jointly. The architecture of the proposed network is discussed in the following sections.

Deep network architecture

Our quartet-based network architecture is shown in Fig. 5.4, which jointly optimizes the manifold embedding function f and the two Mahalanobis matrices. The network takes quartet samples as input. Each quartet sample contains an image set $Q = \{I_i^m, I_i^n, I_j^m, I_j^n\}$, which are the images of the person i and j at his m and n age stage. The images are firstly passed through a weight-shared convolutional layers, which can extract prolific and robust age and identity information from a facial image while preserving the locality. The deep convolutional network takes the joint manifold cost as the loss function. Subsequently, the distance between the outputs of the deep architecture, for example, $F(I_i^m)$ and $F(I_i^n)$ are measured via two independent metrics, which are namely, age metric and individual metric. With the distances between each image pairs, the quartet loss are thus optimized and the gradients are back-propagated to update the $\mathbf{M}$.

Deep convolutional layer. In our model, the deep convolution layer is trained to explore the joint manifold of the age and identity. As discussed in the Section 3.1, the joint manifold is supposed to keep the locality structure, thus the Eq. 5.2 and Eq. 5.4 are taken as the joint manifold cost. In the experiment, we first compute the similarity matrix across the whole dataset while for each input batch, only the involved locality constrains need to be satisfied during training, which leads to a great computation saving. As a fact, the linear embedding can already reflect the joint manifold, however we employ the deep learning for a better performance.

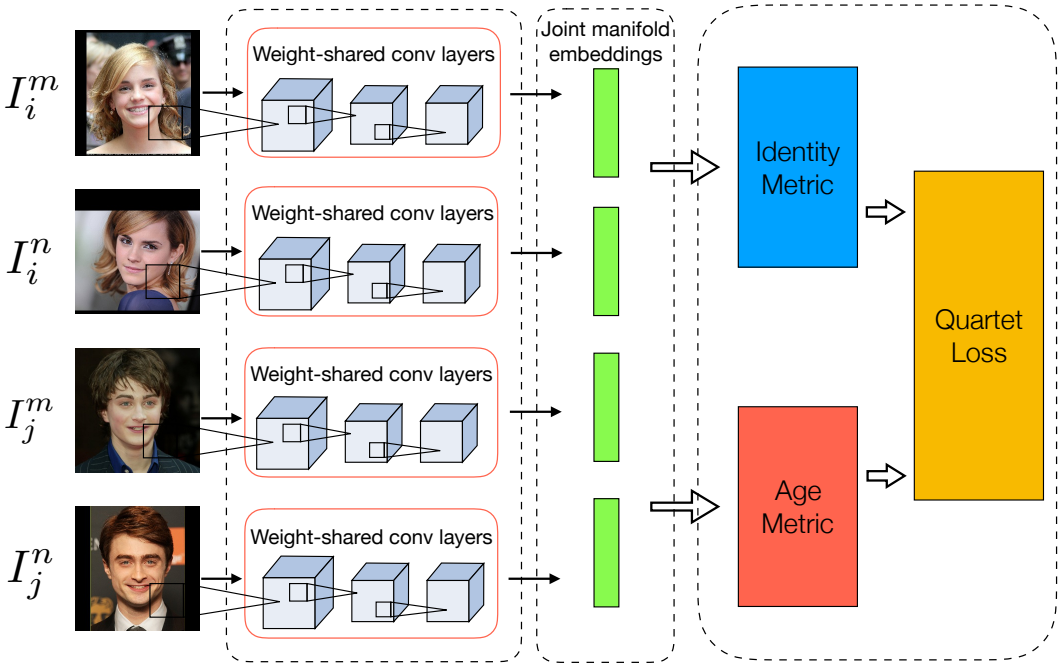


Fig. 5.4 The architecture of our proposed deep network. The network takes quartet samples as input, and the joint manifold embeddings are obtained after the images are forward propagated through four weight-shared convolutional layers. Subsequently, the distances between embedded images on the joint manifold are measured by two independent metrics – individual metric(blue) and age metric(red). Finally the distances are feed to the last layer to optimize the quartet loss.

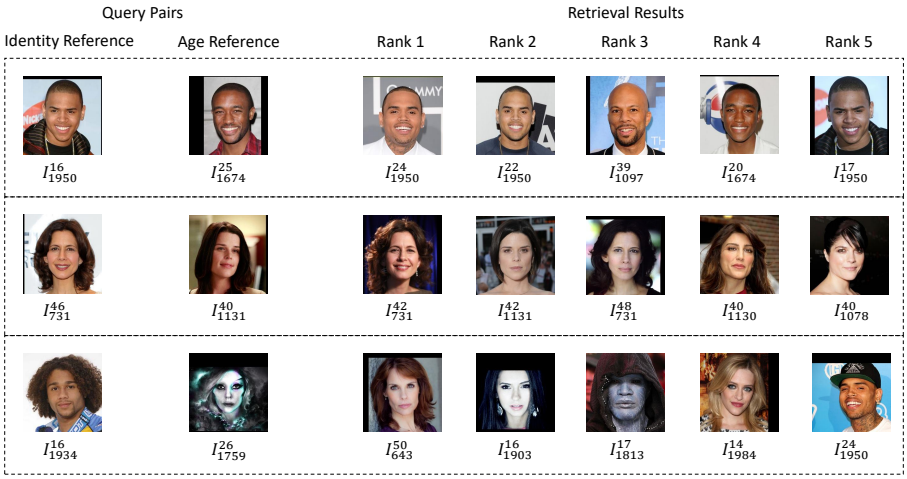


Fig. 5.5 Experimental results on CACD dataset. The first row and second row are selected two convincing retrieval results and the third row is a picked bad retrieval example. However, the failure shown here is because that the age reference image contains too much noisy and even a human cannot correctly figure out the age of the subject, thereby such noisy data influenced the similarity measurement both on the age metric and the identity metric.

Individual metric and age metric. At the end of the deep architecture module, the facial images are represented by a d-dimensional feature. To measure the distances between each image, we introduce two Mahalanobis matrices $\mathbf{M}_{\text{age}}$ and $\mathbf{M}_{\text{ind}}$. Since Mahalanbis matrices are semi-definite positive, $\mathbf{M}$ can be factorized as $\mathbf{M} = \mathbf{P}^\top\mathbf{P}$. In other words, to learn the individual metric and age metric is equally to learn two projections $\mathbf{P}_{\text{ind}}$ and $\mathbf{P}_{\text{age}}$ as:

$$\begin{aligned}\Phi(F(x_i^m), F(x_j^n)) &= d_{ij}^{mn\top}\mathbf{M}d_{ij}^{mn} \\ &= ||\mathbf{P}F(x_i^m) - \mathbf{P}F(x_j^n)||_2^2\end{aligned} \tag{5.13}$$

In our architecture, the two metrics layer are inner product layers with independent weights. The eucledean distance in the projected space is the corresponding Mahalanbis distance. It is not hard to update the matrix $\mathbf{P}$ via the loss function Eq. 5.12 while how to ensure $\mathbf{M}$ being semi-positive is a problem. Inspired by [118], we take a trick when updating on $\mathbf{P}$ happens. After $\mathbf{P}$ is updated by the network, we check all the eigenvalue of the matrix $\mathbf{P}$ and change the most negative eigenvalue to zero and then update $\mathbf{P}$ again to make it closer to a semi-positive matrix. The algorithm is shown in Alg. 1

Algorithm 1: Dual Reference Face Retrieval

Data: Training set X
Init: Compute similarity S_i and S^n
while ***NOT*** *convergence* **do**
  Construct batch $B = \{I_1, I_2, I_3, \cdots, I_b\}$.
  Forward propagation to obatin embedding $\{x_1, x_2, x_3, \cdots, x_b\}$.
  Construct quartet Q that violates Eq. 5.5 and Eq. 5.7.
  Feed Q into network to obtain loss in Eq. 5.12 and update network weights.
end

5.4 Experiment

As the dual-reference face retrieval is a newly explored task, there are few datasets where each individual's images have a wide age distribution. However, we emphasize that the scarcity of suitable datasets does not mean the task is unnecessary. On the contrary, it supports our motivation that using dual reference images to indicate multiple semantic information is reasonable when merged by the huge volume of unlabelled online images.

In the experiment, we evaluate our DRFR on three face recognition and age estimation datasets: Cross-Age Celebrity Dataset(CACD) [22], FGNet [77], and MORPH [110]. The statistics of these datasets are shown in Table. 5.1. As the CACD contains the most images among the three, we trained our deep neural network and conducted our main experiments on the CACD. Apart from that, we evaluated the robustness of our joint manifold model on FGNet and cross-dataset validate on the MORPH.

Dataset	Images	Subjects	Images/sub.	Age gap
CACD	163446	2000	81.7	0-9
FGNet	1002	82	12.2	0-45
MORPH	55134	13618	4.1	0-5

Table 5.1 Statistics of the Datasets

5.4.1 Experiment on CACD

Settings The Cross-Age Celebrity Dataset is collected for the cross age face retrieval task in [22], and it contains 163446 images from 2000 celebrities with the age ranging from 16 to 62. The large scale data with high age variations supplies the DRFR a high quality experiment environment. However, it is noteworthy that although the age ranges from 16 to 62, the maximum age gap for each celebrity is 9 years old, as all the collected images are

taken from 2003 to 2014. In details, the age gaps are stepping at 1 year old from $(14-23)$ to $(53-62)$, thus there are 40 age gaps in total. On average, each age gap contains 4000 images of 50 celebrities. Following the settings in [22], we take 60% data as training data and the remaining for the test. The training data is picked uniformly from each age gap to ensure all the age gaps are covered. For the test data, as there are averagely 8 different images for each celebrity at each age, we further split the test data into 8 subsets for the following evaluation. To train our deep network on DRFR, the weights of two Mahalanobis matrices were initialized as identity matrices. For the hyper-parameters, we set the ε in Eq. 5.11 as 5 to calculate the similarity matrix set S, and the embeddings' size on the joint manifold is set as 128. The triplet selection scheme can heavily impact the convergence speed of the network training, so does the quartet samples selection. An effective triplet selection can avoid poor training and reduce the influences caused by the mislabelled data, we employed an online quartet selection protocol which is inspired by [25]. During training, the images of an entire mini batch are firstly propagated forward to extract the embeddings with the current model, then those quartets which violate the average margin in this mini batch will be selected to train the network.

Evaluation Metrics and Comparison As DRFR can be regarded as a fine-grained retrieval, we use the top-K retrieval accuracy[136] as the evaluation metric. Since there are no works on this task in the literature before, we combined the existing face retrieval approaches with the age estimation methods to form a hierarchical framework and made the comparison. In the combined hierarchical framework, the face retrieval was first conducted regarding the first reference image as query. Subsequently, we estimated the age of the second reference image and the top 100 candidate images from the face retrieval session. Finally these 100 images are ranked according to the estimated ages. For the facial representation, we choose eigenfaces, LBP, CARC [22], which encodes the images with a set of celebrities, and the deep learning feature extracted from the FaceNet [117]. For the following age estimation, we selected support vector regression (SVR) as well as canonical-correlation analysis (CCA). For the FaceNet feature, we used the same training data as our DRFR's.

Results and conclusion We conducted DRFR and the 8 hierarchical methods on the 8 testing subsets and compute the average top-K retrieval accuracy. The results are shown in Table. 5.2. It shows that when the K is small(less than 6), our proposed DRFR outperformed the other 4 three methods. It is interesting to note that when the allowed output image increases, the accuracy of CARC+CCA is slightly higher than ours. The reason is that CACD is the original dataset which CARC designed, and in our settings, each subset only contains approximately 10 images for each subject, it is reasonable for a high accuracy if the face retrieval system can retrieve all the images of the correct identity.

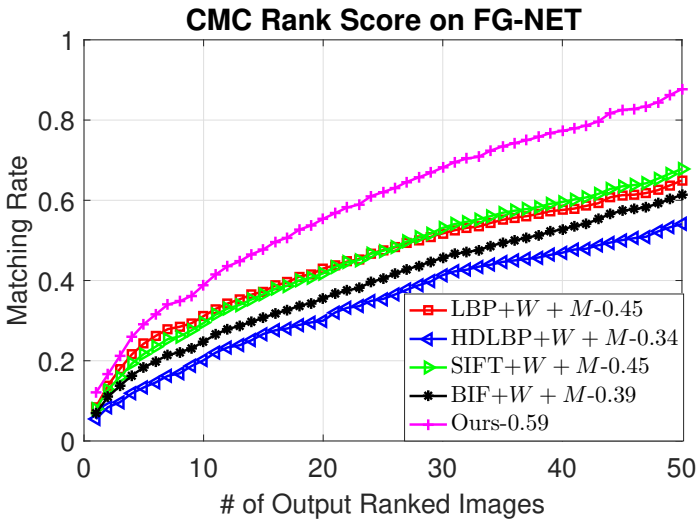


Fig. 5.6 The results of the experiment on FGNet.

Table 5.2 Experimental results on CACD dataset.

Accuracy% @ top-K	K=1	K=2	K=3	K=4	K=5	K=8	K=10
eigenfaces+SVR	14.43	17.25	17.42	17.87	18.5	19.10	19.20
eigenfaces+CCA	14.97	17.73	18.21	18.53	18.71	19.24	19.35
LBP+SVR	17.58	20.32	20.86	21.52	21.85	23.45	24.53
LBP+CCA	17.98	21.44	22.13	22.13	22.22	24.78	25.71
CARC+SVR	18.34	22.45	23.02	23.64	24.30	25.70	26.20
CARC+CCA	18.57	22.25	23.50	23.85	24.50	**26.12**	**26.42**
FaceNet+SVR	19.76	23.20	23.33	23.77	23.64	24.70	26.33
FaceNet+CCA	19.63	23.48	24.12	24.37	24.54	25.38	26.40
DRFR(Ours)	**20.67**	**23.75**	**24.33**	**24.87**	**24.90**	25.80	26.23

5.4.2 Experiment on FGNet

Dataset Setting FGNet dataset consists of 1002 images of 82 subjects in total. As it is tiny while has high age variations, we conduct experiments using different feature on it to evaluate performance of the joint manifold embedding function f of our proposed framework. Similar to the experiment setting on CACD, we split FGNet into training and test set, avoiding the situation that the same subject shows in both sets. The training set contains 60% images while the rest is left for test.

Comparison with linear embedding methods To evaluate the robustness of our joint manifold embedding, we extracted the embeddings, which is shown as green stripes in Fig. 5.4, from the model trained in above CACD experiments. And we chose four other feature descriptors, which includes: LBP, BIF, SIFT and HDLBP, to make the comparison. To get the corresponding embeddings of these hand-crafted features, we employed PCA as the embedding technique, whose projection matrix is denoted as W. Subsequently, the age and identity metrics $\mathbf{M}_{\text{age}}$ and $\mathbf{M}_{\text{ind}}$ were trained for each embeddings based on the quartet loss. And finally the retrieval was conducted on the learned metrics.

Results and conclusion Fig. 5.6 shows the results of our experiments on FGNet. It can be seen that the CMC rank score of our joint manifold outperforms others. Since the $\mathbf{M}_{\text{age}}$ and $\mathbf{M}_{\text{ind}}$ are learned with respect to each embeddings, we can draw the conclusion that: firstly, the joint manifold embedding function trained on CACD has robust generalization. Secondly, the proposed joint manifold preserves more information of the age and identity locality.

5.4.3 Cross Dataset Validation on MORPH

The MORPH dataset has 55134 images of 13618 subjects. Though both the images and subjects are in big amount, the number of images for each subject is only 4.1, which is not sufficient to compromise the quartet samples for training. Thereby instead of training a new model, we conduct a cross-validation on MORPH. We used the model trained on the CACD dataset directly on the MORPH dataset and the results are shown in Table. 5.3. It is shown that the results are very close to those on CACD. One reason of the minor backward can be the divergence of the age distribution between MORPH and CACD. Another reason is that the images in MORPH are over-cropped and some parts of the forehead and the chin in the image are absence, while the images are all of the full face in CACD.

Acc% @ top-K	K=1	K=3	K=5	K=10
MORPH	18.26	20.81	22.99	23.17
CACD	20.67	24.33	24.90	26.23

Table 5.3 Cross dataset validation on MORPH.

5.5 Summary

In this chapter, we proposed a dual-reference face retrieval framework, which tackles the problem of retrieving a person's face image at a 'given' age. In the proposed framework, the retrieval is conducted on a joint manifold and based on two similarity metrics. We have systematically evaluated our approach on CACD, FGNet and MORPH, and the corresponding results show that the proposed approach achieves promising results on this new task and the framework is stable and robust.

For the future work, a larger dataset with wider age range can be collected to further improve our algorithm. Also, the dual-reference retrieval framework can be extended to other retrieval tasks besides the face.

Chapter 6

Aging Image Synthesis

6.1 Introduction

Age synthesis has received many research interests for its importance to wide range applications, for example, finding missing people, face verification, security surveillance and entertainment. Classical methods try to find the average depict (texture, shape *etc.*) of different ages for different gender then render it to a face image to yield an age-progressed result [69]. [18, 68] generate a new facial texture with identical second order statistics with a given sample texture but not concern on the age [124]. [145, 114, 130] focus on both age group feature representation and identity preservation, which describe the difference of texture/shape/*etc.* Between different age groups with different genders and re-render it with the personal characters to realize face aging.

Nowadays, age synthesis has got great progress benefited from the success of General Adversarial Network (GAN) [137, 121, 5, 152, 76, 91, 144]. Wang*et al.* [137] proposed a recurrent face aging (RFA) framework considering the in-between evolving states between the adjacent age groups. Shu*et al.* [121] proposed a Kinship-Guided Age Progression (KinGAP) approach which can generate personalized aging images by computing average face and using the senior family members as a prior guidance. These methods need pair-wise samples (requires faces of the same person at different ages) to learn the personalize identity and need the target images in training set be labeled with the true age for optimizing model. As we all know, it is hard to collect pair-wise data or so many labeled images. To tackle the shortage of pair-wise data, Yang*et al.* [144] presented a framework to simulate aging effect without pair-wise training data, which preserves the identity feature through both pixel level loss and adversarial loss, and in order to realize the young or senior aging effect they feed a discriminator with generated image and groups of young faces and senior faces to guide the face aging. However, "young" and "senior" are abstract conceptions, different people

has different views on it, moreover, the needed "young" or "senior" faces has no uniform baseline. These methods are illustrated in Fig. 6.1(a), and synthesizing age with two given images is still a challenge in face aging area.

(a) Conventional GAN-based age synthesis framework: Age features are learned from training images with age annotations, then an identity reference image and a target age which is given as a numeral or an one-hot-vector are fed in generator.

(b) Dual-reference Age Synthesis framework: The system synthesize face images which is not only of the same identity in the identity reference image, but also at the similar age reflected in the age reference image.

Fig. 6.1 Comparison between conventional age synthesis framework and our proposed dual-reference age synthesis framework

In this work, we first investigate age synthesis and address the challenges. Then we propose a new framework as shown in Fig. 6.1(b). In the new framework, two images are taken as inputs and an image is generated which shares the same personality of the given identity reference image and in the similar age group with the age reference image. In the proposed framework, we employ a joint manifold as well as an orthogonal projection to disentangle the age and personality information.

There are three contributions of this chapter.

1. We propose a new age synthesis framework which concerns on the age synthesis. We combine the identity disentangle space projection with the age disentangle space projection as a joint manifold feature, which can present the personnel feature and age feature at the same training time.

2. The final images are generated by training a generative adversarial network and augmented with two preserving functions. So our framework can be extended to other application not only age feature learning and age synthesis.

6.2 Related Work

Aging Synthesis As far back as 2002, Lanitis*et al.* described how the effects of aging on facial appearance can be explained using learned age transformations [79]. Then traditional age synthesis focus on facial muscle structure or skin's texture changes *etc.*, or learn those features

from the average faces of different age groups for age pattern transfer. Those models are usually very complex or neglects the differences between different persons [152]. Kaur*et al.* [68] proposed face texture transfer (FaceTex) framework augmented the prior work. FaceTex suppress facial texture comprising skin texture details around facial meso-structures (*e.g.* eyes, nose and mouth) and synthesize a facial image with different facial textures while maintaining the identity of the original one. Further, Makhzani*et al.* [94] proposed an Adversarial Autoencoders (AAE) using an adversarial training procedure to learn the latent vector, which inspired most of the state-of-the-art face aging methods. [5, 76, 91] investigate age synthesis based on GAN and AAE which can generate the personalized aging images at the tender age. The state-of-the-art GANs combined GAN with AAE, which can learn the intangible character through an encoder and generate images with photo realistic. Zhang*et al.* [152] proposed a conditional adversarial auto-encoder (CAAE) and described a synthesis protype based on GAN and AAE: personalized identities are indicated by map the original face image to a latent vector via an encoder, then these identities and a corresponding numeral (age) are fed into the generator to synthesize face images. Antipov*et al.* [5] proposed an Age Conditional Generative Adversarial Network (Age-GAN) to generate identity-preserving synthetic images within required age categories, which use the Facenet to optimize latent vectors, and it can be considered as a part of CAAE. Recently, Wang*et al.* [139] proposed an identity-preserved conditional generative adversarial networks (IPCGANs) , which use an age classifier forces the generated face with the target age and use the multi-layer feature of age classifier as identity feature. Recently, Li*et al.* [87] proposed a Wavelet-domain Global and Local Consistent Age Generative Adversarial Network (WaveletGLCA-GAN) which adopt wavelet transform to depict the textual information in frequency-domain with given age labels, WaveletGLCA-GAN abstract age information from local patches of a given age face image and generate an image with the target age, but it needs forehead, eyes and mouth local patches of the target age images and consists five sub-networks which are complex. Despite of focusing on face aging synthesis,Expression Generative Adversarial Network (ExprGAN) [32] and StarGAN [26] can also be used for face aging synthesis. A style-based generator architecture [67] for generative adversarial networks, which leads to an automatically learned, unsupervised separation of high-level attributes (*e.g.* pose and identity when trained on human faces) and stochastic variation in the generated images (*e.g.* freckles,hair), and it enables intuitive, scale-specific control of the synthesis.

Generative Adversarial Networks GAN [38] and its variants [154, 7, 61, 99] have shown the impressive success in computer vision applications, especially the conditional GANs (CGANs) [108, 103]. They have been adopted for image generation [32, 154, 7, 61, 108, 103], image super-resolution [123, 83] and image translation [17, 63, 55, 26, 67]. The

classical GAN consists of two parts: discriminator and generator, which train the two parts alternately. The adversarial loss function (as Eq. 6.1) forces discriminator to try to classify the fake image and real image, and makes generator try to generate indistinguishable images. However GAN cannot decide what kind of image be generated.

$$\min_G \max_D \mathbb{E}_{p\sim p_{data}(x)} log[D(x)] + \mathbb{E}_{z\sim p(z)} log[1 - D(G(z))] \tag{6.1}$$

So CGAN introduces condition y into Eq. 6.1 which can control the generated results with the given condition, and this controllable character made CGAN as the main network for age synthesis. Loss function of CGAN is as Eq. 6.2.

$$\min_G \max_D \mathbb{E}_{p\sim p_{data}(x)} log[D(x|y)] + \mathbb{E}_{z\sim p(z)} log[1 - D(G(z|y))] \tag{6.2}$$

6.3 Method

In this section, we first describe the pipeline of our proposed method, then two main modules of the framework are discussed in Sec.3.2 and Sec.3.3 respectively. Finally, the objective functions are introduced.

6.3.1 Overview

With a given reference face image, can you imagine what did Emma Stone look like when she was young? Or a baby will look like when he/she grows up? We can tell you in Fig. 6.2.

Synthesizing face aging or rejuvenating only with reference faces is more difficult than with a numerical age, but this challenge is more reasonable because a reference image can express what kind of age effect exactly and include more age information than a number. To tackle this task, age synthesis under unsupervised manner, our framework consists of three parts: an age agent, an identity agent and a GAN. The age agent is used to learn the reference age feature, the identity agent is used to preserve the identity information and a GAN which has the fabulous ability to generate images is used to synthesize the face image. Fig. 6.3 gives a pipeline of the proposed framework.

For convenience, we define I_i^m as an image of the individual with identity i at age m. The main characteristic is that there are two images as inputs in our framework, one is the identity reference image I_i^m and the other is the age reference image I_j^n. We assume that the face image is sampled from two low dimensional manifolds where the identity and age change smoothly along respective dimensions. The identity agent and the age agent map the two input images into two low-dimension features respectively, and they are joint in the manifold

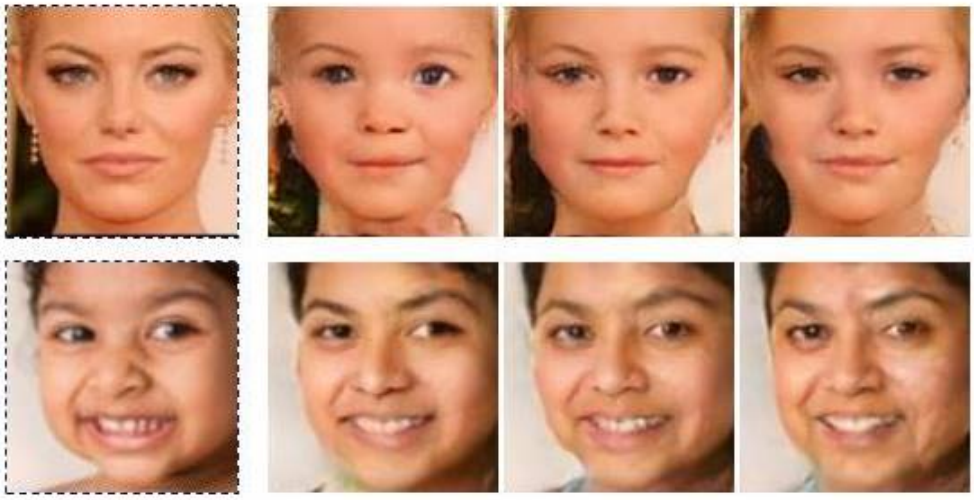


Fig. 6.2 Demonstration of our age synthesis results (images with black dotted box are the original inputs.)

space. Then the joint manifold embedding is fed to the generator and an facial image $\hat{I}_i^n$ is synthesized.

6.3.2 Identity Agent

The identity agent is inspired by [153], the same identity encoder and identity discriminator as [152] are used. Fig. 6.4 shows the detailed architecture of the identity agent.

Aim to abstract the identity feature from identity reference images without pair-wise or labeled training data, a reconstruction loss is used:

$$L_{rec} = ||I_i^m, \hat{I}_i^n||_1 \tag{6.3}$$

And an adversarial process force the identity manifold with no "holes" [152]. Denote p_{data} as the distribution of the realistic data, assume p_z as the prior identity feature distribution. The latent code is trained to approach by:

$$\min_{E_I} \max_{D_I} \mathbb{E}_{z_I \sim p_z} log[D_I(z_{I_{id}})] + \mathbb{E}_{I \sim p_{data}} log[1 - D_I(E_I(I))], \tag{6.4}$$

where $z_I = E_I(I_{id})$ denotes the identity embedding of identity reference image I_{id}.

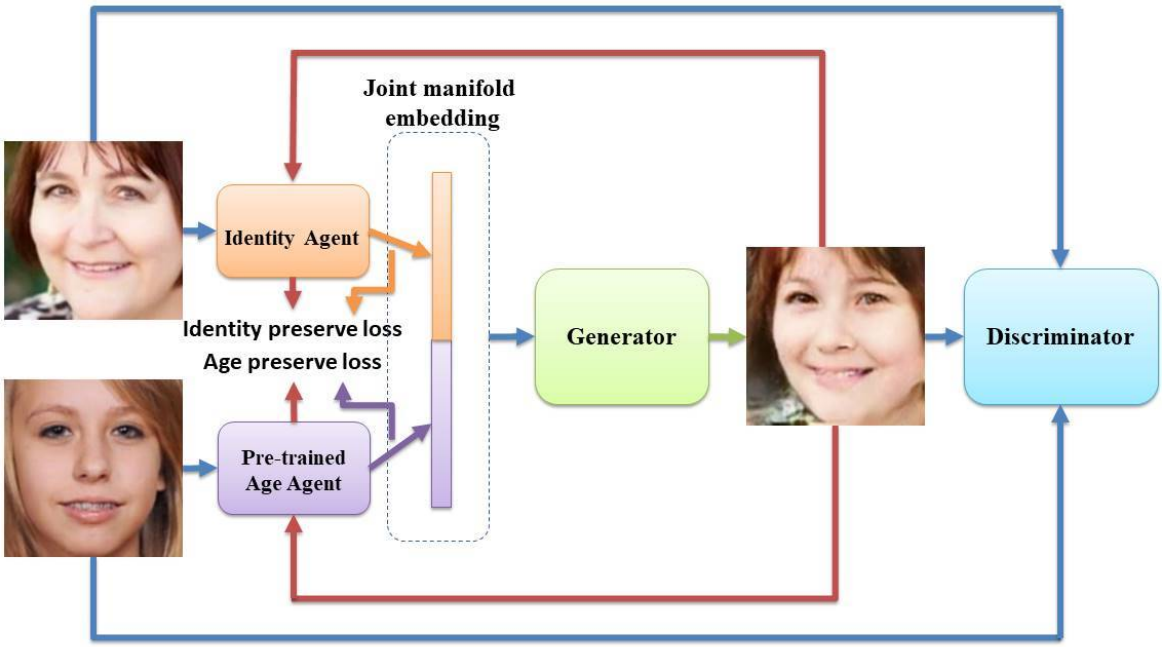


Fig. 6.3 The pipeline of the proposed age synthesis method. The identity agent learns the disentangle depictions of the identity reference image and the age agent learns the age features of the age reference image as well. The identity depictions and age features as a joint manifold embedding is fed into a generator. A discriminator tries to recognize the synthesized image and the two ground-truth inputs which guarantees the synthesized face image looks realistic, and identity preserving loss and age preserving loss guarantee the synthesized face image have the identity information of the identity reference image and the age information of the age reference image.

Furthermore, the synthesized image $\hat{I}_i^n$ should has the same identity with I_i^m, then we designed the identity preserving function as:

$$L_{ID} = ||E_I(I_i^m), E_I(\hat{I}_i^n)||_2 \tag{6.5}$$

6.3.3 Age Agent

An age agent is designed for the proposed framework based on the deep expectation of apparent age (DEX) [111, 112] (the winner of LAP challenge on apparent age estimation). A pre-trained VGG16 model as Fig. 6.5 shows.

Different from many researches, the hierarchy age embedding has more age information than a giving one-hot vector or a numerical label. Moreover, comparing with those congregate

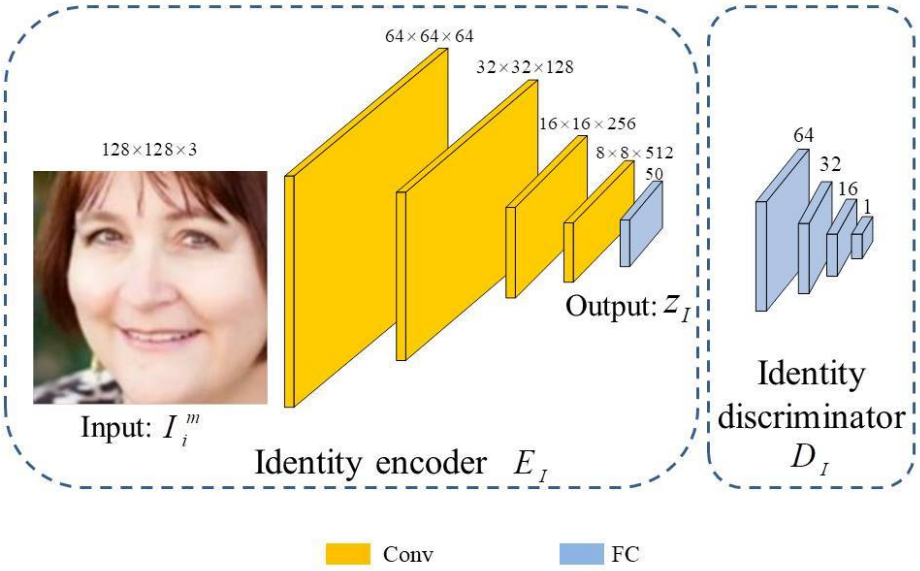


Fig. 6.4 Identity agent consists an encoder E_I and a discriminator D_I. Encoder learns to represent the latent vector z_I and discriminator force z_I to subjects to the uniform distributions.

multi-layers outputs of the VGGnet model as the final feature [144, 32], this 1024-dimension fusion feature makes our framework light-weighed.

Specially, an age preserving function L_A guarantees the synthesized image has the same age with the age reference image.

$$L_{AGE} = ||E_A(I_j^n), E_A(\hat{I}_i^n)||_2 \tag{6.6}$$

6.3.4 Objective function

To generate a photo-realistic face image, a realistic discriminator is employed here to discriminate the two reference images as real and the generated image as fake. Thus the discriminator's objective function can be derived as:

$$\min_G \max_D \mathbb{E}_{I_i^m \sim p_{data}(I)} log[D(I_i^m)] + \mathbb{E}_{I_j^n \sim p_{data}(I)} log[D(I_j^n)] + \mathbb{E}_{I_i^m, I_j^n \sim p_{data}(I)} log[1 - D(\tilde{I}_i^n)] \tag{6.7}$$

where $\hat{I}_i^n = G(E_I(I_i^m), E_A(I_j^n))$.

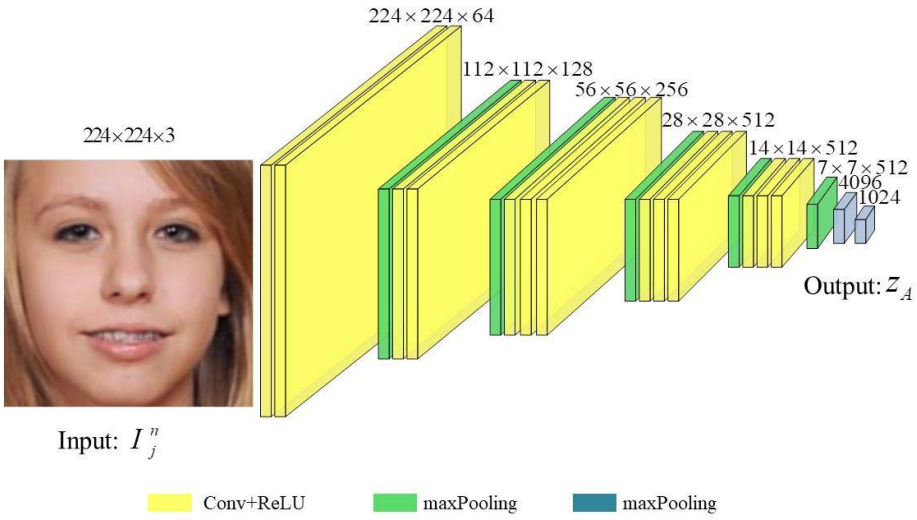


Fig. 6.5 The reference face image goes through the deep convolution network, and the age agent project the the first full-connection layer output to a 1024-dimension vector which is used as age feature.

Finally, the overall objective function is:

$$\begin{aligned}\min_{E_I,G}\max_{D_I,D} \mathscr{L}_{rec} + \mathscr{L}_{ID} + \mathscr{L}_{AGE} + \mathbb{E}_{I_i^m \sim p_{data}(I)} log[D(I_i^m)] + \\ \mathbb{E}_{I_j^n \sim p_{data}(I)} log[D(I_j^n)] + \mathbb{E}_{I_i^m, I_j^n \sim p_{data}(I)} log[1 - D(\tilde{I}_i^n)]\end{aligned} \tag{6.8}$$

6.4 Experiments

6.4.1 Data Description

We perform experiments on two widely used benchmark face datasets UTKFace [152][1] and Cross-Age Celebrity Dataset (CACD) [23][2], all images are aligned and cropped. CACD dataset covers 2,000 celebrities and the age labels were estimated by simply subtract the birth year from the year of which the photo was taken, so we choose those images with rank smaller or equal to six as the homepage said. Specially, comparing face images of UTKFace are in the wild, CACD face images are exquisite photos which have low qualities. UTKFace and CACD are all annotated with real age, we divide images into 10 age groups and made a static on them. From Fig. 6.7, we learn that only UTKFace includes baby(0-5 years old), child(6-10 years old) and those senior people above 70-year-old images. Furthermore, the

[1] https://susanqq.github.io/UTKFace/
[2] http://bcsiriuschen.github.io/CARC/

amount of young person(from 20-year-old to 40-year-old) is about twice than that of other age groups.

6.4.2 Implementation Details

In term of morphology, 0-10 years old children have the different facial appearance from the teenagers and adults. Then aiming to learn the age feature of the children and senior people and to avoid over-fitting in range of 20-40 years old group, we augmented UTKFace and CACD by flip those images not in this group. Then we use 80% images as training data, 10% as validation data and the left 10% as test data. We trained our networks on an NVIDIA TITAN X GPU using a random sample of 100 as one batch from training data.

10 images as shown in Fig. 6.6 from UTKFace was chosen as the age reference images whose appearance ages cover a range of baby to senior people and are distinct from the training images.

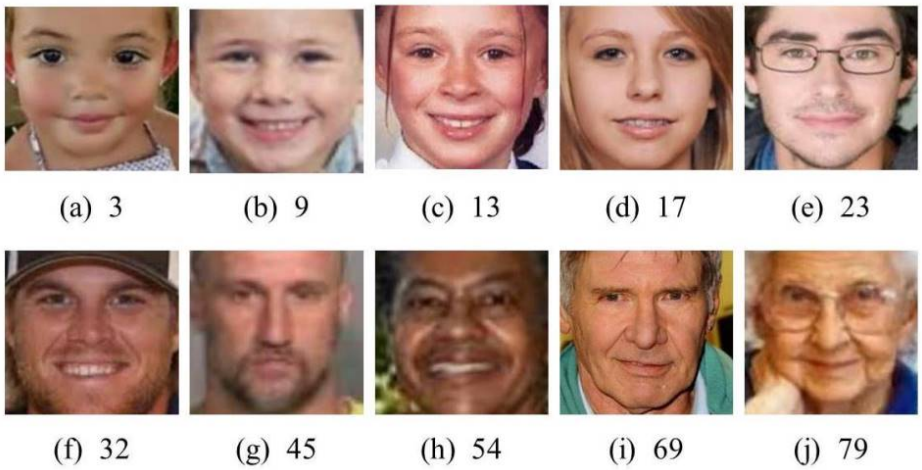


Fig. 6.6 Numbers under each reference images is real ages range from 3 to 79. Their apparent ages are different with their real age, *e.g.* men in (f),(g) and (h) are at different real ages but they look like at the same age.

The synthesized image $\hat{I}_i^n$ is supposed to have the following characteristics:

- has the same personality information with identity reference image I_i^m
- be at the same age group with the age reference image I_j^n
- be photo-realistic.

Empirically, it is hard to archive good performance if train the model with multiple loss functions in Eq.6.8 directly. To tackle this difficulty, we propose joint-training strategy for

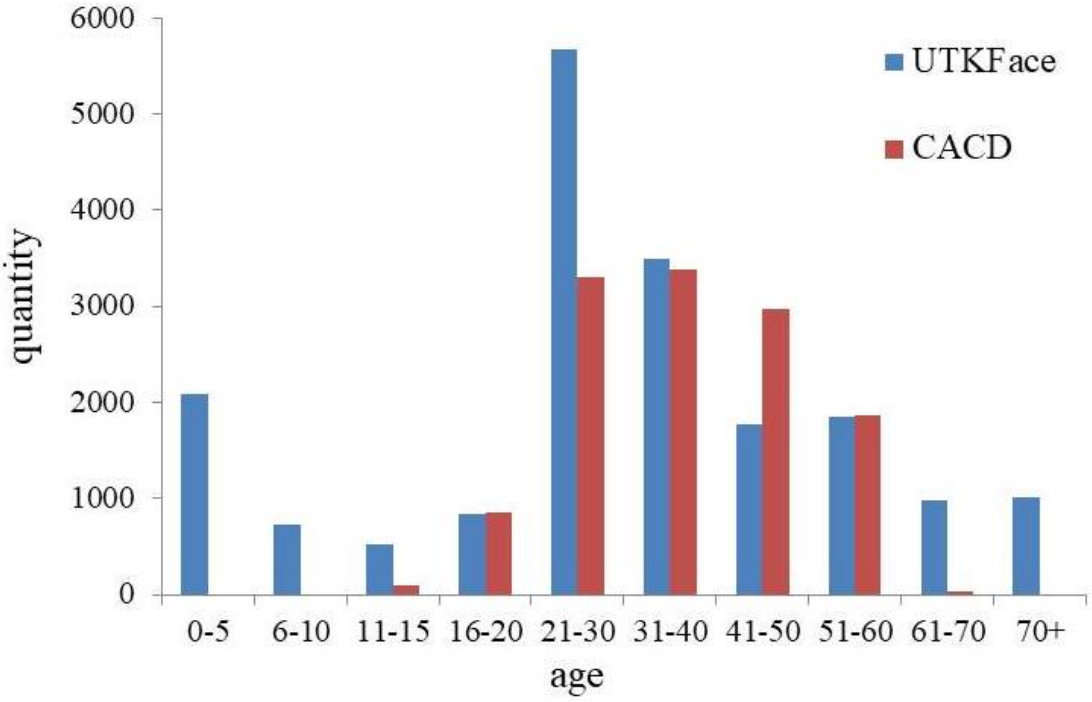


Fig. 6.7 Age distribution of the two datasets

training. At first, in order to preserve identity information and make sure that the identity manifold covers all feature space, we set $I_i^m = I_j^n$, and then train the identity agent and generator by only minimize reconstruction loss Eq. 6.3 and adversarial loss Eq. 6.4. Then we add the adversarial loss of generator and discriminator Eq. 6.7 in the training loss to guarantee the generated face be photo-realistic. Subsequently, after the identity agent loss converges, we fix E_I and D_I, set $I_i^m \neq I_j^n$, and use the two preserve functions Eq. 6.5,Eq. 6.6 to optimize the generator.

For CAAE, we retrain CAAE on UTKFace with the released code, and then fine tune the CAAE model on CACD, then we can get the 10 age groups images of UTKFace and CACD. For IPCGAN, we get the test results on UTKFace and CACD with the released pretrained model which were trained on CACD and can only synthesize 5 age groups images (11-20, 21-30, 31-40, 41-50 and 50+).

6.4.3 Experimental Performance and Analysis

Experimental Results As in most existing GAN works [153, 148], subjective evaluation on the quality of sythesized images is the mainstream, in this section, we mainly evaluate our proposed method on subjective manner. Age synthesis results are shown in Fig. 6.8

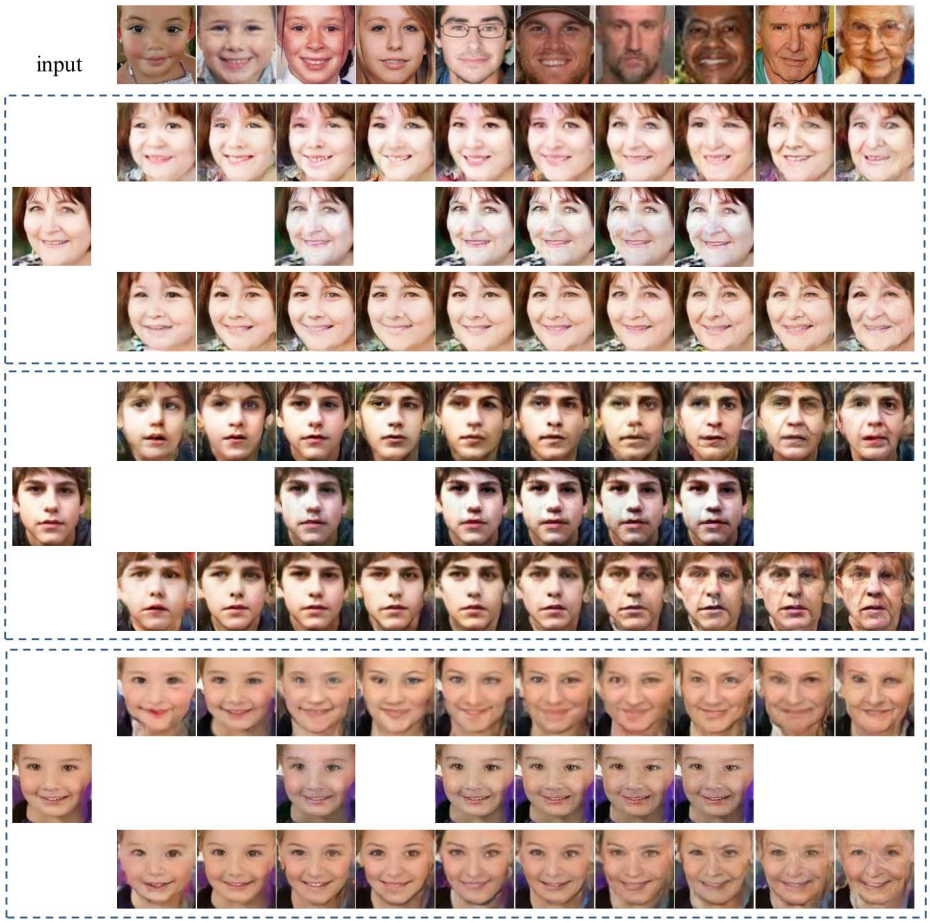


Fig. 6.8 Some synthesized faces on UTKFace. Each dotted box denotes one person's image. In each box, from top to bottom, they are images generated by CAAE, IPCGAN and DRAS.

and Fig. 6.9. The first left column is the identity reference images for DRAS (input images for CAAE and IPCGAN), and the first row is the age reference images for DRAS and the corresponding 10 age groups for the other two methods. The identity reference images include senior person, young person and toddle.

Fig. 6.8 and Fig. 6.9 show the synthesized faces in different age groups. We can see that age appearances change slightly and the generated images look the same as the input images in IPCGAN. There is one alexnet in IPCGAN, which is used to abstract identity feature and recognize age groups. However, identity preserving and age classifier sharing the same convolution layers is not reasonable. There has no identity or age preserving strategy in CAAE, and the synthesized images in CAAE look blurry and have artifacts, for example, in the first dotted box in Fig. 6.8, pupils of CAAE images changed and the mid-age one looks the same as young one. Compared with CAAE and IPCGAN, DRAS can synthesize higher quality face images with the same identity and age with the references images.

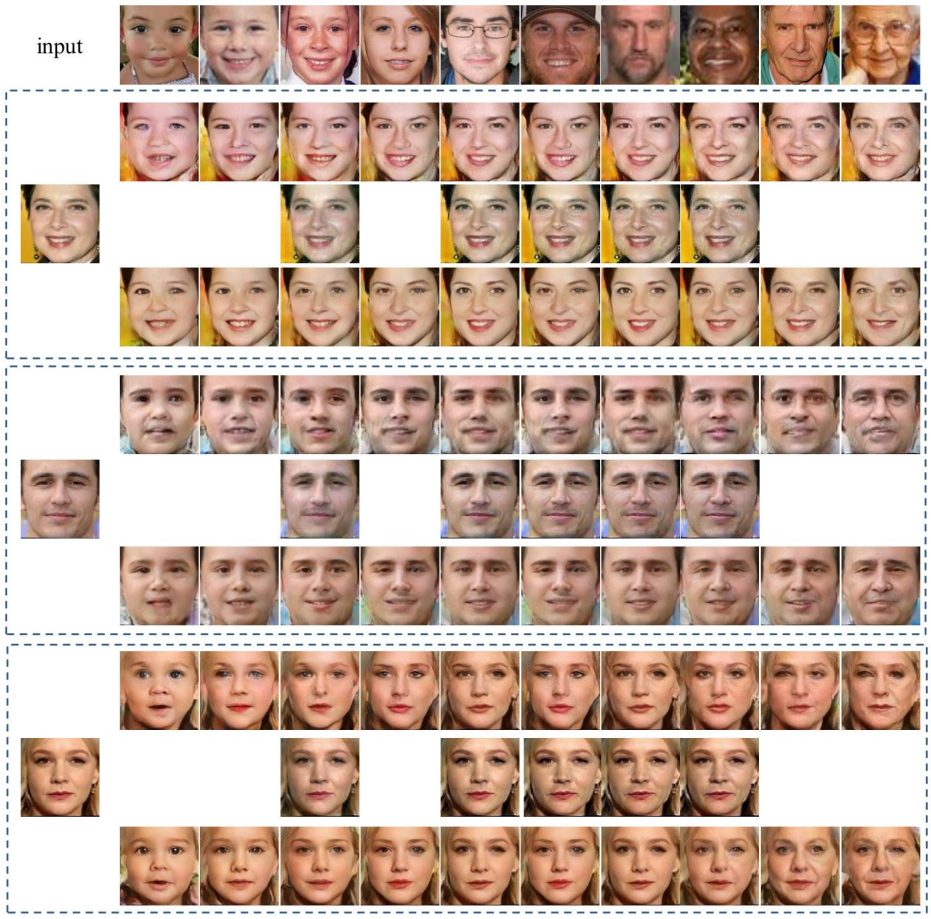


Fig. 6.9 Some synthesized faces on CACD. Each dotted box denotes one person's image. In each box, from top to bottom, they are images generated by CAAE, IPCGAN and DRAS.

Fig. 6.10 Synthesized faces of UTKFace with identity reference images and their own ages. The first row is the ground truth. In each box, from top to bottom, they are images generated by CAAE, IPCGAN and DRAS.

Furthermore, we compare our results with those ground truth. First, we use the same image as identity reference image and age reference image, and reconstruct the reference image as Fig. .6.10 and Fig. .6.11. The reconstruction images of IPCGAN have the best performance on some details, for example the illumination and curl hair on face (show in green boxes). Images of DRAS are mostly look like real images (show in blue boxes).

Then we synthesize Emma Waston and Isabella Rossellini's images with different age reference images, choose the real images from CACD, and the real images are in the same age group with age reference images. The IPCGAN can generate images which most like the identity reference images, but still has slight age apperance difference. Our proposed model can synthesize images which have the same identity feature and age feature with the reference images (as Fig. .6.12 shows).

No matter the reconstruction results or the synthesis results, CAAE has the lowest performance.

Ablation Study For analyzing the two preserving functions, we set four training scenarios as Table 6.1.

Under the four training scenarios, we trained our model and got test results as Fig. .6.13 and Fig. .6.14.Those face images with red box look younger or older than the reference age, which means the trained model without age preserving function will lose the age information. And those images with yellow box look different with the reference identity, which means the model without identity preserving function will lose the personality information in some degree. Moreover, by checking those images synthesized under scenario S2 and S3, images

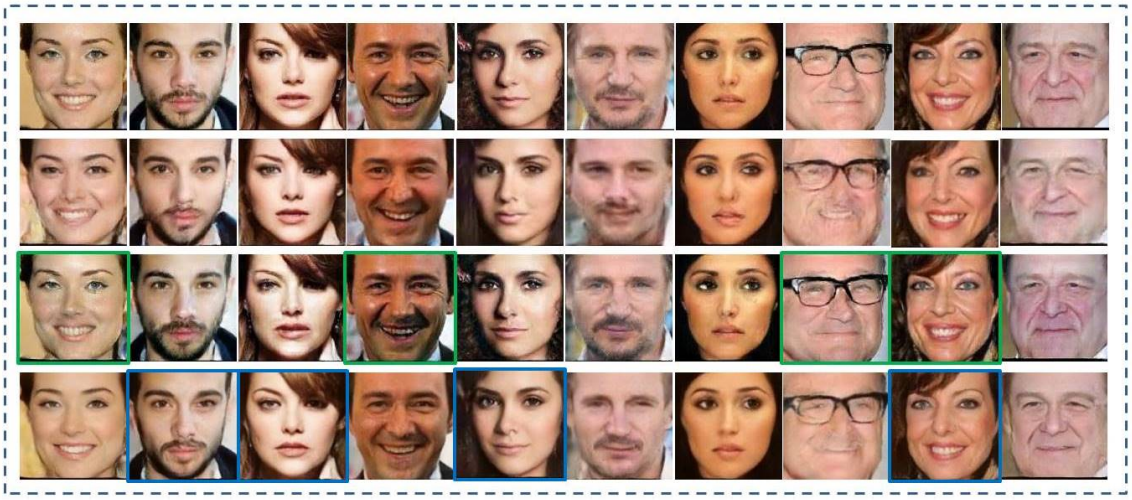


Fig. 6.11 Synthesized faces of CACD with identity reference images and their own ages. The first row is the ground truth. In each box, from top to bottom, they are images generated by CAAE, IPCGAN and DRAS.

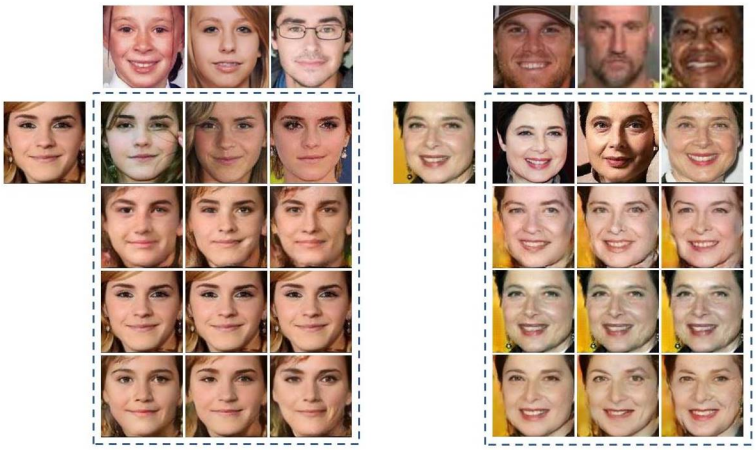


Fig. 6.12 Synthesized faces with different identity reference images and age reference images. The first row is the age reference images and the first left column is the identity reference images. In each box, from top to bottom, they are real images in the same age with age reference images, and images generated by CAAE, IPCGAN and DRAS.

Table 6.1 Description of four training scenarios.

Scenario	Description
S1	with the two preserving functions
S2	without the two preserving functions
S3	with only identity preserving function
S4	with only age preserving function

with age preserving function have artifacts but those S2 images have not. Age information usually appears in local facial parts, such as wrinkles at the eye corners, the width between two eyes is large and the face shape is round like an apple for baby, therefore age feature consists texture features, shape features *etc.* which can be seen as artifact by human eyes.

Fig. 6.13 The effective of identity and age preserving functions. There are face images generated under S1, S2, S3 and S4 respectively from top to bottom.

6.5 Summary

This chapter studied a new age synthesis task: dual-reference age synthesis. By given two reference face images, synthesize a face image which has the same identity information with one image and at the age of the other one, what is a challenge. In the proposed framework, a joint manifold embedding is abstracted from training data through the identity agent and age agent, then the proposed model is trained on adversarial way with the joint manifold embedding, and is optimized with identity preserving function and age preserving function. Detailed training strategy was discussed in this chapter, the experimental results on UTKFace

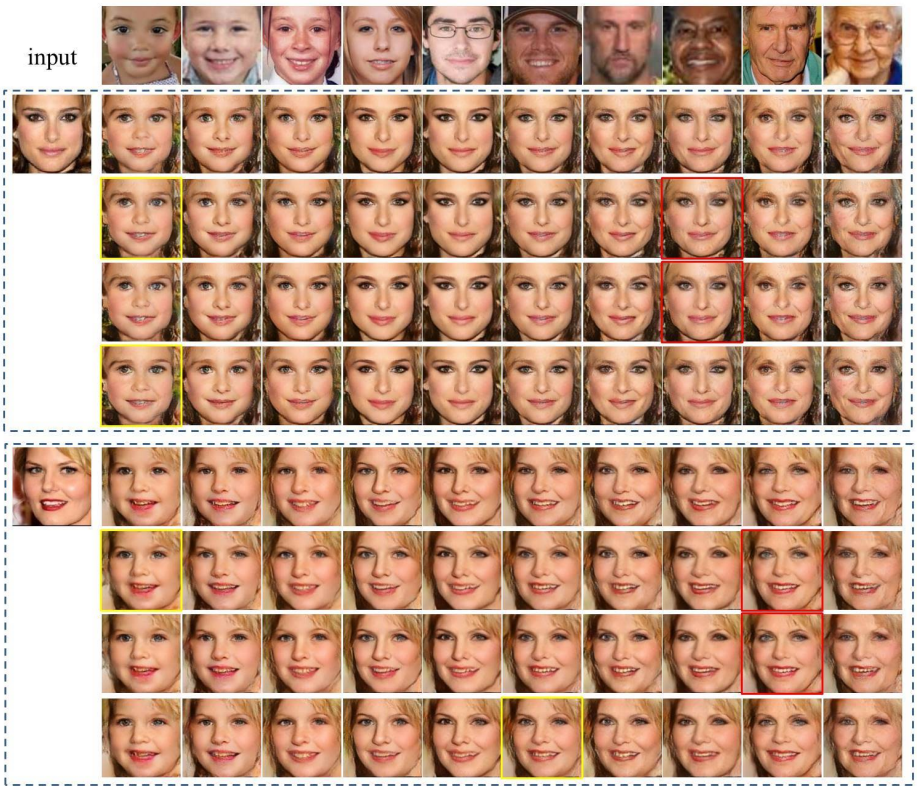


Fig. 6.14 The effective of identity and age preserving functions. There are face images generated under S1, S2, S3 and S4 respectively from top to bottom.

and CACD are given and are analyzed thoroughly. The corresponding results show that the proposed approach achieves promising results on this new task.

Chapter 7

Gait Recognition Based on Generative Adversarial Networks

7.1 Introduction

Gait is a behavioural biometric characteristic which can be used for remote human identification. Compared with recognition technologies based on other biometric characteristics like fingerprint or iris, gait recognition can be applied at a much larger (longer) distance without subjects' cooperation. Nowadays, surveillance cameras are widely installed in public places such as airports, government buildings, streets and shopping malls, which makes gait recognition a useful tool for crime prevention and law enforcement. Gait analysis has contributed to evidence for convictions in criminal cases in some countries like Denmark [80] and UK [16].

For automated gait recognition, there are two main approaches: model-based, and appearance-based. Model-based methods aim to model the human body structure parameters, while appearance-based approaches extract gait features directly from gait sequences regardless of the underlying body structure. This work falls in the latter category, which can also work well on low-quality gait videos, when the body structure parameters are difficult to extract precisely.

The average silhouette over one gait cycle, known as Gait Energy Image (GEI, as shown in Fig. 7.1) is widely used in recent appearance-based gait recognition systems because of its simplicity and effectiveness [51]. In [64], several gait templates were evaluated on a gait dataset consisting of more than 3000 subjects and it was found that directly matching GEI can yield very good performance, when gaits from the probe (*i.e.*, query gait) and gallery (*i.e.*, reference gait) are in the same walking conditions. Yet in real-world scenarios, there exist

covariate factors such as shoe type, carrying condition, clothing, speed, or camera viewpoint, which may affect the recognition performance significantly. Various machine learning algorithms were proposed [40, 97, 3] to learn gait features that are robust to covariates which only partially change the gait appearance. Large camera angle, however, was considered the most challenging factor— which may affect the gait features in a global manner.

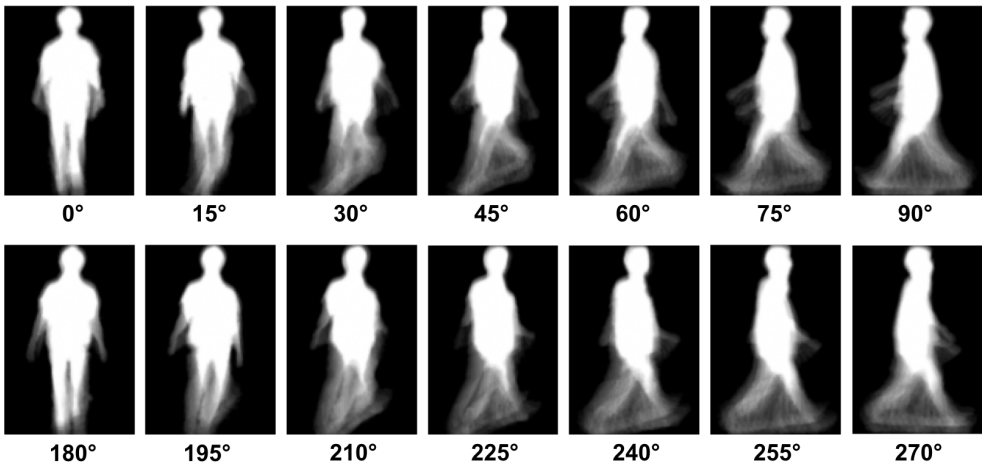


Fig. 7.1 Gait Energy Images (GEIs) in OU-MVLP dataset.

Fig. 7.1 demonstrates several GEI samples from the OU-MVLP dataset [128] from different view angles, and we can see that view changes may substantially alter the visual features of gaits, causing recognition difficulties. Recently, various deep learning approaches (*e.g.*[120, 142]) were applied to learn view-invariant features, which show superb performance on datasets with a small number of subjects (*e.g.* on CASIA-B [149]). However, the performance of these deep approaches do not scale well to large datasets, *e.g.* on the OU-MVLP dataset [128] with more than 10000 subjects. Moreover, the black-box nature of these deep CNNs makes it hard for real-world applications, *e.g.* when evidence is required. To address this issue, in this work we proposed a Discriminant Gait Generative Adversarial Network (DiGGAN) framework, which not only can scale well for large-scale cross-view gait identification tasks, but also can generate all the possible views for evidence. Our contribution can be summarised as follows:

- Algorithm: A generative adversarial network based model (named DiGGAN) is proposed in this paper. With the mechanisms of two independent discriminators, the proposed network can generate GEIs at unseen views while preserving the identity information. Besides, a triplet loss is handily introduced in our framework to enhance the discriminability of the feature learned.

- Application: Large-scale cross-view gait identification is challenging, and our proposed DiGGAN effectively solves the issue. Moreover, it can generate the all-view evidence, which is important for forensic applications.
- Performance: On the world's largest OU-MVLP dataset (with more than 10000 subjects), our method outperforms other algorithms significantly on many real-world gait identification scenarios (*e.g.* cooperative/uncooperative mode). It also has the best results on the popular CASIA-B dataset and shows strong generalisation ability across datasets.

7.2 Related Work

7.2.1 Cross-view Gait Recognition

Cross-view gait recognition methods can be roughly divided into three categories. Cross-view gait recognition methods can be roughly divided into three categories. The first category, for example, [6] is based on reconstructing 3D gait model from multiple calibrate cameras. These branch of methods have very obvious drawbacks — they rely on multiple cameras which are fully controlled and working cooperatively. Such requirements are very challenging to satisfied in real-world applications.

The second category typically achieve cross-view gait recognition by performing view normalization. For example, [37] first estimates the poses of lower limbs and then extracts the rectified angular measurements as well as trunk spatial displacements as features for gait recognition. However, such method is not always feasible especially when the lower limbs are not clearly visible hence the poses are difficult to be estimated. To tackle this problem, [74] proposed a view normalization framework based on domain Transformation obtained through Invariant Low-rank Textures (TILT), where the gait images are normalised to the side view without knowing the prior pose of the gait. Nevertheless, the performances of such method is not promising when the gait images are captured in front and back view as there is a large view angle gap with the side view.

The third category is to learn a common space where the gait images from different view angles are mapped into a same feature space and then a metric is learnt to measure the similarities then perform the matching. For instance, [95] introduced the SVD-based View Transformation Model (VTM) to project gait features from one view into another. This method is improved by Kusakunniran et al. by using Truncated SVD (TSVD) [71] to avoid oversizing and overfitting of VTM. Instead of using the global features (e.g.,[95], local

Region of Interest (ROI) was selected based on local motion relationship to build VTMs through Support Vector Regression (SVR).

There are also some variations in the third category. For example, Bashir et al. [10] used Canonical Correlation Analysis (CCA) to project gaits from two different views into two subspaces with maximal correlation. The correlation strength was employed as the similarity measure for identification. In [73], after claiming there may exist some weakly or non-correlated information on the global gaits across views [10], motion co-clustering was carried out to partition the global gaits into multiple groups of gait segments.

Most recently, deep learning approaches [120, 142, 148], [53] were applied for gait recognition, which can model the non-linear relationship between different views. In [120], the basic CNN framework, namely GEINet was applied on a large gait dataset, and the experimental results suggested its effectiveness when the view angle changes between probe and gallery are small. To combat large view changes, a number of CNN structures were studied in [142] on the CASIA-B dataset (with 11 views from 0 to 180), and Siamese-like structures were found to yield the highest accuracies. However, this dataset only includes 124 subjects, and the most recent work [128] found these CNN structures do not generalise well to a large number of subjects. In [148] and [53], GAN approaches are applied to generate gait features/images to a common view or a target view for matching. However, the generative nature of both GAN models limit the recognition accuracies, although they are more interpretable than the discriminant CNN-based approaches [53].

7.2.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs)[39] introduces a novel self-upgrading system. By keeping a balanced competition between a generator and a discriminator, fake data can be synthesised. While early work focuses on preventing low-quality, instability and model collapse problems, *e.g.* WGANs[8, 41] and DCGANs[109], recent applications utilise various supervision to control the generated data. Conditional GANs [100] can generate samples according to provided label information. The assumption is that the data is generated by interpolating conditional variations along a low-dimensional manifold. By modifying different manifold assumption, GANs have been successfully applied to interpolating facial poses, ages. GaitGAN[148] and MGANs[53] are close related work that uses GANs for gait recognition. However, compared with their methods, our method can 1) extract more discriminant view-invariant features, which is robust for large cross-view gait recognition tasks and 2) generate GEI images at unseen view angles, which can be used as important evidence for forensic applications.

7.3 Method

In this section, we describe the framework of the proposed DiGGAN and discuss the details of each component respectively. For a convenient discussion, in the rest of the paper, we use x_i^k to denote the GEI image of the i^{th} subject captured at angle k, thus $i \in \{1,2,...,N_s\}$ and $k \in \{1,2,...,N_v\}$, where N_s is the number of subjects and N_v is the number of the views in the dataset.

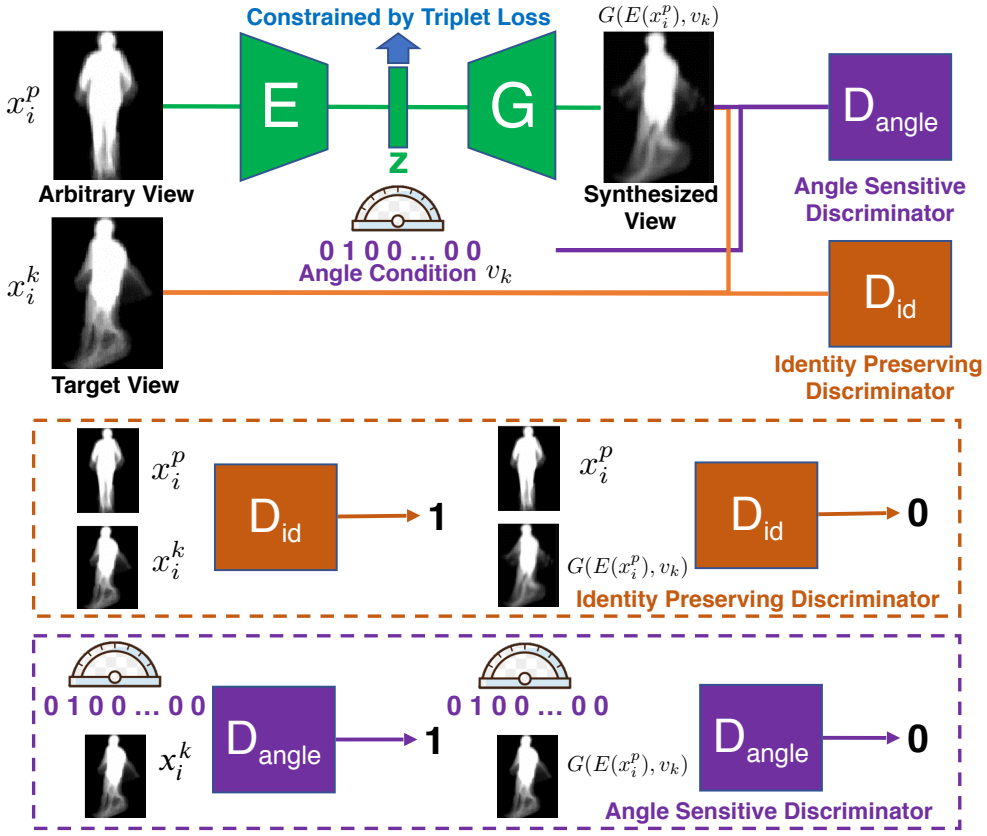


Fig. 7.2 The illustration of the proposed DiGGAN.

7.3.1 Framework Overview

Fig. 7.2 illustrates the pipeline of the proposed DiGGAN. The network is trained to transfer a GEI image x_i^p at an arbitrary view p to GEI image x_i^k with the target view k. As the input GEI image x_i^p and the target GEI image x_i^k are supposed to share the same identity, an auto-encoder E is first applied on x_i^p to disentangle the view angle information and the identity information thus project the images to an identity preserving latent space and yield the latent code $z = E(x_i^p)$. To involve the target view information k, the latent code z is concatenated

with a one-hot vector label $v \in \{0,1\}^{N_v}$, followed by a generator that takes the concatenated vector as input and generate the image $\hat{x}_i^k = G(E(x_i^p, v))$. Finally, two discriminators on angle and identity are employed to impose the angle and identity information. Additionally, to enhance the discriminability of the embeddings in latent space, a triplet loss is introduced to constrain z.

7.3.2 Angle Sensitive Discriminator

We assume that the GEI image is sampled from a low dimensional manifold where the identity and angle change smoothly along respective dimensions. As the latent code z is constrained to contain the identity information only, we can easily manipulate the angle of the generated image by concatenating different angle labels to z. Thus it is intuitive to employ a conditional discriminator D_{id} to ensure the view angle of the generated image. Mathematically, for a given training pair (x_i^p, x_i^k), the angle sensitive discriminator can be trained by:

$$\min_{E,G} \max_{D_{angle}} \mathbb{E}_{x_i^k, v_k \sim p_{data}}[\log D_{angle}(x_i^k, v_k)] + \mathbb{E}_{x_i^p, v_k \sim p_{data}}[\log(1 - D_{angle}(G(E(x_i^p), v_k), v_k)]. \tag{7.1}$$

It is worth noting that in D_{angle}, the one-hot vector label is concatenated after the first convolutional layer to obtain a better performance according to [107].

7.3.3 Identity Preserving Discriminator

One of the drawbacks in original GANs is the poor diversity in generated samples, for example, the model tends to remember samples in the training set hence outputs averaged images without differentiating identities. To tackle this problem, we introduce an identity preserving discriminator D_{id} in our framework. Inheriting the similar idea of D_{angle}, the D_{id} is designed as a conditional discriminator which takes two images as input and is expected to predict 1 if two inputs share a same identity and 0 otherwise. Thus the objective function can be derived as:

$$\min_{E,G} \max_{D_{id}} \mathbb{E}_{x_i^p, x_i^k \sim p_{data}}[\log D_{id}(x_i^p, x_i^k)] + \mathbb{E}_{x_i^p, x_i^k \sim p_{data}}[\log(1 - D_{id}(x_i^p, G(E(x_i^p), v_k)))]. \tag{7.2}$$

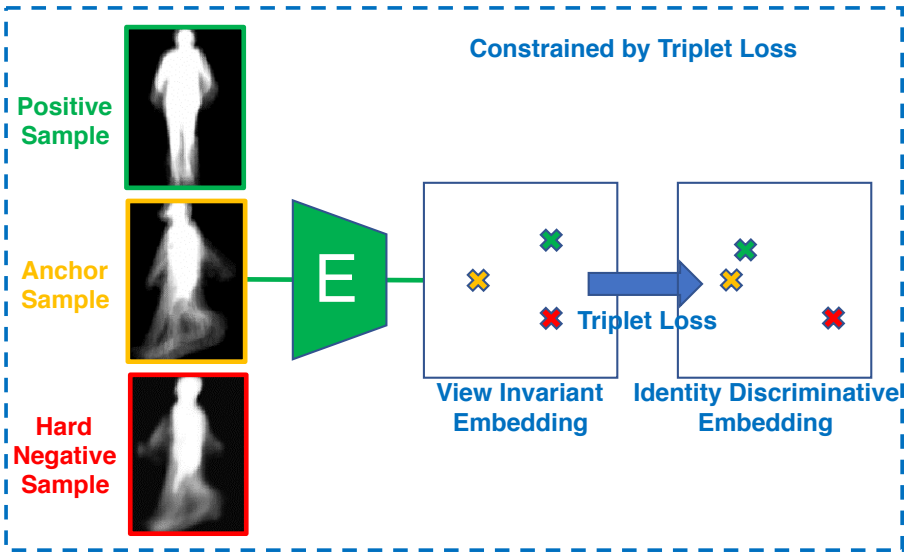


Fig. 7.3 The illustration of the triplet loss employed in DiGGAN. The triplet loss is introduced to push the negative samples away from the anchor samples while pulling the positive samples closer.

7.3.4 Triplet Constraints on z

Although the generated images can be directly used for gait recognition, *e.g.* direct matching on the pixels [149], searching on the latent space has been widely adopted by most of existing works [142] [53] for its higher performance and efficiency. However, the identity preserving discriminator does not directly constrains the latent code, which may result in the distribution of the latent code z exhibiting a 'hole'. Inspired by [153], which employs an extra discriminator to impose a uniform distribution on z, we introduce the triplet loss to enhance the discriminability of z. Concretely, as shown in Fig. 7.3, a triplet sample consists of an anchor, a positive and a negative sample, where the positive sample shares the same identity with the anchor while the negative has a different one. A hinge loss is employed here to push the negative sample away from the anchor, and at the same time, to pull the positive closer. For example, in triplet (x_i^p, x_i^k, x_j^q), the objective function below is to be minimized:

$$\mathscr{L}_{triplet} = \max(d(E(x_i^p), E(x_i^k)) - d(E(x_i^p), E(x_j^q)) + \delta, 0), \tag{7.3}$$

where $d(\cdot,\cdot)$ can be $\mathscr{L}_2$ distance and the δ is the margin to be ensured.

7.3.5 Objective Function and Training Strategies

Reconstruction loss Besides the adversarial loss, the pixel-wise reconstruction loss is also introduced to enhance the sharpness of the generated image:

$$\mathscr{L}_{rec} = ||G(E(x_i^p), v_k), x_i^k)||_1 \tag{7.4}$$

Overall objective function Based on Eq. 7.1 to 7.4, we can define the overall objective function as follows:

$$\begin{aligned}\min_{E,G}\max_{D_{id},D_{angle}} \quad & \mathscr{L}_{triplet} + \mathscr{L}_{rec} + \\ & \mathbb{E}_{x_i^k, v_k \sim p_{data}}[\log D_{angle}(x_i^k, v_k)] + \\ & \mathbb{E}_{x_i^p, v_k \sim p_{data}}[\log(1 - D_{angle}(G(E(x_i^p), v_k), v_k)] + \\ & \mathbb{E}_{x_i^p, x_i^k \sim p_{data}}[\log D_{id}(x_i^p, x_i^k)] + \\ & \mathbb{E}_{x_i^p, x_i^k \sim p_{data}}[\log(1 - D_{id}(x_i^p, G(E(x_i^p), v_k)))].\end{aligned} \tag{7.5}$$

Training strategies Empirically, training such a model with multiple loss functions in Eq. 7.5 is challenging thus always leads to poor results. To tackle this difficulty, we propose a step-by-step strategy for training. In the frist step, we only train the angle sensitive discriminator with artificial batches that are generated from the realistic GEI images. Specifically, we randomly sample n GEI images from the training set to form a batch and train the D_{angle}. In each batch, half of the images are assigned with wrong angle labels while the rest are assigned with the correct ones. Training with realistic images rather than generated ones helps the angle sensitive discriminator to converge quickly. After the D_{angle} converges, we subsequently train the network without the triplet loss in two sub stages. In the first sub stage, we set $x_i^k = x_i^p$, which means a same image is fed into the network as the (input, ground truth) pair, therefore enables the network to learn to recover the input image first. Then in the second sub stage, we feed different images to teach the model to generate images with different angles. Finally, we take the triplet loss in and fine tune the whole network. Fig. 7.4 shows the generated images at different stages of the training process. The model learns to generate averaged images at the initial stage. After that, with different images being fed into the network, the model learns to generate images of new angles. Finally the model learns to generate images with more details of the identity information from the triplet loss.

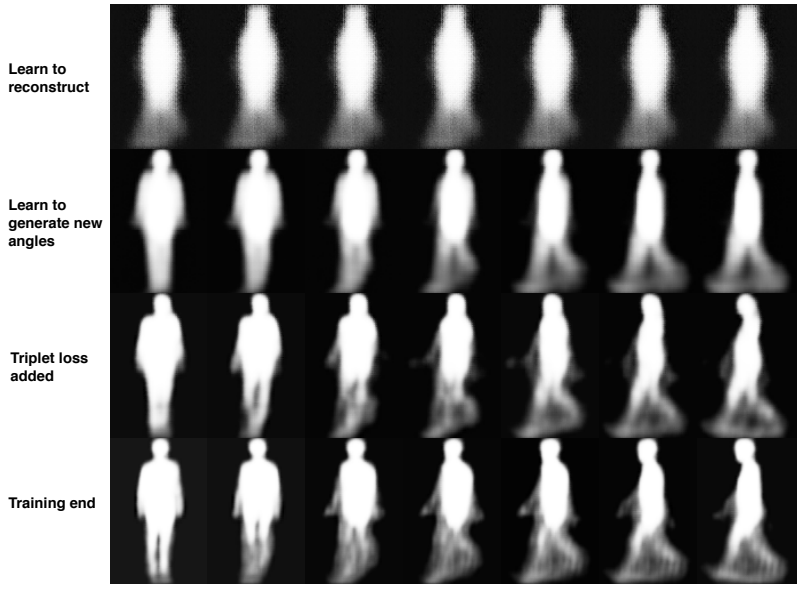


Fig. 7.4 The generated images at different stages of the training process. First row: initial stage of the model. The model outputs averaged image. Second row: the model learns to generate images with new angles from (x_i^p, x_i^k). Third row: after adding triplet loss into training, the model learns more identity details. Last row: model converges.

7.4 Experiments

In this section, we systematically evaluated our method on two datasets, the OU-ISIR Gait Database, Multi-View Large Population Dataset (OU-MVLP) [128] and CASIA-B [155]. It is worth noting that the OU-LP [64] and USF [115] datasets are not used in this paper due to lack of large view changes.

To evaluate the performance of our proposed method, we mainly focus on the cross-view identification under the *cooperative setting* [128], where the gallery has a uniform camera view angle. We also studied the *uncooperative setting* [128], where the gallery contains unknown views and following [128], we randomly select one out of all the view angles for each test subject in gallery. Moreover, we explored the effect of the triplet-loss to the performance of our framework. Specifically, we also demonstrated the generated gait images for unseen views, which may serve as important evidence for forensic application. To the best of our knowledge, this is the first work that is flexible (any-to-any view generation) at such a fine level.

In the following, we will in turn introduce each of them.

Datasets OU-MVLP is the world's largest cross-view gait dataset [128]. It contains 10,307 subjects (5,114 males and 5,193 females with various ages, ranging from 2 to 87 years) and 14 different view angles 0°, 15°, 30°, 45°, 60°, 75°, 90°, 180°, 195°, 210°, 225°,

Table 7.1 Rank 1 identification rate (%) for all baselines in cooperative setting on OU-MVLP datase.

(a) VTM [95]

	Probe				
Gallery	0	30	60	90	Mean
0	68.8	0.5	0.2	0.1	17.4
30	0.7	82.2	2.1	0.8	21.4
60	0.3	3.2	77.6	5.4	21.6
90	0.2	1.1	4.2	80.9	21.6
Mean	17.5	21.7	21.0	21.8	20.5

(b) GEINet [120]

	Probe				
Gallery	0	30	60	90	Mean
0	75.9	32.1	7.0	7.4	30.6
30	17.3	89.6	43.7	22.7	43.3
60	4.0	43.4	86.5	55.4	47.3
90	3.4	21.5	50.2	90.7	41.5
Mean	25.2	46.6	46.8	44.0	40.7

(c)Siamese [150]

	Probe				
Gallery	0	30	60	90	Mean
0	52.7	23.7	11.1	11.3	24.7
30	18.4	78.6	32.6	27.6	39.3
60	8.0	33.5	76.1	39.6	39.3
90	7.9	26.5	36.5	82.1	38.2
Mean	21.8	40.6	39.1	40.1	35.4

(d)CNN-MT [142]

	Probe				
Gallery	0	30	60	90	Mean
0	70.7	16.7	4.4	3.9	23.9
30	14.1	88.1	36.9	17.0	39.0
60	4.0	39.2	85.7	44.2	43.3
90	3.2	16.2	43.4	89.3	38.0
Mean	23.0	40.0	42.6	38.6	36.1

(e)CNN-LB [142]

	Probe				
Gallery	0	30	60	90	Mean
0	74.4	16.5	3.5	2.8	24.3
30	13.6	89.3	36.0	16.2	38.8
60	2.9	36.2	88.4	44.7	43.0
90	2.2	14.0	41.2	**91.7**	37.3
Mean	23.3	39.0	42.3	38.9	35.9

(f)DM [149]

	Probe				
Gallery	0	30	60	90	Mean
0	68.8	0.8	0.1	0.0	17.4
30	1.2	82.2	1.4	0.3	21.3
60	0.1	1.1	77.5	5.6	21.1
90	0.0	0.2	4.1	80.9	21.3
Mean	17.5	21.1	20.8	21.7	20.3

(g)MGANs [53]

	Probe				
Gallery	0	30	60	90	Mean
0	72.0	9.6	6.8	2.4	22.7
30	9.4	83.2	30.3	10.7	33.4
60	5.3	30.6	80.3	21.0	34.3
90	2.1	12.0	22.0	85.9	30.5
Mean	22.2	33.8	34.8	30.0	30.2

(h)DiGGAN(Ours)

	Probe				
Gallery	0	30	60	90	Mean
0	**79.0**	**62.1**	**46.5**	**47.7**	**58.8**
30	**58.1**	**89.8**	**64.8**	**58.5**	**67.8**
60	**44.1**	**66.0**	**88.7**	**67.2**	**66.5**
90	**44.6**	**58.9**	**66.0**	90.0	**64.8**
Mean	**56.4**	**69.2**	**66.5**	**65.8**	**64.5**

240°, 255° and 270°. The subjects repeat forward and backward walking twice of each, such that two sequences are generated in each view. The wearing conditions of subjects are various due to the collection process ranging different seasons. The size-normalized GEIs used in this paper are 88×128 pixels. Some examples from OU-MVLP dataset are illustrated in Fig. 7.1. CASIA-B is another widely used cross-view gait dataset that consists of 124 subjects with 11 different view angles range from 0° to 180° with an interval of 18° [155]. For each subject, there are six sequences of normal walking, two sequences with bags and two sequences with different clothes.

Settings For the experiments in OU-MVLP, we follow the settings in [128]. The 10,307 subjects in OU-MVLP dataset are split into two disjoint groups —- 5153 subjects for training

our DiGGAN model and 5154 for testing (*i.e.*, probe and gallery). Similarly, for the CASIA-B dataset, we choose the first 62 subjects for training and the rest 62 subjects for testing.
Technical Details: Due to the page limitation, the details of our network architecture as well as the implementation code can be found at our Github[1] repository after the review. For the parameters, the dimension of the latent code z is set as 512 for OU-MVLP and 128 for CASIA-B; and the δ in Eq. 7.3 is set as 0.2 for all the experiments.
Performance Measurement: Rank-1 identification rate (*i.e.*, recognition accuracy) is used as the evaluation metric. Features are extracted from the trained DiGGAN, before nearest neighbour classifier can be applied for different cross-view gait recognition tasks.

7.4.1 Experimental Results on Cooperative Setting

Experimental Results on OU-MVLP Since two GEIs with 180° view difference are mostly considered as those from the same-view pair based on perspective projection assumption [96], we focus on four typical view angels (0°, 30°, 60°, 90°) in this section. We compared our DiGGAN framework with some state-of-the-art baselines, including classical ones: direct matching (DM)[149], VTM[95], CNN-based methods: GEINet [120], Siamese[150], CNN-MT[142], CNN-LB[142], and the most recent GAN-based approach: MGANs[53]. In the cooperative mode, the rank 1 identification rates of all four view angles are reported in Table 7.1, from where we can see:

- Our method outperforms other methods significantly on cross-view gait identification tasks. Our overall rank-1 accuracy is 64.5%, and that is 23.8% higher than the second best GEINet.
- Our method is more robust on cross-view gait identification. In this cooperative mode, although accuracy decreases w.r.t. increasing view angles differences, they are less significant when compared with other algorithms. Our DiGGAN can yield very competitive performances even when the view difference is 90°, which indicate our method can extract robust view-invariant features.
- Most of the methods suffered from gallery in view 0°, yet our DiGGAN can achieve a reasonable accuracy of 58.8%, much higher than the second best.

In Table 7.2, we also report the average rank 1 accuracies on cross-view gait identification excluding the identical views (between probe and gallery). We can see other algorithms do not generalise well in this large-scale cross-view gait recognition evaluation, while our DiGGAN can still remain very competitive results.

[1] http://www.github.com/anomynous

Table 7.2 Average rank 1 identification rates (%) under Probe 0°,30°,60°,90° excluding identical view (cooperative mode) on OU-MVLP dataset.

	Probe				
Method	0°	30°	60°	90°	Mean
VTM[95]	0.4	1.6	2.2	2.1	1.6
GEINet[120]	8.2	32.3	33.6	33.6	26.9
Siamese[150]	11.4	27.9	26.7	26.2	23.1
CNN-MT[142]	7.1	24.0	28.2	21.7	20.3
CNN-LB[142]	6.2	22.2	26.9	21.2	19.1
DM[149]	0.4	0.7	1.9	2.0	1.3
MGANs[53]	5.6	17.4	19.7	11.4	13.5
DiGGAN(ours)	**48.9**	**62.3**	**59.1**	**57.8**	**57.0**

Effect of Triplet Loss and Identity Discriminator To explore the effect of the triplet loss, we trained two separate models on OU-MVLP: one with the triplet constrains on z and another without the triplet constrains. We compared them with the state-of-the-art method GEINet[120]. The results are shown in Table. 7.3. Although without the triplet loss, our method still outperforms the state-of-the-art, the improvement by introducing the triplet loss is significant as illustrated.

Table 7.3 Average rank 1 identification rates (%). (w/o T) indicates the model without triplet loss.

	Probe				
Method	0°	30°	60°	90°	Mean
GEINet[120]	25.2	46.6	46.8	44.0	40.7
DiGGAN(w/o T)	37.6	50.8	52.7	51.3	48.1
DiGGAN	56.4	69.2	66.5	65.8	64.5

Experimental Results on CASIA-B CASIA-B is a relative small dataset. We evaluated our model and report the average recognition accuracies on CASIA-B in Table 7.4. The comparison is conducted under the probe views 54°, 90° and 126° and with several methods such as VTM [72], C3A [10], ViDP [59], CNN [142] and MGANs [53] The results show that our method yields the competitive performance under probe 54° while getting significant improvements under probe 90° and 124°, which indicates our framework works well on small scale datasets.

Cross Dataset Evaluation In this section, we evaluated the generalisation ability of our model. We trained three models, among which the first model (M_O) is trained on OU-MVLP dataset only, the second model (M_C) is trained on CASIA-B dataset and the last model (M_{O+C}) is first trained on OU-MVLP and then fine-tuned on CASIA-B. We report the

Table 7.4 Average rank 1 identification rates (%) under Probe 54°, 90° and 126° excluding identical view (cooperative mode) on CASIA-B dataset.

	Probe			
Method	54°	90°	126°	Mean
VTM[95]	55.0	46.0	54.0	51.0
C3A[10]	75.7	63.7	74.8	71.4
ViDP[59]	64.2	60.4	65.0	63.2
CNN[142]	**94.6**	88.3	93.8	92.2
MGANs[53]	84.2	72.3	83.0	79.8
DiGGAN(ours)	94.4	**91.2**	**93.9**	**93.2**

average rank 1 identification rates of each model on the 62 subjects in CASIA-B's test set, and the results are shown in Table 7.5. We can see that the model M_O trained on OU-MVLP yields a promising identification rate on CASIA-B dataset. We can also find that pre-training on OU-MVLP dataset helps the model M_{O+C} to achieve the best results among the three because of its massive number of training samples.However, we noticed that M_{O+C} does not benefit much from a large pretrain set. A possible reason is that the view angles as well as the nationalities of the subjects in OU-MVLP and CASIA-B are very different. Nevertheless, the experimental results suggest it is not harmful to use the large OU-MVLP for representation learning. In fact, based on the learned representation, even without local fine tuning, our model M_O can outperform all the existing methods except the CNN[142], which shows our framework has a very strong generalisation ability.

Table 7.5 Average rank 1 identification rates (%) under Probe 0°,30°,60°,90° excluding identical view (cooperative mode) on CASIA-B.

	Probe			
Model	54°	90°	126°	Mean
M_O	86.2	82.2	84.7	84.4
M_C	94.4	91.2	93.9	93.2
M_{O+C}	**94.6**	**91.3**	**93.9**	**93.3**

7.4.2 In-depth Analysis

To better understand the success of our proposed model, this section provides detailed discussions and verifies some key statements in our methodology. All experimental results are based on OU-MVLP dataset.

Uncooperative Setting Results Compared with cooperative mode, this scenario is more challenging since the gallery views are non-uniform. Following the settings in [128], we

randomly select one from the 14 view angles for each test subject in gallery. Furthermore, considering the cost of collecting full-view training samples, it would be more practical to train the model with less views but can generalise to more. In this paper, we thus add an extra challenge and use the same model that is trained by only 4 angles and the rest of 10 angles in the test gallery are assumed as unseen. To the best of our knowledge, this is the first attempt to match gait images from unseen view angles in the test gallery. In Table 7.6, our model significantly outperforms state-of-the-art approaches that are trained by full 14 views.

Table 7.6 Rank 1 identification rate (%) for all baselines in uncooperative setting on OU-MVLP dataset.

	Probe				
Method	0°	30°	60°	90°	Mean
GEINet[120]	15.7	41.0	39.7	39.5	34.0
Siamese[150]	15.6	36.2	33.1	36.5	30.3
CNN-LB[142]	14.2	32.7	32.3	34.6	28.5
CNN-MT[142]	11.1	31.5	31.1	29.8	25.9
DM[149]	7.1	7.4	7.5	9.7	7.9
DiGGAN(ours)	**30.8**	**43.6**	**41.3**	**42.5**	**39.6**

Performance on Small-scale Gallery In many realistic applications, such as indoor office, the gallery size can be smaller. In Fig. 7.5, we can see the performance tends to be higher with smaller gallery. At 100-identity scale, the accuracies under all views exceed 90%, which is in line with the experimental results on the small-scale CASIA-B. Given the high performance, our model has many potential industrial values..

Any-to-Any View Gait Evidence Generation One of the advantages of our proposed model is that we can generate gait images from arbitrary view angle to all target angles whereas existing approaches can only achieve 1-to-1 generation [53]. Such an extension helps the understanding to humans when the identification is based on the latent features and thus improve the user's trust.

Fig. 7.9 shows the generated 14 views (0° - 90°, 180° -270°) given an input image at 0°. We also show the generated gait images of four typical views (0° , 30°, 60° ,90°) using these four angles as input (Fig. 7.6). We can see that the generated gait images have a high similarity with the ground truth, even for a large view variance. These cross-view generated gait images can be used as evidence in surveillance and forensic applications.

We also conduct a new scenario that has not been considered in previous works, in which the training dataset does not contain certain views that appear in test dataset. In Fig. 7.7, we show the generated view 60° which is completely unseen during training. We can see that

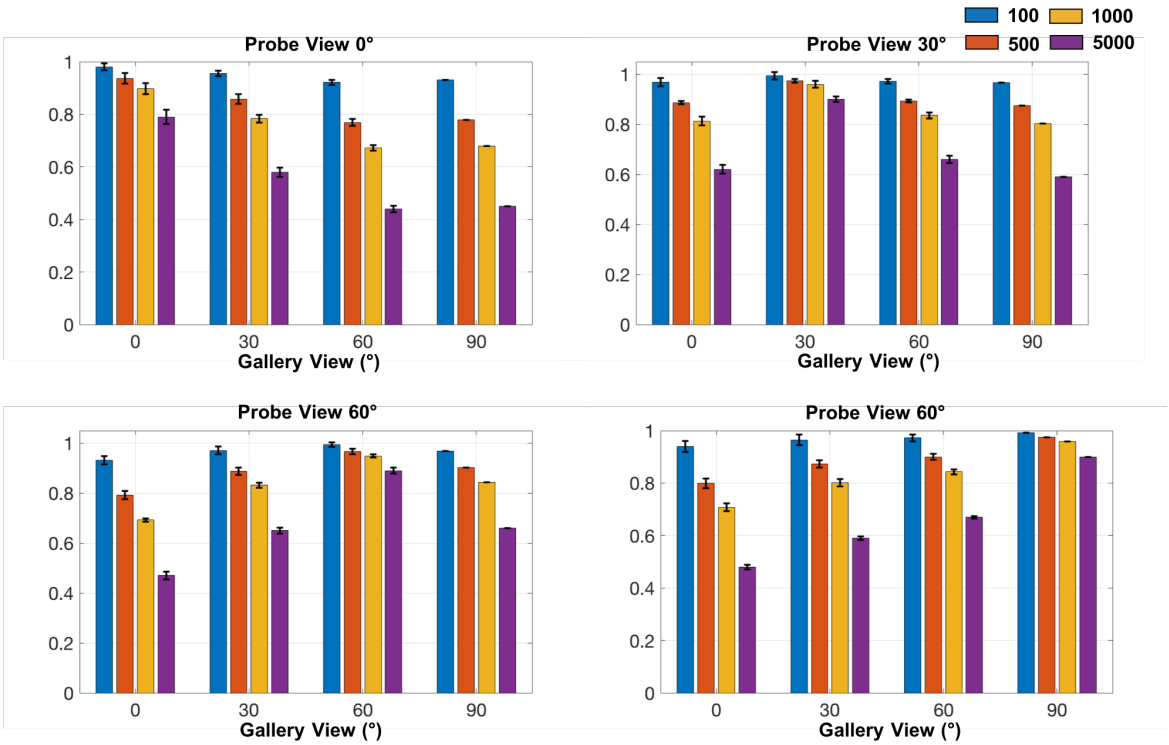


Fig. 7.5 Performance *w.r.t.* the size of gallery (in cooperative mode) on OU-MVLP.

both the identity and angle information can be generated, although some details (*e.g.* hands and feet) are missing.

Consistency Evaluation between Latent Code Searching and Evidence To evaluate the effectiveness of our generated evidence, we combine it with the results of latent code searching (rank 5). As Fig. 7.8 shows, the generated gait evidence of the subject identified by searching in latent space have high similarity with the input probe image, which indicates that good consistency is achieved in both latent space and generated image level. Moreover, this demonstrates the generated evidence is effective, which could be applied to the real-world forensic situations. On the other hand, we can see that the generated images for the first five nearest subjects are similar (Fig. 7.8), which means if the latent codes are close with each other, the generated images also look similar. It also proves the effectiveness of our model.

7.5 Summary

This chapter studied a challenging large-scale cross-view gait recognition problem. Using GANs to generate different views, the learnt latent embedding achieved remarkable cross-view transferability. The model effectively incorporated three modules. The D_{angle} loss provided an interactive interface through which a given arbitrary view could be used to

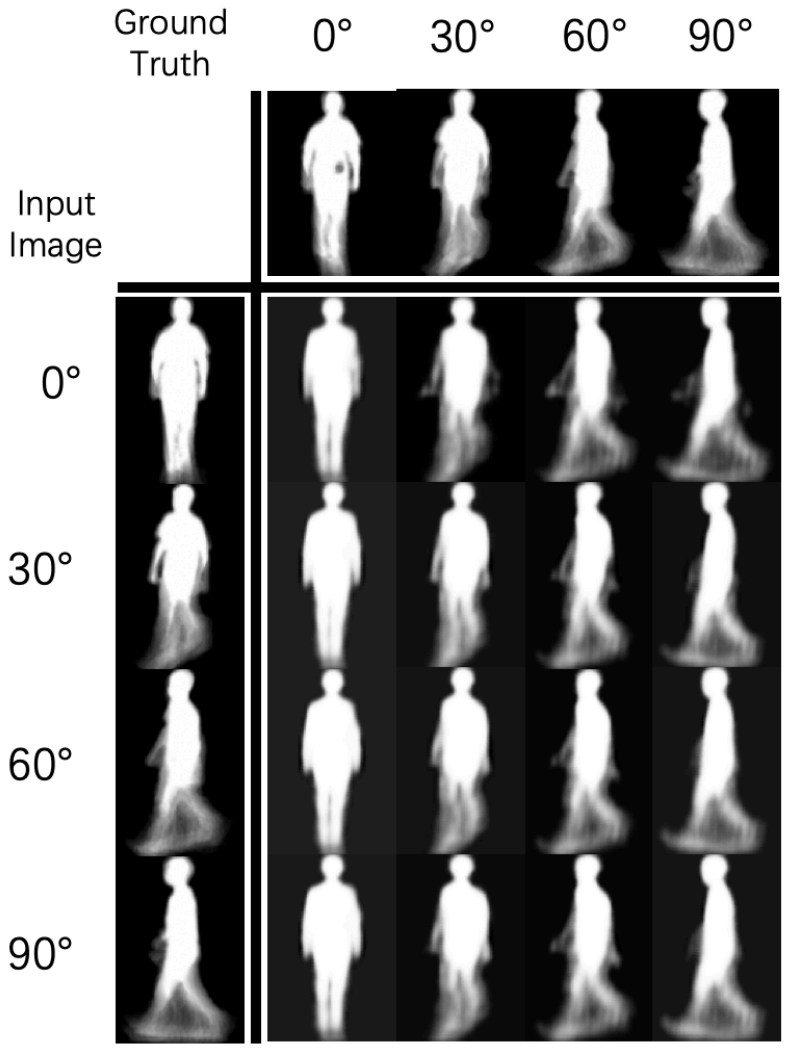


Fig. 7.6 Generated images at 0°, 30°, 60° and 90° with different input views. The top row shows the ground truth GEIs from the target views in the gallery. The first column shows the input GEIs from the probe. The images in bottom right 4×4 matrix are the generated GEIs.

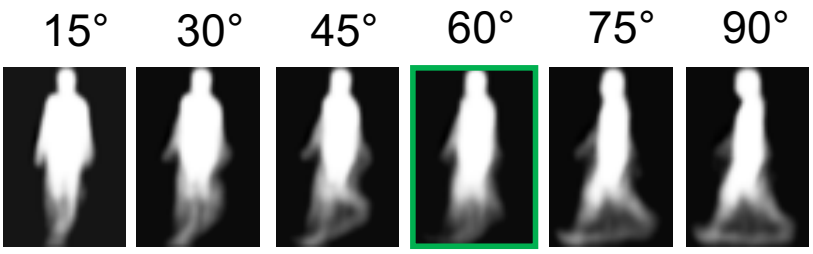


Fig. 7.7 Gait view generation: 6 generated GEI images with corresponding views. 60° is completely unseen during training.

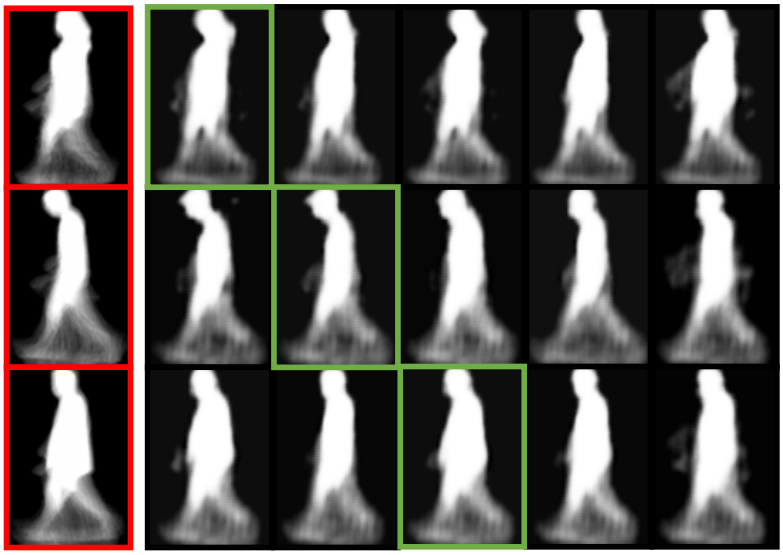


Fig. 7.8 Qualitative analysis of the evidence generation. The first column marked with red box illustrates three different GEIs in the probe. The rest five images in each row are generated GEIs based on the 5 most similar reference(*i.e.* latent code) templates.

generate all of other views. The D_{id} loss preserves identity sensitive information in the generated images. To further discriminate a large number of identities, triplet constraint was introduced onto the latent embedding. Moreover, since the triplet training incorporated images from different views, the inter-identity distance was enlarged, which further decorrelated effects of the cross-view problem. Extensive experiments manifested promising improvements over the state-of-the-arts. Our method also achieved the best results in the non-cooperative scenario, which has non-uniform views in the gallery. More reliable performance was achieved in small-scale datasets (*i.e.* CASIA-B) and we further show our DiGGAN framework can effectively take advantage of large dataset for cross dataset generalisation. Detailed training strategy was discussed so as the model could benefit both gait recognition domain and experts of other domains who would use GANs to solve their problems. Overall, this chapter made a breakthrough towards reliable cross-view gait recognition at a very large scale with generated evidence for practical applications.

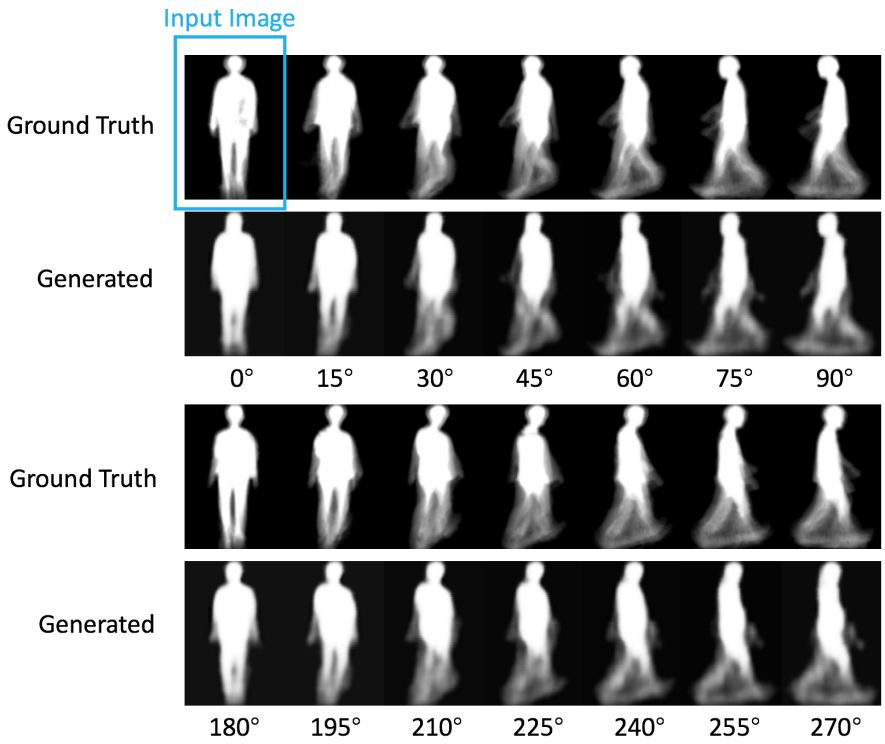


Fig. 7.9 The generated images of 14 views with the input image at 0°, which is indicated by the blue box on top left.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

In this thesis, four fundamental problems are studied in biometric information analyses. In chapter 3 and 4, we proposed two novel models for age estimation problem. In chapter 5, we proposed a new task as well as a new framework for face retrieval. Followed by chapter 6, a novel aging progression model is proposed and the facial image at an arbitrary age can be generated, in promising quality. Lastly in chapter 7, we investigated the biometric information analysis on gait, as an aid for those on face. Detailed conclusions will be given in following subsections.

8.1.1 Age estimation

In chapter 3 and 4, we proposed two novel models for age estimation problem. We first proposed a clustering based age grouping protocol. Different with conventional methods that divided ages into groups empirically, using clustering results as the age groups can greatly reveal the intrinsic structure. For example, the self formed age groups in our framework may have overlapping. Our understanding of this is, although subject to the evaluation metric of MAE, one age is just a numeral, the variations within each age label are still non-ignoble.

Inspired by this fact, we proposed a metric learning based method for age estimation in chapter 4. Because neither classification nor regression can reflect these intraclass variations (Here we take each age as a class for an easier and better understanding). We proposed a quartet based loss that can capture the relationships between distances thus the natural constraints of human aging progress are imposed on the learned metric; We conducted the experiments on three benchmarks, and the experimental results show that our method outperforms the state-of-the-art methods. Additionally, the experiment on FGNet dataset

shows that our method still works on small dataset as quartet model exploring deeper structural information.

8.1.2 Face retrieval

In chapter 5, we proposed a dual-reference face retrieval framework, which tackles the problem of retrieving a person's face image at a 'given' age. In the proposed framework, the face images are first project on a joint manifold, where the identity and age changes smoothly along respect directions. Then two metrics are learned to measure the similarities subject to identity and age. The final retrieval can be conducted by using a nearest neigh our search on the manifold. We have systematically evaluated our approach on CACD, FGNet and MORPH, and the corresponding results show that the proposed approach achieves promising results on this new task and the framework is stable and robust. In this chapter, we also realized that the image can contain much more information than the text or numeral. This finding inspired our work in the next chapter.

8.1.3 Aging progression

In chapter 6 we proposed a new age synthesis task: dual-reference age synthesis. By given two reference face images, synthesize a face image which has the same identity information with one image and at the age of the other one, what is a challenge. In the proposed framework, a joint manifold embedding is abstracted from training data through the identity agent and age agent, then the proposed model is trained on adversarial way with the joint manifold embedding, and is optimized with identity preserving function and age preserving function. Detailed training strategy was discussed in this chapter, the experimental results on UTKFace and CACD are given and are analyzed thoroughly. The corresponding results show that the proposed approach achieves promising results on this new task.

8.1.4 Gait recognition

This chapter studied a challenging large-scale cross-view gait recognition problem. Using GANs to generate different views, the learnt latent embedding achieved remarkable cross-view transferability. The model effectively incorporated three modules. The D_{angle} loss provided an interactive interface through which a given arbitrary view could be used to generate all of other views. The D_{id} loss preserves identity sensitive information in the generated images. To further discriminate a large number of identities, triplet constraint was introduced onto the latent embedding. Moreover, since the triplet training incorporated

images from different views, the inter-identity distance was enlarged, which further decorrelated effects of the cross-view problem. Extensive experiments manifested promising improvements over the state-of-the-arts. Our method also achieved the best results in the non-cooperative scenario, which has non-uniform views in the gallery. More reliable performance was achieved in small-scale datasets (*i.e.* CASIA-B) and we further show our DiGGAN framework can effectively take advantage of large dataset for cross dataset generalisation. Detailed training strategy was discussed so as the model could benefit both gait recognition domain and experts of other domains who would use GANs to solve their problems. Overall, this paper made a breakthrough towards reliable cross-view gait recognition at a very large scale with generated evidence for practical applications.

8.2 Future Work

There can be many future directions upon this thesis.

Unified Facial Image Analysis First, the age estimation, face retrieval and aging progression can be unified into one framework, as these tasks share a same information source — human face. Although age estimation and conventional face retrieval are two antithesis. In age estimation, we try to avoid the biases introduced by personal information while for face retrieval problem, the age information is a distraction. However, a potential solution is to find a common space that disentangles the age and identity information, for example, the age and identity are lied on two orthogonal spaces. We are going to investigate this in the close future.

Multi Biometric Attribute Retrieval In chapter 5, we proposed a dual reference face retrieval task, where the system takes two inputs and gives an output with the same attribute a, namely identity with the first input and another same attribute b, namely age with the second input. It is intuitive to extend this to multi biometric attribute retrieval. For example, when a victim cannot remember what exactly the criminal looks like but can describe him/her with several celebrities. It can be combined with the first future direction.

Attribute Interpolation for Generative Adversarial Network In chapter 6, we proposed a GAN to generate human faces at different ages. In chapter 7, we employed a GAN to generate gait energy images at different view angles. The experimental results in these works and the generated images show that the proposed models are very effective and robust. However, existing GAN based models cannot generate unseen attributes. Because the generative model learns and mimics the distribution from training set, it is very unlikely to generate samples without seeing their distribution. But for attributes with ordinal information such as age and

view angle, is it possible to infer their unseen distributions? We leave this question for the future work.

References

[1] (1999). Computer history - 1600's.

[2] (2019). Image uploaded to internet everyday.

[3] Aggarwal, H. and Vishwakarma, D. K. (2018). Covariate conscious approach for gait recognition based upon zernike moment invariants. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):397–407.

[4] Ahonen, T., Hadid, A., and Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041.

[5] Antipov, G., Baccouche, M., and Dugelay, J.-L. (2017). Face aging with conditional generative adversarial networks. *arXiv preprint arXiv:1702.01983*.

[6] Ariyanto, G. and Nixon, M. (2011). Model-based 3d gait biometrics. In *IJCB*, pages 1–7.

[7] Arjovsky, M., Chintala, S., and Bottou, L. (2017a). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223.

[8] Arjovsky, M., Chintala, S., and Bottou, L. (2017b). Wasserstein generative adversarial networks. In *ICML*, pages 214–223.

[9] Bagherian, E. and Rahmat, R. W. O. (2008). Facial feature extraction for face recognition: a review. In *2008 International Symposium on Information Technology*, volume 2, pages 1–9. IEEE.

[10] Bashir, K., Xiang, T., and Gong, S. (2010). Cross-view gait recognition using correlation strength. In *BMVC*, pages 1–11.

[11] Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

[12] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):711–720.

[13] Bhattacharjee, D., Halder, S., Nasipuri, M., Basu, D. K., and Kundu, M. (2011). Construction of human faces from textual descriptions. *Soft Computing*, 15(3):429–447.

[14] Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

[15] Bledsoe, W. W. (1964). The model method in facial recognition. *Technical Report*.

[16] Bouchrika, I., Goffredo, M., Carter, J., and Nixon, M. S. (2011). On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889.

[17] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7.

[18] Bukar, A. M., Ugail, H., and Hussain, N. (2017). On facial age progression based on modified active appearance models with face texture. In *Advances in Computational Intelligence Systems*, pages 465–479. Springer.

[19] Burt, D. M. and Perrett, D. I. (1995). Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 259(1355):137–143.

[20] Cao, D., Lei, Z., Zhang, Z., Feng, J., and Li, S. Z. (2012). Human age estimation using ranking svm. In *Chinese Conference on Biometric Recognition*, pages 324–331. Springer.

[21] Chang, K.-Y., Chen, C.-S., and Hung, Y.-P. (2011). Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*, pages 585–592. IEEE.

[22] Chen, B.-C., Chen, C.-S., and Hsu, W. H. (2014a). Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, pages 768–783. Springer.

[23] Chen, B.-C., Chen, C.-S., and Hsu, W. H. (2014b). Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[24] Chen, S., Zhang, C., Dong, M., Le, J., and Rao, M. (2017a). Using ranking-cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5183–5192.

[25] Chen, W., Chen, X., Zhang, J., and Huang, K. (2017b). Beyond triplet loss: a deep quadruplet network for person re-identification. *arXiv preprint arXiv:1704.01719*.

[26] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001a). Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685.

[28] Cootes, T. F., Edwards, G. J., Taylor, C. J., et al. (2001b). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685.

[29] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

[30] Dalal, N. and Triggs, B. (2005a). Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society.

[31] Dalal, N. and Triggs, B. (2005b). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

[32] Ding, H., Sricharan, K., and Chellappa, R. (2017). Exprgan: Facial expression editing with controllable expression intensity. *arXiv preprint arXiv:1709.03842*.

[33] Escalera, S., Fabian, J., Pardo, P., Baró, X., Gonzalez, J., Escalante, H. J., Misevic, D., Steiner, U., and Guyon, I. (2015). Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9.

[34] Geng, X., Smith-Miles, K., and Zhou, Z.-H. (2008). Facial age estimation by nonlinear aging pattern subspace. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 721–724. ACM.

[35] Geng, X., Yin, C., and Zhou, Z.-H. (2013). Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412.

[36] Geng, X., Zhou, Z.-H., and Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2234–2240.

[37] Goffredo, M., Bouchrika, I., Carter, J., and Nixon, M. (2010). Self-calibrating view-invariant gait biometrics. *IEEE TSMC*, 40(4):997–1008.

[38] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[39] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014b). Generative adversarial nets. In *NIPS*, pages 2672–2680.

[40] Guan, Y., Li, C., and Roli, F. (2015). On reducing the effect of covariate factors in gait recognition: a classifier ensemble method. *TPAMI*, 37(7):1521–1528.

[41] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *NIPS*, pages 5767–5777.

[42] Gunay, A. and Nabiyev, V. V. (2008). Automatic age classification with lbp. In *2008 23rd International Symposium on Computer and Information Sciences*, pages 1–4. IEEE.

[43] Guo, G., Fu, Y., Dyer, C. R., and Huang, T. S. (2008a). Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188.

[44] Guo, G., Fu, Y., Huang, T. S., and Dyer, C. R. (2008b). Locally adjusted robust regression for human age estimation. In *2008 Ieee Workshop on Applications of Computer Vision*, pages 1–6. IEEE.

[45] Guo, G. and Mu, G. (2011). Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*, pages 657–664. IEEE.

[46] Guo, G. and Mu, G. (2013). Joint estimation of age, gender and ethnicity: Cca vs. pls. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE.

[47] Guo, G., Mu, G., Fu, Y., and Huang, T. S. (2009a). Human age estimation using bio-inspired features. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119. IEEE.

[48] Guo, G., Mu, G., Fu, Y., and Huang, T. S. (2009b). Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 112–119. IEEE.

[49] Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 1735–1742. IEEE.

[50] Han, H., Otto, C., and Jain, A. K. (2013). Age estimation from face images: Human vs. machine performance. In *Biometrics (ICB), 2013 International Conference on*, pages 1–8. IEEE.

[51] Han, J. and Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE TPAMI*, 28(2):316–322.

[52] He, X., Yan, S., Hu, Y., Niyogi, P., and Zhang, H.-J. (2005). Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340.

[53] He, Y., Zhang, J., Shan, H., and Wang, L. (2019). Multi-task gans for view-specific feature learning in gait recognition. *IEEE TIFS*, 14(1):102–113.

[54] He, Z., Li, X., Zhang, Z., Wu, F., Geng, X., Zhang, Y., Yang, M.-H., and Zhuang, Y. (2017). Data-dependent label distribution learning for age estimation. *IEEE Transactions on Image processing*, 26(8):3846–3858.

[55] Heljakka, A., Solin, A., and Kannala, J. (2018). Recursive chaining of reversible image-to-image translators for face aging. *arXiv preprint arXiv:1802.05023*.

[56] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[57] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[58] Hu, B., Zheng, F., and Shao, L. (2018). Dual-reference face retrieval. In *AAAI*.

[59] Hu, M., Wang, Y., Zhang, Z., Little, J., and Huang, D. (2013). View-invariant discriminative projection for multi-view gait-based human identification. *TIFS*, 8(12):2034–2045.

[60] Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[61] Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. (2017). Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 4.

[62] Hutton, T. J., Buxton, B. F., Hammond, P., and Potts, H. W. (2003). Estimating average growth trajectories in shape-space using kernel smoothing. *Medical Imaging, IEEE Transactions on*, 22(6):747–753.

[63] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *arXiv preprint*.

[64] Iwama, H., Okumura, M., Makihara, Y., and Yagi, Y. (2012). The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE TIFS*, 7(5):1511–1521.

[65] Jain, A. K., Klare, B., and Park, U. (2012). Face matching and retrieval in forensics applications. *IEEE multimedia*, 19(1):20.

[66] Jones, J. P. and Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, 58(6):1233–1258.

[67] Karras, T., Laine, S., and Aila, T. (2018). A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*.

[68] Kaur, P., Zhang, H., and Dana, K. J. (2017). Photo-realistic facial texture transfer. *arXiv preprint arXiv:1706.04306*.

[69] Kemelmacher-Shlizerman, I., Suwajanakorn, S., and Seitz, S. M. (2014). Illumination-aware age progression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3334–3341.

[70] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[71] Kusakunniran, W., Wu, Q., Li, H., and Zhang, J. (2009). Multiple views gait recognition using view transformation model based on optimized gait energy image. In *ICCVW*, pages 1058–1064.

[72] Kusakunniran, W., Wu, Q., Zhang, J., and Li, H. (2010). Support vector regression for multi-view gait recognition based on local motion feature selection. In *CVPR*, pages 974–981.

[73] Kusakunniran, W., Wu, Q., Zhang, J., Li, H., and Wang, L. (2014). Recognizing gaits across views through correlated motion co-clustering. *IEEE TIP*, 23(2):696–709.

[74] Kusakunniran, W., Wu, Q., Zhang, J., Ma, Y., and Li, H. (2013). A new view-invariant feature for cross-view gait recognition. *IEEE TIFS*, 8(10):1642–1653.

[75] Kwon, Y. H. and Lobo, N. D. V. (1994). Age classification from facial images. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 762–767. IEEE.

[76] Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., et al. (2017). Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5969–5978.

[77] Lanitis, A. and Cootes, T. (2002). Fg-net aging data base. *Cyprus College*.

[78] Lanitis, A., Taylor, C., and Cootes, T. (1994). Automatic tracking, coding and reconstruction of human faces, using flexible appearance models. *Electronics Letters*, 30(19):1587–1588.

[79] Lanitis, A., Taylor, C. J., and Cootes, T. F. (2002). Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455.

[80] Larsen, P., Simonsen, E., and Lynnerup, N. (2008). Gait analysis in forensic medicine. *Journal of Forensic Sciences*, 53:1149–1153.

[81] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[82] LeCun, Y. et al. (2015). Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20:5.

[83] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*.

[84] Levi, G. and Hassner, T. (2015a). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42.

[85] Levi, G. and Hassner, T. (2015b). Age and gender classification using convolutional neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*.

[86] Li, C., Liu, Q., Liu, J., and Lu, H. (2012). Learning ordinal discriminative features for age estimation. In *Computer vision and pattern recognition (cvpr), 2012 ieee conference on*, pages 2570–2577. IEEE.

[87] Li, P., Hu, Y., He, R., and Sun, Z. (2018). Global and local consistent wavelet-domain age synthesis. *arXiv preprint arXiv:1809.07764*.

[88] Liu, B., Xia, Y., and Yu, P. S. (2000). Clustering through decision tree construction. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 20–29.

[89] Liu, C. and Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4):467–476.

[90] Liu, K.-H., Yan, S., and Kuo, C.-C. J. (2015). Age estimation via grouping and decision fusion. *IEEE Transactions on Information Forensics and Security*, 10(11):2408–2423.

[91] Liu, S., Sun, Y., Zhu, D., Bao, R., Wang, W., Shu, X., and Yan, S. (2017). Face aging with contextual generative adversarial nets. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 82–90. ACM.

[92] Long, Y., Zhu, F., Shao, L., and Han, J. (2018). Face recognition with a small occluded training set using spatial and statistical pooling. *Information Sciences*, 430:634–644.

[93] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[94] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

[95] Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., and Yagi, Y. (2006a). Gait recognition using a view transformation model in the frequency domain. In *ECCV*, volume 3953, pages 151–163.

[96] Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., and Yagi, Y. (2006b). Which reference view is effective for gait identification using a view transformation model? In *CVPRW*, pages 45–45. IEEE.

[97] Makihara, Y., Suzuki, A., Muramatsu, D., Li, X., and Yagi, Y. (2017). Joint intensity and spatial metric learning for robust gait recognition. In *CVPR*, pages 6786–6796.

[98] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[99] Mirza, M. and Osindero, S. (2014a). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

[100] Mirza, M. and Osindero, S. (2014b). Conditional generative adversarial nets. *arXiv*.

[101] Mu, G., Guo, G., Fu, Y., and Huang, T. S. (2009). Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 112–119. IEEE.

[102] Niu, Z., Zhou, M., Wang, L., Gao, X., and Hua, G. (2016). Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928.

[103] Odena, A., Olah, C., and Shlens, J. (2016). Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*.

[104] Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

[105] O'toole, A. J., Vetter, T., Volz, H., and Salter, E. M. (1997). Three-dimensional caricatures of human heads: distinctiveness and the perception of facial age. *Perception*, 26(6):719–732.

[106] Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *BMVC*, volume 1, page 6.

[107] Perarnau, G., van de Weijer, J., Raducanu, B., and Álvarez, J. M. (2016). Invertible conditional gans for image editing. *arXiv*.

[108] Radford, A., Metz, L., and Chintala, S. (2015a). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[109] Radford, A., Metz, L., and Chintala, S. (2015b). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*.

[110] Ricanek, K. and Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE.

[111] Rothe, R., Timofte, R., and Gool, L. V. (2015). Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

[112] Rothe, R., Timofte, R., and Gool, L. V. (2016). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*.

[113] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

[114] Sagonas, C., Panagakis, Y., Arunkumar, S., Ratha, N., and Zafeiriou, S. (2016). Back to the future: A fully automatic method for robust age progression. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 4226–4231. IEEE.

[115] Sarkar, S., Phillips, P., Liu, Z., Vega, I., Grother, P., and Ortiz, E. (2005). The humanid gait challenge problem: data sets, performance, and analysis. *IEEE TPAMI*, 27(2):162–177.

[116] Schroff, F., Kalenichenko, D., and Philbin, J. (2015a). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

[117] Schroff, F., Kalenichenko, D., and Philbin, J. (2015b). Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

[118] Shalev-Shwartz, S., Singer, Y., and Ng, A. Y. (2004). Online and batch learning of pseudo-metrics. In *International Conference on Machine Learning*.

[119] Shaw, R., McIntyre, M., and Mace, W. (1974). The role of symmetry in event perception. *Cornell University Press*.

[120] Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., and Yagi, Y. (2016). Geinet: View-invariant gait recognition using a convolutional neural network. In *ICB*, pages 1–8.

[121] Shu, X., Tang, J., Lai, H., Niu, Z., and Yan, S. (2016). Kinship-guided age progression. *Pattern Recognition*, 59:156–167.

[122] Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.

[123] Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. (2016). Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*.

[124] Song, L., Lu, Z., He, R., Sun, Z., and Tan, T. (2017). Geometry guided adversarial facial expression synthesis. *arXiv preprint arXiv:1712.03474*.

[125] Sun, Y., Chen, Y., Wang, X., and Tang, X. (2014). Deep learning face representation by joint identification-verification. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 1988–1996. Curran Associates, Inc.

[126] Suo, J., Min, F., Zhu, S., Shan, S., and Chen, X. (2007). A multi-resolution dynamic model for face aging simulation. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

[127] Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

[128] Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., and Yagi, Y. (2018). Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4.

[129] Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing*, 19(6):1635–1650.

[130] Tang, J., Li, Z., Lai, H., Zhang, L., Yan, S., et al. (2018). Personalized age progression with bi-level aging dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):905–917.

[131] Tang, X. and Wang, X. (2002). Face photo recognition using sketch. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–257. IEEE.

[132] Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., and Theobalt, C. (2017). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1274–1283.

[133] Thompson, D. W. et al. (1942). On growth and form. *On growth and form.*

[134] Todd, J. T., Mark, L. S., Shaw, R. E., Pittenger, J. B., et al. (1980). The perception of human growth. *Scientific american*, 242(2):132–144.

[135] Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86.

[136] Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., and Wu, Y. (2014). Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393.

[137] Wang, W., Yan, Y., Cui, Z., Feng, J., Yan, S., and Sebe, N. (2018a). Recurrent face aging with hierarchical autoregressive memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[138] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[139] Wang, Z., Tang, X., Luo, W., and Gao, S. (2018b). Face aging with identity-preserved conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7939–7947.

[140] Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs.

[141] Wolf, L., Hassner, T., and Taigman, Y. (2008). Descriptor based methods in the wild. In *Workshop on faces in'real-life'images: Detection, alignment, and recognition*.

[142] Wu, Z., Huang, Y., Wang, L., Wang, X., and Tan, T. (2017). A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE TPAMI*, 39(2):209–226.

[143] Yan, S., Liu, M., and Huang, T. S. (2008). Extracting age information from local spatially flexible patches. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 737–740. IEEE.

[144] Yang, H., Huang, D., Wang, Y., and Jain, A. K. (2017). Learning face age progression: A pyramid architecture of gans. *arXiv preprint arXiv:1711.10352*.

[145] Yang, H., Huang, D., Wang, Y., Wang, H., and Tang, Y. (2016). Face aging effect simulation using hidden factor analysis joint sparse representation. *IEEE Transactions on Image Processing*, 25(6):2493–2507.

[146] Yang, H.-F., Lin, B.-Y., Chang, K.-Y., and Chen, C.-S. (2013). Automatic age estimation from face images via deep ranking. *networks*, 35(8):1872–1886.

[147] Yi, D., Lei, Z., and Li, S. Z. (2014). Age estimation by multi-scale convolutional network. In *Asian conference on computer vision*, pages 144–158. Springer.

[148] Yu, S., Chen, H., Reyes, E. B. G., and Poh, N. (2017). Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *CVPRW*, pages 532–539.

[149] Yu, S., Tan, D., and Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444.

[150] Zhang, C., Liu, W., Ma, H., and Fu, H. (2016). Siamese neural network based gait recognition for human identification. In *ICASSP*, pages 2832–2836.

[151] Zhang, L., Shum, H. P., Liu, L., Guo, G., and Shao, L. (2019). Multiview discriminative marginal metric learning for makeup face verification. *Neurocomputing*, 333:339–350.

[152] Zhang, Z., Song, Y., and Qi, H. (2017a). Age progression/regression by conditional adversarial autoencoder. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

[153] Zhang, Z., Song, Y., and Qi, H. (2017b). Age progression/regression by conditional adversarial autoencoder. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

[154] Zhao, J., Mathieu, M., and LeCun, Y. (2016). Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*.

[155] Zheng, S., Zhang, J., Huang, K., He, R., and Tan, T. (2011). Robust view transformation model for gait recognition. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2073–2076. IEEE.