Fusion of Hand-crafted Descriptors with CNN-based Features for Facial Age Estimation

Shahram Taheri

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

Eastern Mediterranean University
January 2019
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Assoc. Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies the requirements of thesis for the degree of Doctor of Philosophy in Computer Engineering.

Prof. Dr. Hadi Işık Aybay
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Computer Engineering.

Assoc. Prof. Dr. Önsen Toygar
Supervisor

Examining Committee

1. Prof. Dr. Hasan Demirel _____

2. Prof. Dr. Fikret S. Gürgen _____

3. Prof. Dr. Bilge Günsel Kalyoncu _____

4. Assoc. Prof. Dr. Önsen Toygar _____

5. Assist. Prof. Dr. Yıltan Bitirim _____

ABSTRACT

Age estimation from facial images is an important application of biometrics. In contrast to other facial variations like occlusions, illumination, misalignment and facial expressions, ageing variation is affected by human genes, environment, lifestyle and health which make age estimation a challenging task. In this thesis, we propose three new age estimation systems for automatic facial age estimation. These systems utilize different type of feature descriptors, varying from hand-crafted ones to automatically learned features and combine them in different level of information fusion.

In the first proposed system, an integration of different feature extraction algorithms is utilized. This integration is performed by using two-level fusion of features and scores with the help of feature-level and score-level fusion techniques. In our proposed method, the advantage of using different types of features such as biologically-inspired features, texture-based features and appearance-based features is used. Feature-level fusion of biologically-inspired and texture-based methods is integrated into the proposed method and their combination is fused with an appearance-based method using score-level fusion.

The second proposed system exploits multi-stage features from a trained Convolutional Neural Network (CNN), and precisely combines these features with a selection of age-related hand-crafted features. This method utilizes a decision-level fusion of estimated ages by two different approaches; the first one uses feature-level fusion of different hand-crafted local feature descriptors for wrinkle, skin and facial

component while the second one uses score-level fusion of different feature layers of a CNN for age estimation.

In the third system, we propose a new architecture of deep neural networks namely Directed Acyclic Graph Convolutional Neural Networks (DAG-CNNs) for age estimation which automatically combine multi-stage features from different layers of a CNN. This system is constructed by adding multi-scale output connections to an underlying backbone from two well-known deep learning architectures, namely VGG-16 and GoogLeNet. DAG-CNNs not only fuse the feature extraction and classification stages of the age estimation into a single automated learning procedure, but also utilize multi-scale features and perform score-level fusion of multiple classifiers automatically. Experiments on the publicly available Morph-II and FG-NET databases prove the effectiveness of our novel method.

Keywords: Age estimation, convolutional neural networks (CNN), directed acyclic graph CNN (DAG-CNN), feature-level fusion, score-level fusion.

ÖZ

Yüz görüntülerinden yaş tahmini yapılması biyometri alanında önemli bir uygulamadır. Yüzün belli bir bölümünün kapatılması, ışıklandırma değişiklikleri, yanlış ayarlama ve yüz ifadesi değişikliklerine nazaran yaşlanma etkileri yüz görüntülerini farklı sebeplerden dolayı etkiler. İnsan genleri, çevre, yaşam kalitesi ve sağlık koşullarından dolayı etkilenen yüz görüntülerinden yaş tahmini yapılması zor bir işlemdir. Bu tezde, yüz görüntülerinden otomatik olarak yaş tahmini yapan üç yeni sistem önerilmiştir. Bu sistemler, el yapımı ve otomatik olarak öğrenilen değişik özniteliklerden yararlanıp, farklı bilgi kaynaşımı yöntemleriyle öznitelikleri birleştirir.

İlk önerilen yöntemde, farklı öznitelik çıkarıcı algoritmaların entegrasyonundan yararlanılmıştır. Bu entegrasyon, öznitelik-seviyesi ve skor-seviyesi kaynaşımları kullanan iki seviyeli kaynaşım tekniğiyle yapılmıştır. Önerilen yöntemde, biyolojik, dokusal ve görünüme dayalı değişik özniteliklerin avantajı kullanılmıştır. Biyolojik ve dokusal yöntemlerin öznitelik-seviyesi kaynaşımı ve bunların görünüme dayalı bir yöntemle skor-seviyesi kaynaşımı kullanarak sisteme entegre edilmesi sağlanmıştır.

İkinci önerilen yöntem, eğitilmiş evrişimli sinir ağı (CNN) yöntemiyle elde edilen çok aşamalı öznitelikleri kullanıp, bunlarla birlikte seçilen el yapımı öznitelikleri birleştirir. Bu yöntem, iki farklı yaklaşımla tahmin edilen yaşları karar-seviyesi kaynaşımı kullanarak birleştirir. İlk yaklaşım, kırışıklık, cilt ve yüz bileşenleri için farklı el yapımı yerel öznitelik tanımlayıcılarını öznitelik-seviyesi kaynaşımı ile

birleştirir. İkinci yaklaşımda ise farklı CNN öznitelik seviyelerini skor-seviyesi kaynaşımı ile birleştirir.

Üçüncü önerilen sistemde ise yaş tahmini için farklı CNN seviyelerinden elde edilen çok aşamalı öznitelikleri otomatik olarak birleştiren yönlendirilmiş çevrimsiz çizge evrişimli sinir ağı (DAG-CNN) olarak isimlendirilen yeni bir derin sinir ağları yapısı kullanılmıştır. Bu sistem, VGG-16 ve GoogleNet isimli iki tane iyi bilinen derin öğrenme yapısı omurgasına çok aşamalı çıktı bağlantıları eklemekle kurulmuştur. DAG-CNN yapıları, yaş tahmini sisteminin öznitelik çıkartma ve sınıflandırma aşamalarını tek bir otomatik öğrenme işlemine dönüştürürken, aynı zamanda da çok aşamalı özniteliklerden yararlanıp çoklu sınıflandırıcıları otomatik olarak skor-seviyesi kaynaşımı ile birleştirir. Morph-II ve FG-NET veritabanları üzerinde yapılan deneyler, yeni yöntemin etkisini ispatlamıştır.

Anahtar Kelimeler: yaş tahmini, evrişimli sinir ağları (CNN), yönlendirilmiş çevrimsiz çizge CNN (DAG-CNN), öznitelik-seviyesi kaynaşım, skor-seviyesi kaynaşım.

ACKNOWLEDGMENT

I would like to thank my doctorate advisor at Eastern Mediterranean University, Assoc. Prof. Dr. Önsen Toygar, for the privilege of starting my research career. Under her guidance, I was given the freedom to pursue my research in my own way and I greatly appreciated that freedom. It was a great pleasure to work and discuss with her. Thanks for supporting and encouraging me through these years.

I owe my sincere gratitude and a very special thanks to my wife, Dr. Zahra Golrizkhatami, for bearing me through the thick and thin of my Ph.D. tenure. If it wasn't for her support, co-operation and encouragement, this endeavor would not have been possible. The noble one who stood by me all through with patience and tolerance. My words of thanks cannot compensate her contribution, yet with all humility I thank her for her noble gesture and splendid support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAM	Active Appearance Models
AGES	Ageing Pattern Subspace
ANN	Artificial Neural Networks
ASM	Active Shape Model
BIF	Biologically Inspired feature
CAM	Contourlet Appearance Models
CCA	Canonical Correlation Analysis
CNN	Convolutional Neural Network
CPNN	Conditional Probability Neural Network
CS	Cumulative Score
DAG	Directed Acyclic Graph
DCT	Discrete Cosine Transform
FG-NET	Face and Gesture recognition Network
HOG	Histogram of Oriented Gradient
KFA	Kernel Fisher Analysis
KNN	K-Nearest Neighbor
KPLS	Kernel Partial Least Square
LARR	Local Adjusted Robust Regression
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LHI	Lotus Hill Research Institute
MAE	Mean Absolute Error
MLP	Multi Layer Perceptron

MRELBP	Median Robust Extended Local Binary Patterns
NN	Neural Networks
NN	Nearest Neighbor
ODFL	Ordinal Deep Feature Learning
OH	Ordinal Hyperplanes
OLPP	Orthogonal Locality Preserving Projections
PCA	Principal Component Analysis
PLS	Partial Least Square
ReLU	Rectified Linear Units
SVM	Support Vector Machine
SVR	Support Vector Regressor
VGG	Visual Geometry Group
YGA	Yamaha Gender and Age

Chapter 1

INTRODUCTION

Human faces are inherently associated with one's identity, gender, ethnicity, age group, etc. Human perception studies reveal that attributes derived from one's appearance such as one's emotional state, attractiveness, perceived age, etc. tend to significantly influence the interpersonal behavior [1]. The human face contains a great deal of information related to personal characteristics, including identification, emotion, age, gender and race.

Humans possess explicit, cue-based, and often culturally determined systems for perceiving the facial appearance of their peers [2]. Facial appearance is a primary source of information regarding the person's identity, gender, ethnicity, affective state, head pose, age and kinship relations. Hence, the perception of facial attributes governs person perception, interpersonal attraction, and consequently pro social and social behavior [2].

Among many age-related traits, facial appearance might be the most common one that people rely on for age estimation in daily life. As the typical example shown in Figure 1.1, the appearance of human faces exhibits remarkable changes with the progress of ageing. However, the human estimation of facial age is usually not as accurate as other kinds of facial information, such as identity, expression, and gender. Hence, developing automatic facial age estimation methods that are

comparable or even superior to the human ability in age estimation has become an attractive yet challenging topic emerging in recent years.

Figure 1.1: The Ageing Faces of One Subject in the FG-NET Database

Research towards the development of more detailed computational facial models that capture properties of facial cues related to ageing and kinship increasingly attracts the attention of the community. Indeed, by capitalizing on recent advances in machine learning, computer vision, and the available massive collections of facial data, significant progress has been made towards addressing the following problems:

1) Age Progression: is the process of transforming a facial visual input, in order to model it across different ages. The change of the age can be bidirectional, so that the facial output can appear either younger or older than the input.

2) Age Estimation: refers to the process of labeling a facial signal with an age or age group. The input signal can be 2D, 3D or image sequences. The problems that fall into this category can be divided further into two subcategories, depending on the labels of the training data: (a) real age or (b) apparent age estimation, which refers to the age that is inferred by humans based on the individual's appearance.

3) Age-Invariant Facial Characterization: involves the process of building a signal representation that is invariant to the facial transformations and appearance changes caused by ageing.

4) Kinship Verification: is defined as the process of determining whether the individuals in a pair of facial visual inputs are blood related.

Facial age estimation attempts to predict the real age value or age group based on facial images. Automatic age estimation task aims to use machine learning algorithms to estimate a person's age based on features extracted from face image. While extensive efforts have been devoted, facial age estimation still remains a challenging problem due to two aspects:

1) Lack of sufficient training data where each person should contain multiple images in a wide range of ages.

2) Large variations such as lighting, occlusion and cluttered background of face images which were usually captured in wild conditions.

Ageing is a complex problem because at different age points different types of changes occur in the human face. From childhood to teenage the changes are mostly related to craniofacial growth. At maturity the changes are mostly related to the skin color changes and texture effects, with facial skin starting to become slack and less smooth. So ageing is a mixture of all of these components. Moreover, ageing is a slow, irreversible, and a process that is unique to every human being. Many factors affect the ageing process. For example every person has different genes, blood group, life style and belongs to a particular ethnic group.

1.1 Problem Definition

Age estimation is a technique of automatically labeling the human face with an exact age or age group. This age can be either actual age, appearance age, perceived age, or estimated age. Actual age is the number of years one has accumulated since birth to

date, denoted as a real number. Appearance and perceived age are estimated based on visual age information portrayed on the face while estimated age is a subject's age estimated by a machine from the facial visual appearance. Appearance age is assumed to be consistent with actual age although there are variations due to the stochastic nature of ageing among individuals.

Age estimation can be approached as a multi-class classification problem or a regression problem or as an ensemble of both classification and regression in a hierarchical manner.

1.2 Motivation

There are many popular real-world applications related to facial ageing. Age estimation by machine is useful in applications where we do not need to specifically identify the user, but want to know his or her age when accessing restricted content.

- Potential applications: Machine-based age estimation methods could figure in a wide range of applications involving man-machine interfaces such as age adaptive interfaces and the enforcement of age-based access restrictions both to physical and electronic sites.
- Soft biometrics: Age estimation is a type of soft biometrics that provides ancillary information of the users' identity information. It can be used to complement existing biometric features, such as fingerprint and iris, to improve the performance of primary (hard) biometrics system.
- Humans are not perfect in the task of age estimation; hence automated age estimates could complement/aid the task of human operators.
- The problem of age estimation, bears similarities with other standard face interpretation/pattern recognition tasks (i.e. face recognition, expression

recognition etc.) hence the overall problem domain is more accessible to researchers.

- Accurate age estimates are usually required for other facial ageing related applications (i.e. age invariant face recognition and age progression) hence the starting point in dealing with facial ageing is usually the task of age estimation.
- For age estimation there are concrete ways to test the performance of different algorithms allowing in that way the efficient comparative evaluation of different algorithms.

1.3 Objectives

The primary aim of this work is to propose novel multi-stage CNN-based learned features fusion for facial age estimation. In order to achieve the primary aim, the following objectives have been established:

1) To explore novel method for automatically combining the multi-stage learned features from CNN.

2) To investigate the consolidation of hand-crafted features and CNN-based learned features in order to improve the accuracy of age estimation system.

3) To investigate the use of discriminative hand-crafted features such as shapes, wrinkles and appearances, for face age estimation.

4) To develop DAG-CNN architecture for age estimation.

5) To evaluate the proposed methods using the benchmarks.

1.4 Proposed Methods

In this study, three different systems are proposed for facial age estimation. In the first proposed system, an integration of different type of hand-crafted feature descriptors is applied by using feature-level and score-level fusion techniques. Feature-level fusion of biologically-inspired and texture-based methods is integrated into the proposed method and their combination is fused with an appearance-based method using score-level fusion.

The second proposed age estimation system is based on three different levels of information fusion. The wrinkle features extracted by Gabor filters and texture-based and shape-based features extracted by MRELBP and BIF are involved during the first level of information fusion (feature-level fusion) process. Multi-stage learned features from different layers of a trained CNN for age are combined together by score-level fusion method. Finally the obtained results of these two approaches are aggregated by using weighted averaging. The aggregation result shows that generic features extracted from CNN are enhanced by combining them with domain-specific features.

The third proposed system deploys a novel CNN architecture for age estimation which is based on automatic multi-stage fusion of information. Multi-stage learned features from different layers of a CNN are automatically combined together by score-level fusion method by using DAG-CNN architecture. We showed that DAG-CNN can improve the discrimination capability of a deep neural network by allowing its layers to share their learned features and work collaboratively for classification.

1.5 Contributions

The fundamental contributions of this thesis are:

1) A novel age estimation system is proposed by employing two levels of information fusion. Feature-level fusion of hand-crafted features such as shapes, skin and wrinkles is combined by appearance-based features in score-level, as presented in Chapter 5.

2) A novel age estimation approach is proposed by utilizing hand-crafted and CNN-based features in three levels of information fusion: Feature-level fusion of hand-crafted features such as shapes, skin and wrinkles is combined by multi-stage CNN-based features in score-level, as described in Chapter 5 [3].

3) A modern CNN architecture, namely, Directed Acyclic Graph-CNN is proposed for age estimation. In DAG-CNN, the feature extraction and regression stages of the age estimation are fused into a single automated learning procedure. Additionally, instead of using the last layer's feature, it utilizes multistage learned features which are extracted from different intermediate layers of the CNN [4]. Moreover, instead of manually combined different layers' features by feature-level fusion approach, it performs score-level fusion of multiple classifiers automatically, as discussed in Chapter 7.

4) Extension of DAG-CNN architecture for effective and accurate age estimation.

5) The study demonstrates that performing score-level fusion of multi-stage CNN-based features improves achieved accuracies.

1.6 Applications

In this section we present the most significant applications of modeling facial ageing in biometrics, forensics, medicine, cosmetology, business and entertainment.

1.6.1 Biometrics

The physical, physiological or behavioral cues based on which a person is recognized, e.g., iris, fingerprint, face are referred to as biometrics. Age and kinship comprise soft biometrics [2] as they can be used to boost the effectiveness of recognition. Besides improving face recognition accuracy there is a need for robustness towards ageing and kinship. Passport checks demand age-invariance in case of large age gap between the passport image and the person in question. Similarly, kinship invariance can potentially boost automatic face recognition, in particular towards distinguishing between kinds that look alike.

1.6.2 Forensics

Forensics includes a set of scientific techniques that are used for crime detection. Among these techniques, forensics art demonstrates the challenging task of producing a lifelike image of a person. In some cases, forensics experts face the need to change the age of a face. Such cases include updating archive images of wanted criminals as well as images of lost children. Additionally, cases such as matching orphaned or lost children and finding kinds of a victim, to name a few, demand the verification of kin relationship of two people. To that end, automatic genealogical research can significantly aid the work of law enforcement agencies.

1.6.3 Medicine and Cosmetology

Being able to model ageing and kinship and simulating the transformations on the face is vital for modern medicine and cosmetology. Medical home systems that are used to monitor elderly people can aid medical diagnosis by detecting premature ageing. On the other hand, automatic rejuvenation of the face can serve as a guide for cosmetic surgery. Particularly in the case of children, the parents' craniofacial ageing patterns can be used to predict the child's head growth, so that injury-related cosmetic surgery can have optimal long-term results.

1.6.4 Commercial Use

The ever-growing usage of social media and availability of personal photos have led to the rapid integration of facial analysis by businesses. Automatically estimating the customers' age can help with efficient customer profiling and age-oriented decision making, e.g., age-oriented advertisements. Likewise, targeted ads can be more effective when taking kin relation into considerations, as people's preferences can be affected by their relatives.

1.6.5 Entertainment

Visual effects that can age or rejuvenate the actors are already being used in the film making industry. These effects are not limited to movies but are also widely applied to photo editing. The imminent integration of such tools into popular design software will make for more realistic retouched of photos. Make-up artists that specialize in transforming the face can leverage the construction of person, age and kin specific morphable models. Guided by those models, the artists will transform the face of the actor for roles that demand sibling-like similarity between actors.

1.7 Thesis Organization

This thesis is organized into two parts as shown in Figure 1.2. The first part includes introductory chapters providing fundamental and background knowledge of the subject area and the state-of-the-art for age estimation. The second part of this thesis includes three contributory chapters where three novel age estimation systems are discussed. Finally, this thesis concludes the works and an insight to future directions of research and development are given.

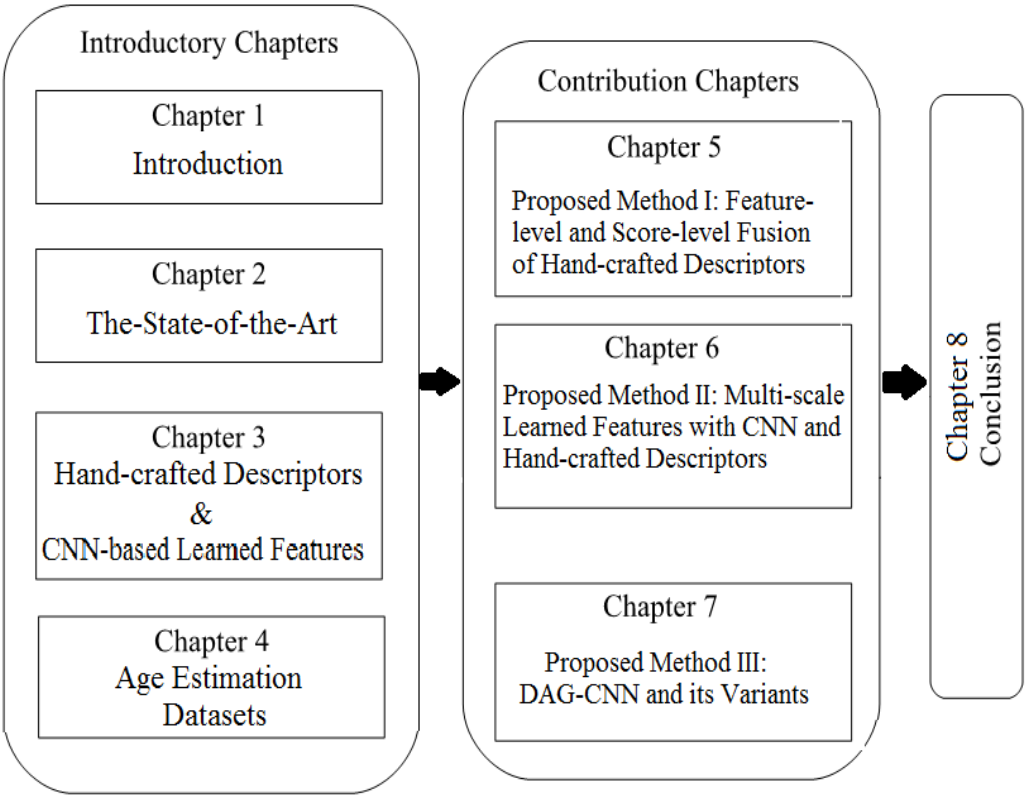


Figure 1.2: Thesis Organization

1.7.1 Introductory Chapters

Chapter 1 provides an overview of this thesis. It defines the problem domain, states the research motivation and specifies the thesis aims and objectives. It further highlights the contributions made by the thesis.

Chapter 2 presents a review of related work in age estimation and ageing feature extraction techniques. This chapter also gives background in facial ageing, factors and challenges affecting age estimation via faces.

Chapter 3 provides an overview of the research methodology. It presents fundamental knowledge about the wrinkle and skin features. It further highlights Biologically Inspired Feature and presents fundamental information about Convolutional Neural Network and its components.

Chapter 4 gives an overview to the age estimation datasets that are used in this study, namely, Morph-II and FG-NET datasets. This chapter also provides statistical overview of the aforementioned datasets and discusses the age estimation evaluation protocol for them.

1.7.2 Contribution Chapters

Chapter 5 presents the first proposed method. Biologically inspired features (BIF) and texture-based features such as MRELBP and HOG are combined by feature-level fusion technique and the result is fused with appearance-based method of KFA by score-level fusion approach. Results are compared with the state-of-the-art methods on Morph-II and FG-NET datasets.

Chapter 6 presents the second proposed method which exploits multi-stage features from a generic feature extractor, a trained convolutional neural network and precisely combined these features with a selection of age-related handcrafted features. FGNET and Morph-II datasets are used for performance evaluation with the state-of-the-art methods.

Chapter 7 presents the third proposed method. In this chapter a new architecture of deep neural networks namely Directed Acyclic Graph Convolutional Neural Networks (DAG-CNNs) for age estimation is introduced which exploits multi-stage features from different layers of a CNN and automatically combines them. Results are compared with the state-of-the-art methods on Morph-II and FG-NET datasets.

Finally, Chapter 8 concludes the thesis with a summary of contributions made by the thesis and an insight into future directions of research.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Age estimation has historically been one of the most challenging problems in the field of facial analysis [2]. Despite its multiple applications in many different areas there are relatively few publications compared to other topics in facial analysis. This difficulty is due to many factors including the following:

- Depending on the application scenario, the age estimation problem can be taken as a multiclass classification problem or a regression problem.
- Large databases are difficult to collect, especially series of chronological image from same individuals.
- The factors that affect the ageing process are uncontrollable and person specific.

The age estimation problem has generally two stages or blocks. The first one is the age representation in the images and the second one the learning algorithm. Several techniques have been published in order to deal with these two stages [2][3][6].

2.2 Age Representation

Age representation, i.e. the extraction of the features that represent the ageing patterns, is a very important step in the age estimation process. A good age representation contains enough variation of the data to express the full complexity of

the problem. There are many ways in the literature to represent ageing factors from a face image. The most significant ones are described in the next subsection.

2.2.1 Anthropomorphic Models

The first known work on age estimation from facial images was done by Kwon and Lobo [5]. Their approach is based in cranio-facial development theory using geometrical ratios between different face regions to classify images into one of three age groups (babies, young adults and senior adults). Figure 2.1 shows some of the points that are used in this approach. They used frontal images in a very strict setup to locate all face components. Ramanathan et al. [7][8] used a similar approach by using 8 ratios rather than the 6 used by Fi [9].

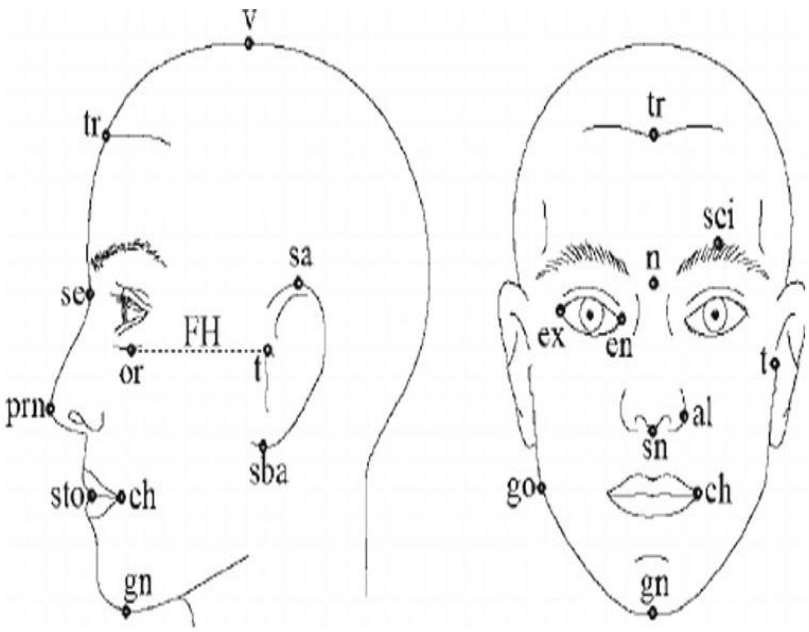


Figure 2.1: Anthropometric Points on the Face [2]

The problem of this model is that it can only be applied to face images of young people in a growing age, since afterwards the facial geometry does not change as much. It is also a problem that the anthropomorphic model requires frontal images, since it limits future applications.

2.2.2 Active Appearance Models

Active Appearance Models (AAM) is a statistical shape model proposed by Cootes et al. [10]. This model contains the shape and grey-level appearance of the object of interest which can generalize almost any valid example. This technique has been used to find the shape of faces by many researchers. Lanitis et al. [11] were the first to use the AAM model for age estimation by defining an ageing function $age = f(b)$, where b is a vector containing the parameters learned by the AAM. Figure 2.2 shows a sample of annotated face and points used for annotation.

Lanitis et al. [11] also tried different classifiers such as Quadratic Functions, Shortest Distance Classifier, Supervised Neural Network and Unsupervised Neural Network. Among all of the utilized classifier, they reported that Quadratic Functions achieved the highest performance.

Later on, Luu et al. [12], used AAM with 68 facial landmarks to classify faces into young or adult classes and then used two specialized functions to finally determine the age. In [12] they further improved the previous method by proposing Contourlet Appearance Models (CAM) which is more accurate and faster at calculating facial landmarks than AAM. This model has the ability of not only capturing global texture information like AAM but also local texture information using Non subsampled Contourlet Transform (NSCT) [13].

Chang et al. in [14] used AAM model with particular ranking formulation of support vectors, OHRank. The approach uses cost-sensitive aggregation to estimate ordinal hyperplanes (OH) and ranks them according to the relative order of ages.

Figure 2.2: Facial Shape (Left) and Appearance Annotation (Right) [2]

This model captures shape and texture information and in general performs better than the Anthropomorphic Models. This method can deal with any range of ages rather than just with young ages like the previous model. However, as suggested by Geng et al. [15], the ageing function is empirically determined, so there is no evidence suggesting that the relation between face and age is described just by a quadratic function.

2.2.3 Ageing Pattern Subspace

The researches presented in Geng et al. [15] were the first ones to explore Ageing pattErn Subspace (AGES) model. They defined an ageing pattern as a sequence of personal face images sorted in time order. Given a grey-scale face image I, where $I(x, y)$ determines the intensity of the pixel (x, y), then an ageing pattern can be represented as a three-dimensional matrix P , where $P\ (x, y, t)$ is the intensity of the pixel (x, y) in the face image at the time t. The images vector is filled with the available face images leaving empty the missing faces in the t axis. Now, the images in the age pattern vector can be processed and transformed into meaningful feature

vectors. Figure 2.3 shows vectorization of ageing pattern with missing images in the ageing pattern vector marked with m.

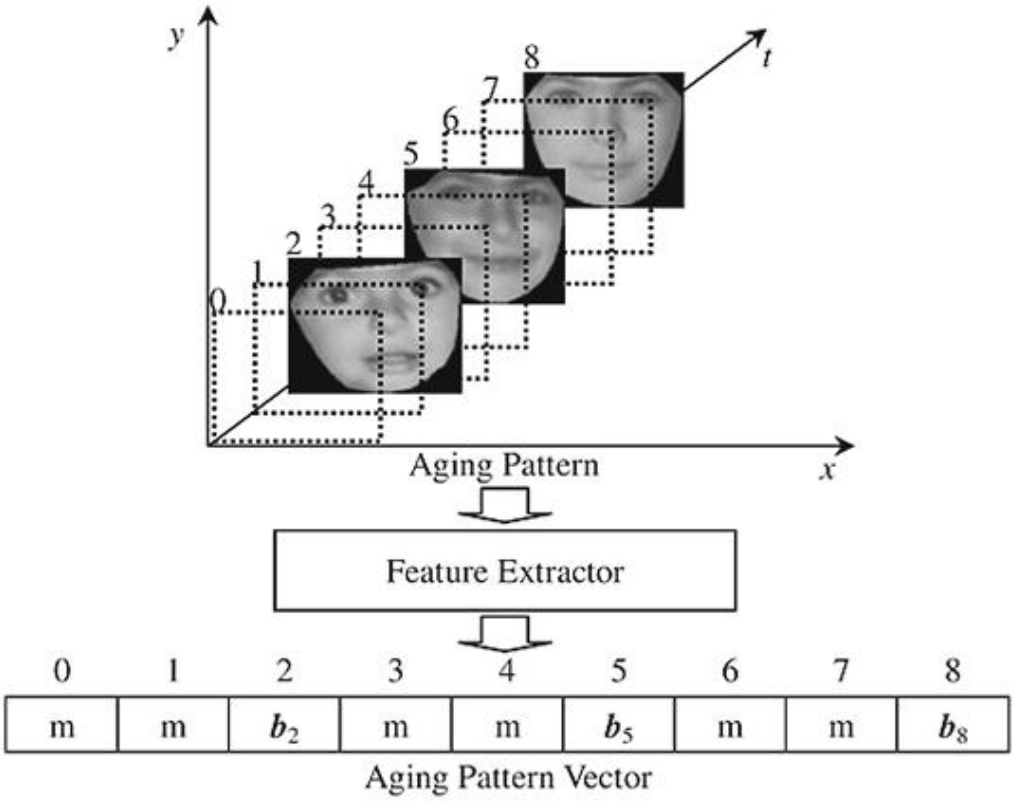


Figure 2.3: Ageing Pattern Vectorization. Age is Marked at the Corner of the Feature[10]

In order to extract the features X. Geng used AAM [15] since he capture the shape and texture of the face images. By representing ageing patterns in this way, the concepts of identity and time are naturally integrated into data without any pre-assumptions.

The principal drawback of the AGES method is that it assumes that there are images of the same individual at different ages, which is not true in many age databases. The databases that fulfill these requirements are small such as FG-NET with just 1000 images and very few representations of the same individual over time.

2.2.4 Age Manifold

Manifold learning methods are applied to find a sufficient embedding space and model the low-dimensional manifold data with a multiple linear regression function. Fu et al. [16] were the first to propose a manifold embedding approach for the age estimation problem.

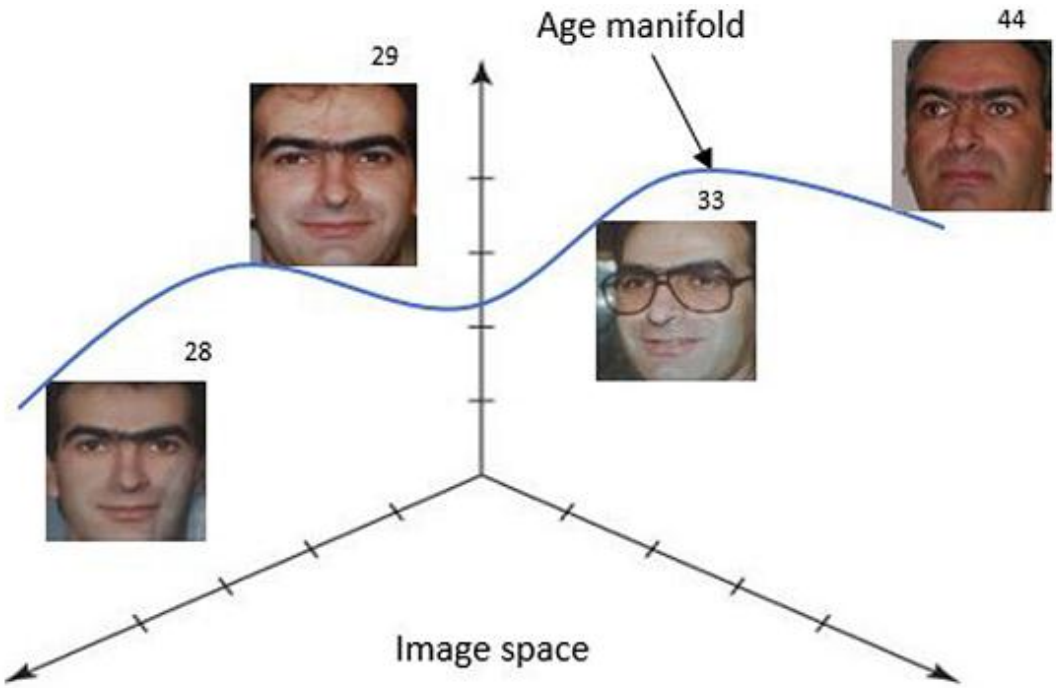


Figure 2.4: Simple Nonlinear Age Manifold [15]

The objective of this method is to find the low-dimensional representation in the embedded subspace capturing the intrinsic data distribution and geometric structure as well as its representation. Figure 2.4 shows a simple nonlinear projection function that models an image space into a 2D age manifold. Guo et al. [17] showed that the Orthogonal Locality Preserving Projections (OLPP) [18] is an effective algorithm to connect the manifold learning with subspace learning. In a posterior work [19], Guo et al. introduced a new approach, using kernel partial least square (KPLS) regression which reduces feature dimensionality and learn the ageing function in a single step.

On the other hand, Wu et al. [20] proposed to model the facial shapes as points on a Grassmann manifold. Age estimation is then considered as regression and classification problems on this manifold. Then, they proposed a method for combining this shape-based approach with other texture-based algorithms.

The main drawback of the age manifold representation is the large number of training instances required to learn the embedded manifold with statistical sufficiency.

2.2.5 Appearance Models

Appearance models focus on wrinkles, face texture and pattern analysis. From the beginning, researchers have tried to capture wrinkles and distinguish them from facial lines. Kwon et al. [21] proposed a wrinkle detector based on snakelets [22] placed into key wrinkle areas of the face. Hayashi et al. [23] combined both shape and texture to estimate age and gender. In Hayashi's proposed approach, the skin is extracted based on a shape model and then a histogram equalization is applied to emphasize wrinkles.

Other researchers have used the texture descriptor Local Binary Patterns (LBP) [24] in the age estimation problem [25][26] obtaining good classification results with Nearest Neighbour (NN) and Support Vector Machines(SVM) classification algorithms. The Gabor filter texture descriptor has [27] also been used in the age estimation task [28], proving to be more discriminative than LBP.

Guo et al. [29] proposed to use Biologically Inspired Features (BIF) [30] for age estimation via faces. The BIF descriptor tries to mimic how the visual cortex works, with a hierarchy of increasingly sophisticated representations. The BIF original

model proposed by Riesenhuber and Poggio [30] is based on a feed-forward model of the primate visual object recognition pathway, the "HMAX" model. The framework of the model contains alternative layers called simple (S) and complex (C) creating in each cycle a more elaborated representation. The S layers are created with a Gabor filtering on the input and the C layers generally operate a "MAX" operator over the previous S layer. Guo et al. [29] modifies the BIF model by changing the operator in the complex layer (C1) from "MAX" to "STD".

Han et al. [31] used the BIF features in a hybrid classification framework improving the previous results with this descriptor. Guo et al. [32] also used the BIF features, and focus to investigate a proposed single-step framework for joint estimation of age, gender and ethnicity. Both the Canonical Correlation Analysis (CCA) [33] and Partial Least Square (PLS) based methods were explored under the joint estimation framework.

In [34], Weng et al. employed a similar ranking technique as the one used by Chang et al. in [14] called MFOR. LBP histogram features are combined with principal components of BIF, shape and textural features of AAM, and Principal Component Analysis (PCA) projection of the original image pixels. Fusion of texture and local appearance descriptors (LBP and HOG features) have also been independently used for age estimation by Huerta et al. in [35].

The outstanding results obtained by some of these works point out the suitability of BIF features for the age estimation via faces task.

2.2.6 Other Models

Depending on the available data, the age estimation problem changes. Many researchers have tackled the age estimation problem with different types of data. A short overview is described below.

Ramanathan et al. [8] studied age progression of individual faces and proposed a method to perform face verification using a Bayesian age-difference classifier to improve the face verification algorithm.

Mikihara et al. [36] used a gait-based database to perform a viability study of age estimation using this type of data. The results show that in future research the combination of gait-based data and face-based age estimation could give very good results.

Xia et al. [37] were the first to attempt age estimation using 3D face images. The obtained results show that the depth dimension has a very discriminative power in this problem.

2.3 Age Estimation Learning Algorithm

Given an age representation, the next step is to determine the individual's age out of the ageing features. Age labels can be seen as a discrete set of classes or as a continuous label space, hence classification and regression learning methods can be used.

2.3.1 Classification Methods

The age estimation problem can be treated as a classification problem, where the solution space is discrete and the objective is to classify each face image into one of the age classes (a class could be an age range of several years or a single year).

Lanitis et al. [38] evaluated the performance of different classifiers such as quadratic function classifier, Artificial Neural Network (ANN) and K-Nearest Neighbors (KNN) classifier with their AAM model. Among the classification methods they tested, they claimed to perform better with the ANN, specifically the Multi Layer Perceptron (MLP), obtaining 4.78 Mean Absolute Error (MAE). The authors also proposed some extensions, for example, training age specific classifiers in a hierarchical fashion. With the extended methods the authors reduced the error to 4.38 MAE with the MLP and 3.82 MAE with the quadratic function (regression).

There have been previous proposals using neural networks, which are able to learn complex mappings and deal with outliers, for age estimation. In [38], Lanitis et al. used AAM encoded face parameters as an input for the supervised training of a neural network with a hidden layer. More recently, Geng et al. [39] tackled age estimation as a discrete classification problem using 70 classes, one for each age. The best algorithm proposed in this work (CPNN - Conditional Probability Neural Network) consists of a three-layered neural network, in which the input to the network includes both BIF features x and a numerical value for age y, and the output neuron is a single value of the conditional probability density function. An extensive comparison of these classification schemes for age estimation has been reported in Fernandez et al. [35]. In [40], Yang et al. used Convolutional Neural Network (CNN) for age estimation under surveillance scenarios and recently Yi et al. [41] and Yan et al. [42] have used CNN in the age estimation problem reporting very promising results in the Morph-II dataset (3.63 MAE).

Ueki [43] classified the images from the WIT_DB database into 11 age groups using Gaussian models in a low-dimensional 2DLDA+LDA feature space using the EM

Algorithm. The accuracy rates they achieved were 46.3%, 67.8% and 78.1% for age groups that were in the 5-year, 10-year and 15-year range respectively.

SVM have been also used for age classification, Guo et al. [17] trained an SVM for each pair of age classes and then using a binary tree search for testing, obtaining 5.55 MAE for females and 5.52 MAE for males in the YGA database and 7.16 MAE in the FG-NET database.

2.3.2 Regression Methods

The age of an individual is nothing else than the time passed from the individual's birth and time is a continuous dimension. Hence, the age estimation problem can be formulated as a regression problem where the objective is to find a regression function that explains the ageing in terms of the feature space.

Lanitis et al. [38] evaluated three regression functions, linear, quadratic and cubic and claimed that the quadratic function was the one which better described the age from their feature space. Y. Fu et al. [33] used linear, quadratic and cubic regression functions as learning algorithms for the manifold age representation. As Lanitis et al. [11] showed, Fu et al. [16] [44] also reported superior performance on the quadratic regression function, pointing out that cubic functions lead to over-fitting while linear functions lead to under-fitting.

Guo et al. [17] compared the performance of Support Vector Regressor (SVR) against the Local Adjusted Robust Regression (LARR) performance in the YGA and the FG-NET databases, concluding that LARR performs a more accurate estimation, achieving 5.25 MAE in YGA for female images, 5.30 in YGA for male images and

5.07 in FG-NET database. In their former work [29], the authors improve the SVR performance in the FG-NET to 4.77 MAE by using BIF features.

2.3.3 Hybrid Methods

Many authors have proposed mixture frameworks, using both classification and regression learning algorithms.

Guo et al. [45] proposed a probabilistic fusion approach. They use Bayes' rule to derive the predictor and then a sequential fusion strategy, so the output of the regressor is used as an intermediate decision which is then fed to the classifier to aid or affect the decision space of the classifier. Their fusion approach has better performance than other single step methods which they compare with.

SVM and SVR were used by Han et al. [31] in a hierarchical fashion. They proposed to use a binary decision tree with SVMs at each node to classify the images into different age ranges, which are coarsely assigned. Later, the age is fine grained by SVRs at the leaves. The SVRs are trained with 5-years-overlap between age ranges in order to reduce the misclassification error.

2.4 Deep Learning Methods for Age Estimation

Recently, many researchers have been using CNN for facial age estimation problems. Based on the number of architecture layers, these deep learning approaches are divided into two classes: shallow architecture and deep architecture.

Modern CNN architectures such as VGG-16 [46] and GoogLeNet [47] are two examples of deep architecture. The deep architecture suffers from over-fitting problem when there is a small number of training data like Morph-II dataset. Recently, the researchers used additional datasets with thousands of annotated

images and transfer learning approach to overcome this problem and achieved the-state-of-the-art results [48][49][50][51]. In the work of Yi et al. [41], they used several shallow multiscale CNNs on different face regions and obtained the MAE of 3.63 on Morph-II dataset. On the other hand, in [52], the authors used CNNs and proposed using a ranking encoding for age and gender and they reported the MAE of 3.5 on Morph-II dataset. Hu et al. [48] proposed a novel learning scheme to embed the age difference information.

Rothe et al. [51] proposed a deep learning solution for age estimation and introduced the IMDB-WIKI dataset which is the largest public age dataset. The authors reported the MAE of 2.68 on Morph-II dataset, which is the best reported result to the best of our knowledge. In [53] they proposed a conditional multitask learning method that architecturally factorizes an age variable into gender-conditioned age probabilities in a deep neural network. In order to overcome the lack of accurate training labels with discrete age values problem, they proposed a label expansion method that increases the number of accurate labels from weakly supervised categorical labels. Liu et al. [52] proposed an ordinal deep feature learning (ODFL) method to learn feature descriptors for face representation directly from raw pixels. They designed an end-to-end ordinal deep learning framework, where the complementary information of both feature extraction and age estimation is exploited to reinforce their model.

One of the age relevant topics is age progression which is defined as aesthetically re-rendering the ageing face at any future age for an individual face. In [55] the authors proposed novel bi-level dictionary learning based personalized age progression method. For each age group, they learned an ageing dictionary to reveal its ageing characteristics (e.g., wrinkles), based on face pairs from neighboring age groups. Shu

et al. [56] used a set of age-group specific dictionaries and a linear combination of these patterns to express a particular personalized ageing process. In [57] the authors presented a novel generative probabilistic model with a tractable density function for age progression. Their model inherits the strengths of both probabilistic graphical model and recent advances of ResNet.

In the few past years, some researchers tried to utilize the multi-scale features learned by different layers of a CNN for different problems. Tang et al. [58] proposed GoogLeNet based multi-stage feature fusion (G-MS2F) for scene recognition. The GoogLeNet model is employed and divided into three parts and the output features from each of the three parts are applied for final decision. In [59] the authors presented an object detection framework based on multi-stage convolutional features for pedestrian detection. Their framework extended the Fast R-CNN framework for the combination of several convolutional features from different stages of the used CNN to improve the network's detection accuracy.

Chapter 3

HAND-CRAFTED DESCRIPTORS AND CNN-BASED LEARNED FEATURES

3.1 Introduction

Feature extraction is an important step in an image classification system. The selection of discriminative feature descriptors is a critical decision and it affects the whole system performance. Therefore, we have chosen different types of well-known and efficient local and global feature extractors that have been successfully used in age estimation and the other applications like face recognition with their discrimination ability, computational efficiency, compact feature space size, and robustness to alignment and illumination variance.

3.2 Wrinkle Features

The ageing process has an impact beyond the facial wrinkles and lines that form on the skin's surface. Ageing affects multiple layers, including the bone, muscles, fat-pads, and skin [60].

3.2.1 Common Types of Wrinkles

There are many areas where wrinkles may start to appear. Figure 3.1 shows a few of the most common include:

- **Eye wrinkles** tend to be the first sign of ageing people notice. In the twenties and thirties, things like crow's feet, tear troughs, and bags underneath the eyes become visible.

- **Forehead wrinkles** can also appear early on, thanks to everyday facial expressions and repeated muscle movement. They typically stretch horizontally across the forehead but may be vertical as well, forming between the eyebrows.
- **Lip wrinkles** are common once a person reaches mid to late thirties. Lines around the mouth develop naturally as the person ages but can be seen earlier (in the twenties and thirties) if he/she is a smoker.

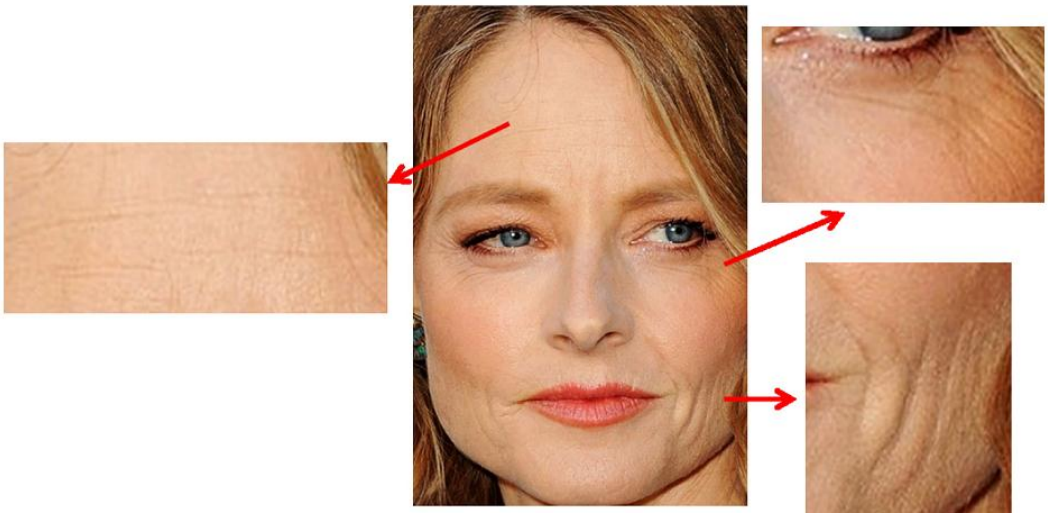


Figure 3.1: Image of a Subject Having Very Deep Wrinkles Around Eyes and Mouth and Light Wrinkles on Forehead [60]

Figure 3.2: Three Subjects of Same Age (52 Years) Haveing Visible Differences in the Appearance of Their Wrinkles [60]

The facial wrinkles provide discriminative information for age estimation, and these facial wrinkles have been considered by many researchers [3][23][60][61][62]. Figure 3.2 shows three different subjects of same age which have significant differences in the appearance of their wrinkles. In order to build pose robust face descriptors, we should detect precise facial landmarks. By extracting the region enclosed to these fiducial points, we obtain some patches which have the same semantics for different subject images. In order to extract wrinkle features effectively, we computed the strength and quantity of wrinkles in several facial patches by using Gabor filter [63] set on the direction of the facial muscles on these patches. The Gabor wavelet is defined as follows:

$$a = (\frac{U_h}{U_l})^{1/(S-1)}, \qquad \sigma_u = \frac{(a-1)U_h}{(a+1)\sqrt{2ln2}}$$

$$\sigma_v = \tan\left(\frac{\pi}{2K}\right)\left[U_h - 2\ln\left(\frac{\sigma_u^2}{U_h}\right)\right]\left[2ln2 - \frac{(2ln2)^2\sigma_u^2}{U_h^2}\right]^{-\frac{1}{2}} \tag{3.1}$$

where U_l and U_h denote the lower and upper average frequencies, respectively. Therefore, the radial frequency of the sinusoid is equal to U_h. Details related to Gabor wavelet are presented in [65].

3.3 Skin Features

Facial skin also provides a lot of discriminative information for estimating the age. Unlike wrinkles, skin ageing appears randomly and non-uniformly for each facial part. Ageing of the facial skin mostly appears in the form of freckles and it reduces the amount of collagen that plays a role in reflecting light, and is distributed non-uniformly on the face. Figure 3.3 shows different skin ageing causes.

Youthful skin is soft, supple, smooth, well hydrated, and rich with cells that renew relatively rapidly. As we age, we experience a loss of facial glands, which results in less oil produced, contributing to less moisture in the skin. We lose collagen and elastin, which can lead to the formation of dynamic wrinkles, like laugh lines, frown lines, and crow's feet. Due to repeated facial movement, dynamic wrinkles eventually become static lines that are gradually etched into the skin over time. Additionally, sagging can occur because skin is no longer able to bounce back as it did in our youth.

Many factors impact the way our skin ages, including lifestyle choices and genetics. Lifestyle choices, like sun exposure, smoking, alcohol use, diet, and stress, can cause brown spots, rough skin, and wrinkles, as well as the premature onset and progression of ageing. Genetics affect all layers of the skin and contribute to thinning, dryness, and loss of elasticity of the skin during the ageing process.

Figure 3.3: Ageing Skin Causes [60]

As a result, the overall tone of the facial skin becomes non-uniform. Since ageing skin has smooth variations, we selected a feature extractor capable of analyzing microstructures of skin texture. For this purpose, we used Median Robust Extended Local Binary Pattern (MRELBP) [64] descriptor which detects very fine details such as edges, lines, spots and flat areas in a computationally efficient manner. In [65] the authors showed that MRELBP outperforms all of the variations of LBP in the texture classification problem. Ageing skin texture has many edges and corner components but young skin has more flat and non-uniform structure. Therefore, the MRELBP descriptor reflects the characteristics of skin ageing and can be used for the skin features.

Figure 3.4 illustrates the MRELBP descriptor and its components. It is constructed by concatenating the histograms of three different descriptors, namely RELBP_CI, RELBP_NI and RELBP_RD descriptors which are defined as follows:

1) Center pixel representation:

$$MRELBP\_CI(x_c) = s\big(\varphi\big(X_{c,w}\big) - \mu_w\big) \tag{3.2}$$

where x_c is a center pixel and φ is a patch Median filter, $X_{c,w}$ denotes the local patch of size $w \times w$ centered at the center pixel x_c, $s()$ is the sign function and μ_w denotes the mean of $\varphi(X_{c,w})$ over the whole image.

2) Neighbour representation:

$$MRELBP_{NI_{r,p}}\,(xc) = \sum_{n=0}^{p-1} s\big(\varphi\big(X_{r,p,w_r,n}\,\big) - \mu_{r,p,w_r}\big)2^n,$$

$$\mu_{r,p,w_r} = \frac{1}{p}\sum_{n=0}^{p-1} \varphi\big(X_{r,p,w_r,n}\big) \tag{3.3}$$

where $X_{r,p,w_r,n}$ denotes a patch of size $w_r \times w_r$ centered on $X_{r,p,w_r,n}$

3) Radial difference representation:

$$MRELBP\_RD_{\mathrm{r,r-1,p},w_r,w_{r-1}}(x_c)$$

$$= \sum_{n=0}^{p-1} s(\varphi(X_{r,p,w_r,n}) - \varphi(X_{r-1,p,w_{r-1},n}))2^n \qquad (3.4)$$

where $X_{r,p,w_r,n}$ and $X_{r-1,p,w_{r-1},n}$ denote the patches centred at the neighbouring pixels $x_{\mathrm{r,p,n}}$ and $x_{\mathrm{r-1,p,n}}$, respectively.

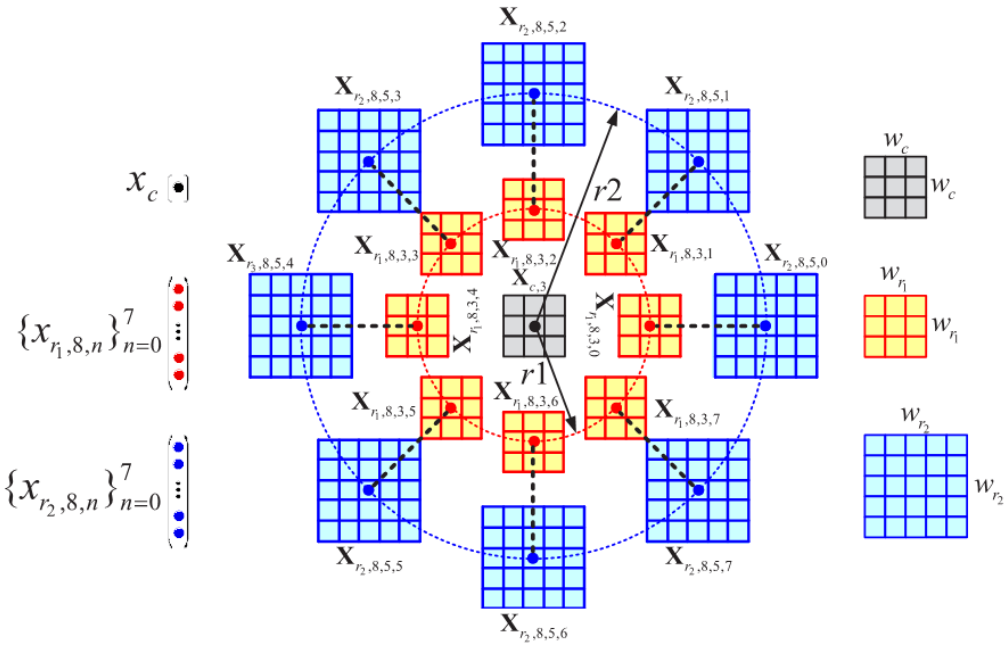


Figure 3.4: MRELBP Descriptor and Its Components [69]

3.4 Biologically Inspired Features (BIF)

Another feature descriptor which was successfully used in previous studies in age estimation field [11] [53] [54] [55] is BIF that tries to model visual processing in the cortex as a stack of increasingly sophisticated layers. Riesenhuber and Poggio [30] proposed a new set of features derived from a feed-forward network of the primary

visual object recognition pathway, called the "HMAX" model. The Model consists of two different types of layers: S units' neurons (simple) and C units' neurons (complex). Specifically, S1 layer is constructed by convolving a set of Gabor filters over the grayscale image at four orientations and 16 scales. Then each pair of adjacent S1 unit is combined together to generate 8 bands of units for each direction. In the next layer, this is called C1, the maximum values within local patches and across the scales within a band is computed. Therefore C1 feature includes 8 bands and 4 orientations. The Gabor functions which are used in the S1 units are in the following form:

$$G(x,y) = exp\left(-\frac{(X^2 + \gamma^2 Y^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}X\right) \tag{3.5}$$

where $X = xcos\theta + ysin\theta$ and $Y = -xsin\theta + ycos\theta$ are the rotations of the Gabor filters with angle θ which varies between 0 and π, σ is the effective width, λ is the wavelength and s is the filter size.

Figure 3.5 shows the overview of a BIF system: the gray-level input image is first analyzed by an array of S1 units at four different orientations and 16 scales. At the next C1 layer, the image is subsampled through a local MAX (M) pooling operation over a neighborhood of S1 units in both space and scale, but with the same preferred orientation. In the next stage, S2 units are essentially RBF units, each having a different preferred stimulus. Note that S2 units are tiled across all positions and scales. A MAX pooling operation is performed over S2 units with the same selectivity to yield the C2 unit responses.

3.5 Convolutional Neural Networks

Convolutional neural networks have been one of the most influential innovations in the field of computer vision. They have performed a lot better than traditional computer vision and have produced state-of-the-art results. These neural networks have proven to be successful in many different real-life case studies and applications, such as image and signal classification, object detection, segmentation, face recognition.

3.5.1 Neural Networks

Neural networks are a type of machine learning models which are designed to operate similar to biological neurons and human nervous system. These models are used to recognize complex patterns and relationships that exist within a labeled dataset. They have following properties:

- The core architecture of a Neural Network model is comprised of a large number of simple processing nodes called Neurons which are interconnected and organized in different layers.
- An individual node in a layer is connected to several other nodes in the previous and the next layer. The inputs form one layer are received and processed to generate the output which is passed to the next layer.
- The first layer of this architecture is often named as input layer which accepts the inputs, the last layer is named as the output layer which produces the output and every other layer between input and output layer is named is hidden layers.

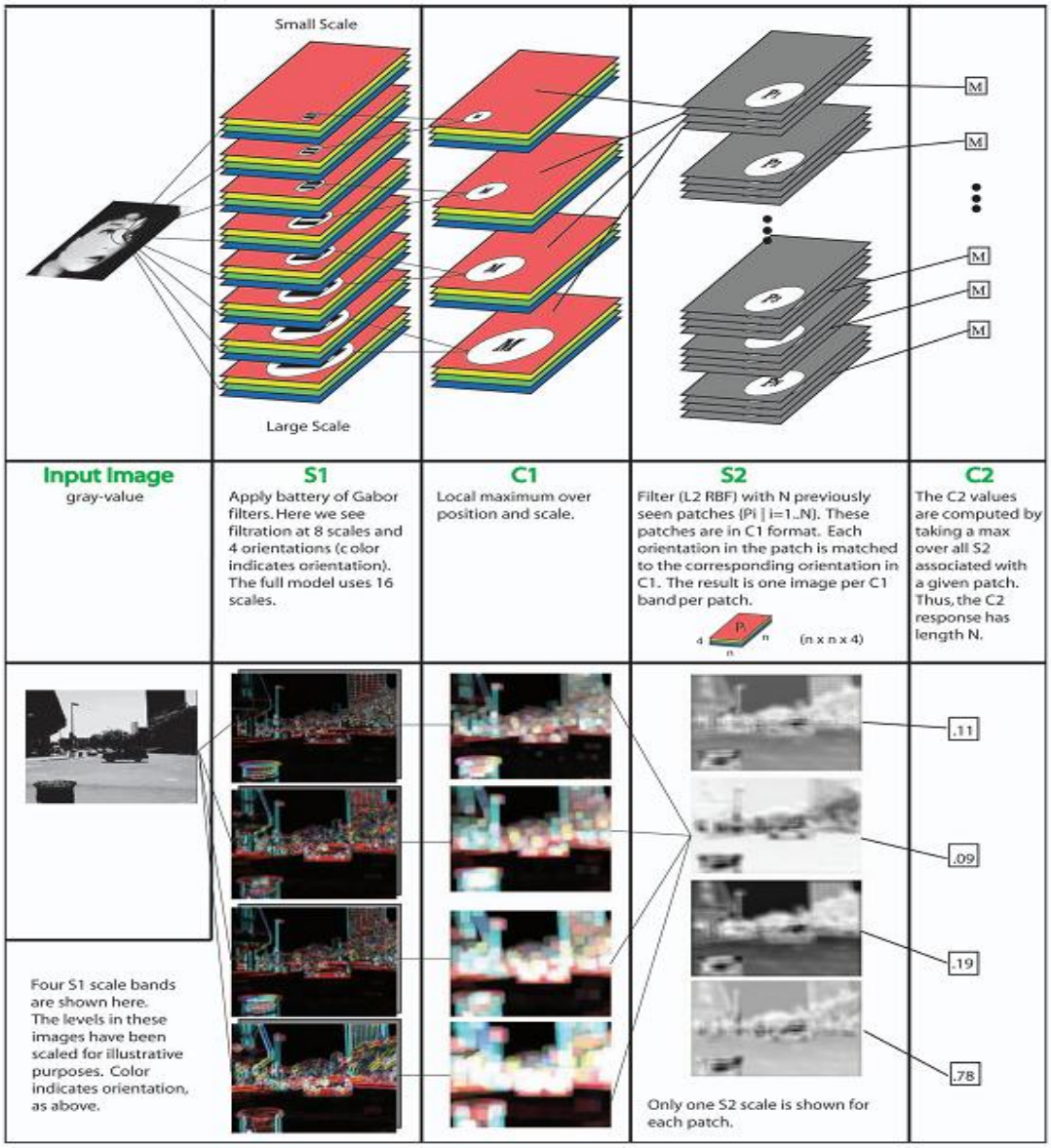


Figure 3.5: Overview of a BIF System [54]

3.5.2 Key Concepts in a Neural Network

The key concepts in a Neural Network are given below.

- Neuron: A Neuron is a single processing unit of a Neural Network which are connected to different other neurons in the network. These connections represent inputs and output from a neuron. To each of its connections, the neuron assigns a "weight" (W) which signifies the importance the input and adds a bias (b) term (Figure 3.6).

- Activation Functions: The activation functions are used to apply non-linear transformation on input to map it to output. The aim of activation functions is to predict the right class of the target variable based on the input combination of variables. Some of the popular activation functions are Relu, Sigmoid, and TanH.
- Forward Propagation: Neural Network model goes through the process called forward propagation in which it passes the computed activation outputs in the forward direction.
- Error Computation: The neural network learns by improving the values of weights and bias. The model computes the error in the predicted output in the final layer which is then used to make small adjustments the weights and bias. The adjustments are made such that the total error is minimized. Loss function measures the error in the final layer and cost function measures the total error of the network.
- Backward Propagation: Neural Network model undergoes the process called backpropagation in which the error is passed to backward layers so that those layers can also improve the associated values of weights and bias. It uses the algorithm called Gradient Descent in which the error is minimized and optimal values of weights and bias are obtained. This weights and bias adjustment is done by computing the derivative of error, derivative of weights, bias and subtracting them from the original values.

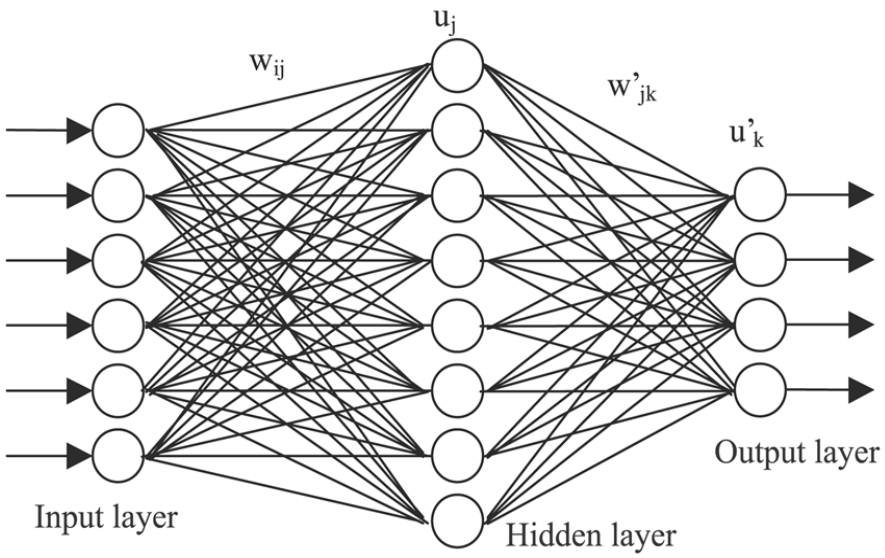


Figure 3.6: NN's Architecture with One Hidden Layer

3.5.3 Convolutional Neural Network Components

A simple ConvNet is a sequence of layers, and every layer of a ConvNet transforms one volume of activations to another through a differentiable function. There are three main types of layers to build ConvNet architectures: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. In order to construct a full ConvNet architecture, a stack of theses layers is used (Figure 3.7).

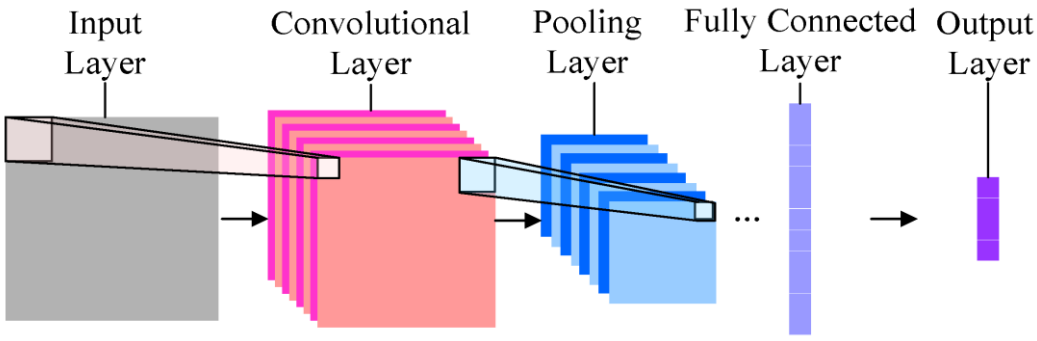
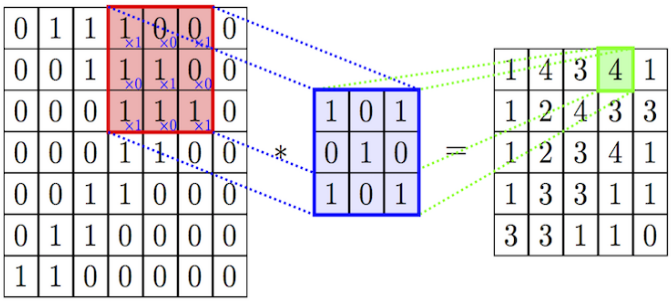


Figure 3.7: Convolutional Neural Network [67]

3.5.4 Convolution Layer

The convolution layer computes the output of neurons that are connected to local regions or receptive fields in the input, each computing a dot product between their

weights and a small receptive field to which they are connected to in the input volume. Each computation leads to extraction of a feature map from the input image. Figure 3.8 shows an overview of the convolution operator and the result of applying it with two separate kernels, over an image to act as an edge detector.



(a)

(b)

Figure 3.8: (a) Convolution Operation (b) Example of Applying Convolution Operation

3.5.5 ReLU Layer

The rectified linear units (ReLU) layer commonly follows the convolution layer. The addition of the ReLU layer allows the neural network to account for non-linear

relationships, i.e. the ReLU layer allows the ConvNets to account for situations in which the relationship between the pixel value inputs and the ConvNet output is not linear. Note that the convolution operation is a linear one. The output in the feature map is just the result of multiplying the weights of a given filter by the pixel values of the input and adding them up:

$$y = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n \tag{3.6}$$

where w is a weight value and x is a pixel value. Figure 3.9 shows the ReLU function which takes a value x and returns 0 if x is negative and x, if x is positive.

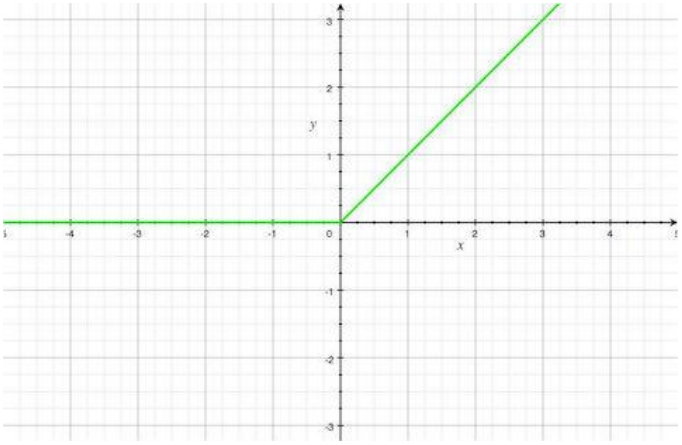


Figure 3.9: ReLU , f(x) = max(0,x)

3.5.6 Pooling Layer

The pooling layer also contributes towards the ability of the ConvNet to locate features regardless of where they are in the image. In particular, the pooling layer makes the ConvNet less sensitive to small changes in the location of a feature, i.e. it gives the ConvNet the property of translational invariance in that the output of the pooling layer remains the same even when a feature is moved a little. Pooling also reduces the size of the feature map, thus simplifying computation in later layers.

One of the techniques of subsampling is max pooling that takes the largest value from the window of the image currently covered by the kernel. For example in Figure 3.10, a max-pooling layer of size 2 x 2 is applied on a 4x4l image.

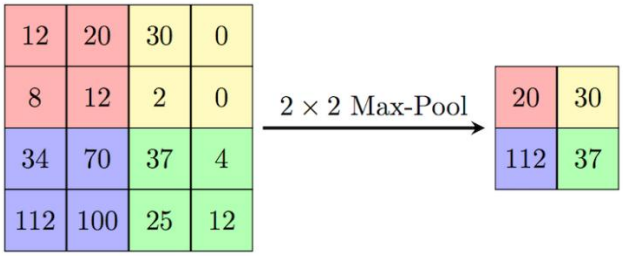


Figure 3.10: Max-Pooling [67]

The objective of the fully connected layer is to flatten the high-level features that are learned by convolutional layers and combining all the features. It passes the flattened output to the output layer where you use a softmax classifier or a sigmoid to predict the input class label.

3.5.7 The Fully-Connected and Loss Layers

The fully-connected layer is where the final "decision" is made. At this layer, the ConvNet returns the probability that an object in a photo is of a certain type. The fully-connected layer has at least 3 parts - an input layer, a hidden layer, and an output layer. The input layer is the output of the preceding layer, which is just an array of values.

Figure 3.11 shows a fully-connected network for classifying the input image into different classes. Following the fully-connected layer is the loss layer, which manages the adjustments of weights across the network. Before the training of the network begins, the weights in the convolution and fully-connected layers are given random values. Then during training, the loss layer continually checks the fully-

connected layer's guesses against the actual values with the goal of minimizing the difference between the guess and the real value as much as possible. The loss layer does this by adjusting the weights in both the convolution and fully-connected layers.

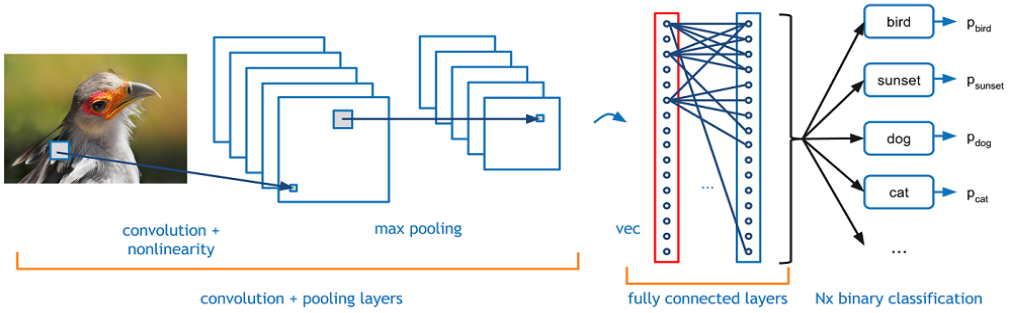


Figure 3.11: Fully-Connected and Loss Layers [67]

3.6 Visualizing Convolutional Neural Networks

In general, the more convolution steps we have, the more complicated features our network will be able to learn to recognize. For example, in Image Classification a ConvNet may learn to detect edges from raw pixels in the first layer, and then use the edges to detect simple shapes in the second layer, and then use these shapes to deter higher-level features, such as facial shapes in higher layers. This is demonstrated in Figure 3.12.

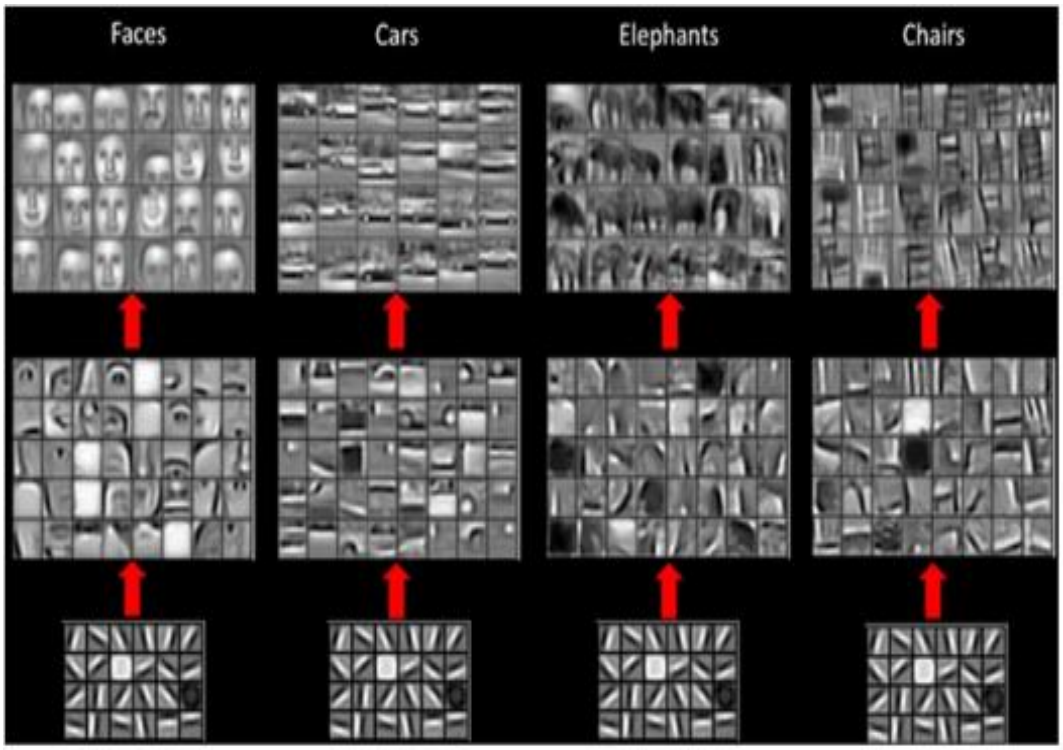


Figure 3.12: Learned Features by Different Layers of a CNN

Chapter 4

AGE ESTIMATION DATABASES

4.1 Introduction

Precise age and age-group estimation requires a database with good quality facial images at different ages. It is hard to collect a large ageing database with a series of chronometric images from an individual. Age and age-group estimation often uses databases collected previously and published. Brief descriptions of these databases are found in [33]. Table 4.1 gives the summary of some of the ageing databases available.

Table 4.1: Summary of facial ageing databases

Database	# of subjects	Database size	Age range (years)
FG-NET	82	1002	0–69
MORPH-II	13,618	55,134	27–68
Yamaha gender and age (YGA)	1600	8000	0–93
Waseda human-computer interaction	26,222	5500	3–85
AI & R Asian	17	34	22–61
Burt's Caucasian Face database	----	147	20–62
Lotus Hill Research Institute (LHI) database	----	50,000	9–89
Human and object interaction processing	300	306,600	15–64
Iranian face database	616	3600	2–85
Gallagher's Web-Collected database	–	28,231	0–66
Ni's Web-Collected database	–	219,892	1–80
BERC database	95	5910	3–83
3D Morphable database	438	–	–

4.2 Morph-II Ageing Database

Morph-II [66] is an ageing database which contains more than 55,000 face images of about 13,000 subjects. These images are captured during 2003 to 2007. Age ranges in this database vary from 16 to 77 years. The age distribution is shown in Figure 4.1 (a) and some examples of different subjects are shown in Figure 4.1 (b). Figure 4.2 shows age progression for two different subjects with different ancestry, a white male and an African-American Female.

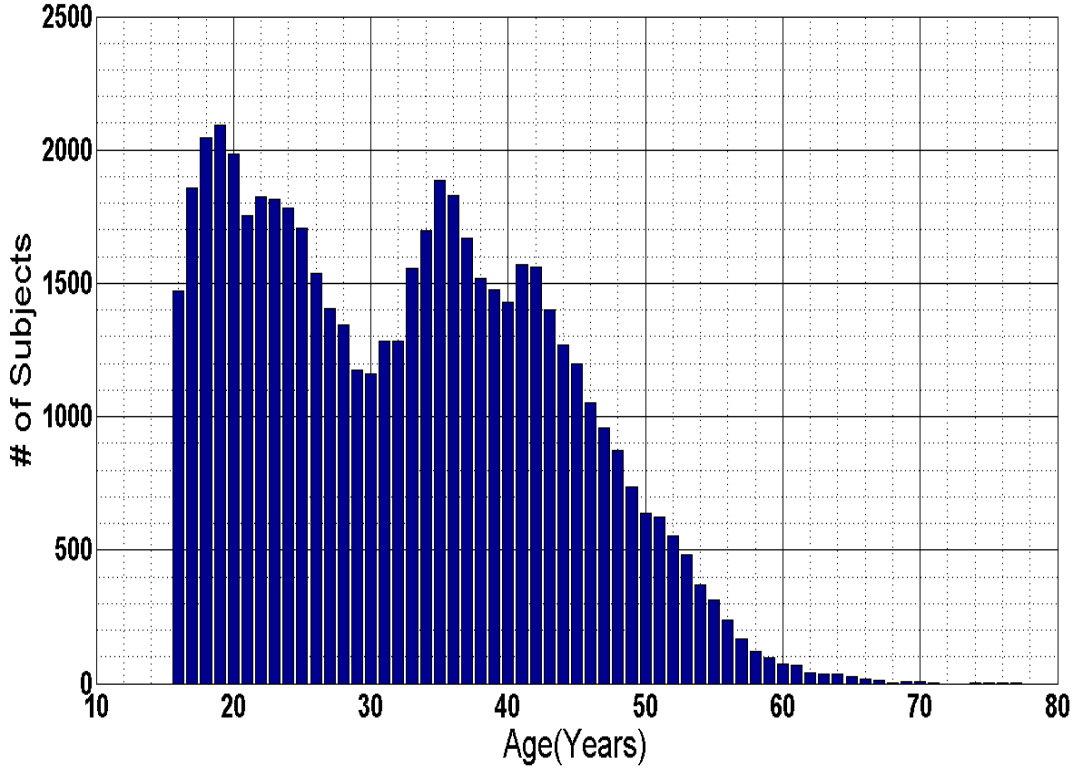
Table 4.2 shows the number of facial images in this release by decade-of-life; Table 4.3 shows the distribution of images by gender and ancestry and Table 4.4 shows the number of additional images that exist from the initial facial image.

Table 4.2: The age and gender information of samples from Morph-II dataset

	<20	**20-29**	**30-39**	**40-49**	**>50**	**Total**
Male	6638	14016	12448	10062	3482	46646
Female	831	2309	2909	1988	453	8490
Total	7469	16325	15357	12050	3935	55136

Table 4.3: Number of facial images by gender and ancestry in Morph-II dataset

	African	**European**	**Asian**	**Hispanic**	**"Other"**	**Total**
Male	36,832	7,961	141	1,667	44	46,646
Female	5,757	2,598	13	102	19	8,490
Total	42,589	10,559	154	1,769	63	55,136



(a)

(b)

Figure 4.1: Age Distributions of Morph-II Dataset. (b) Example of Different Subjects in Morph-II Dataset

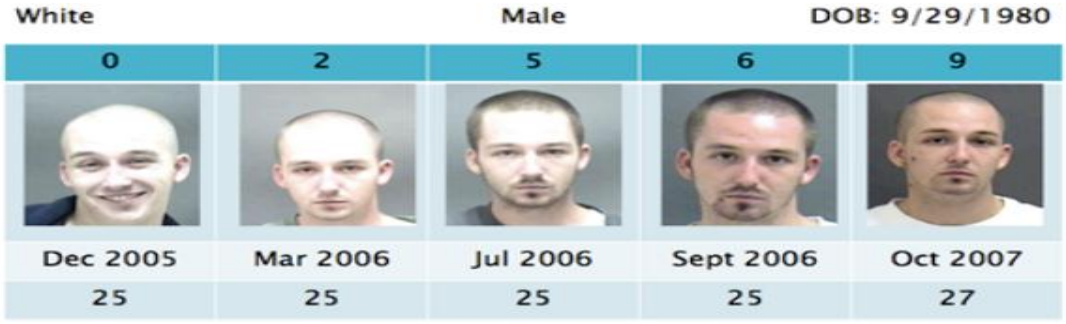
Table 4.4: Number of additional images per subject in Morph-II dataset

	1	**2**	**3**	**4**	**5+**
Male	373	2,350	3,606	1,975	3,155
Female	85	478	712	352	532

4.2.1 Age Estimation Evaluation Protocol for Morph-II

K-fold cross-validation is the basic form of cross-validation. Other forms of cross-validation are just but special cases of k-fold cross-validation or involve repeated rounds of k-fold validation. In k-fold cross-validation, original data is randomly split into k equal subsets. Then, k iterations of training and validation are performed such that in every iteration, a different fold of data is reserved for validation while the remaining k-1 are used to learn a model. The estimated error is the mean of all validation errors. Standard deviation of these errors can be used to approximate the confidence range of the estimate. The main advantage of k-fold cross-validation is that eventually all samples will used for both learning and validating a model.

In order to use the database in a systematic way, we follow the BEFIT protocol described in [67] to split the database into five non-overlapped subsets randomly with a very important criterion: all of the sample images from a specific subject should be in one and only one unique fold each time. The age distributions of different folds are shown in Figure 4.3.



(a)

(b)

Figure 4.2: Image Progression for (a) White Male and (b) African-American Female

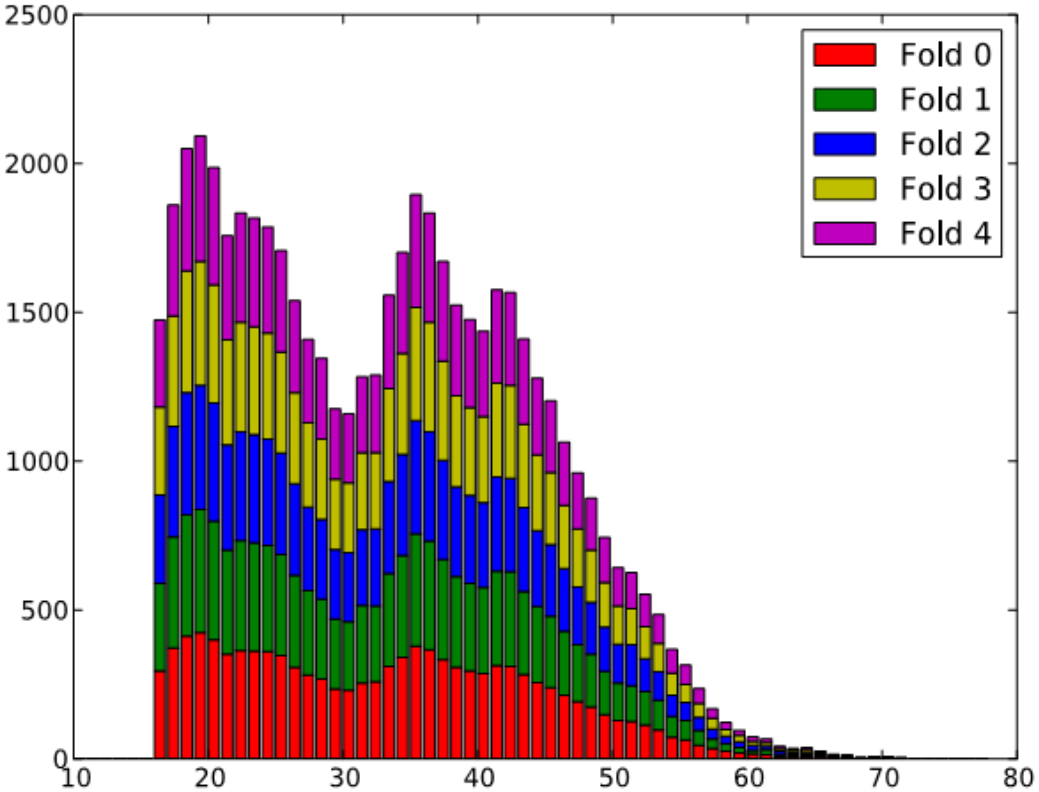


Figure 4.3: Distribution of Morph-II Database Images over Age in the Individual Folds [67]

4.3 FG-NET Ageing Database

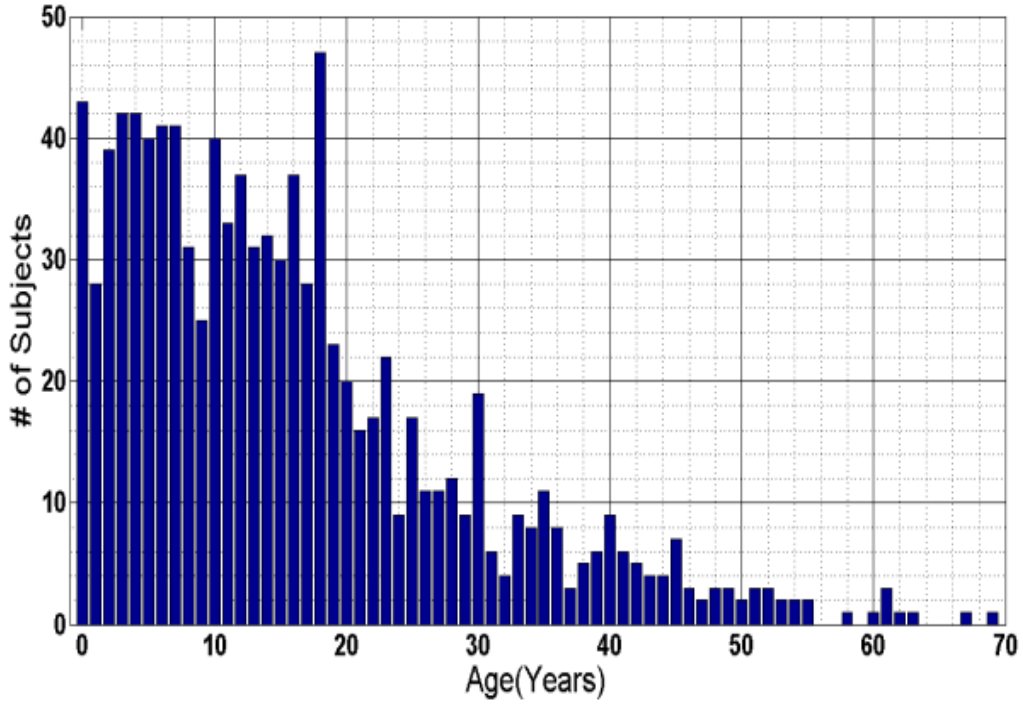
In 2004, the FG-NET ageing database (The Face and Gesture recognition NETwork) was released in order to help researchers who try to understand the effect of ageing on facial appearance [68]. After that, FG-NET was used in many studies in different domains such as in age estimation, age-invariant face recognition, gender classification and age progression [69].

The FG-NET consists of 1002 images from 82 different people with ages varying between 0 to 69 years old. The subjects' ages are not equally distributed and most of the subjects' ages are less than 40 years old in the database. These images were collected by scanning personal photographs of subjects so they display considerable variability in resolution, image sharpness, and illumination in combination with face viewpoint and expression variation. Occlusions in the form of spectacles, facial hair and hats also exist in a number of images.

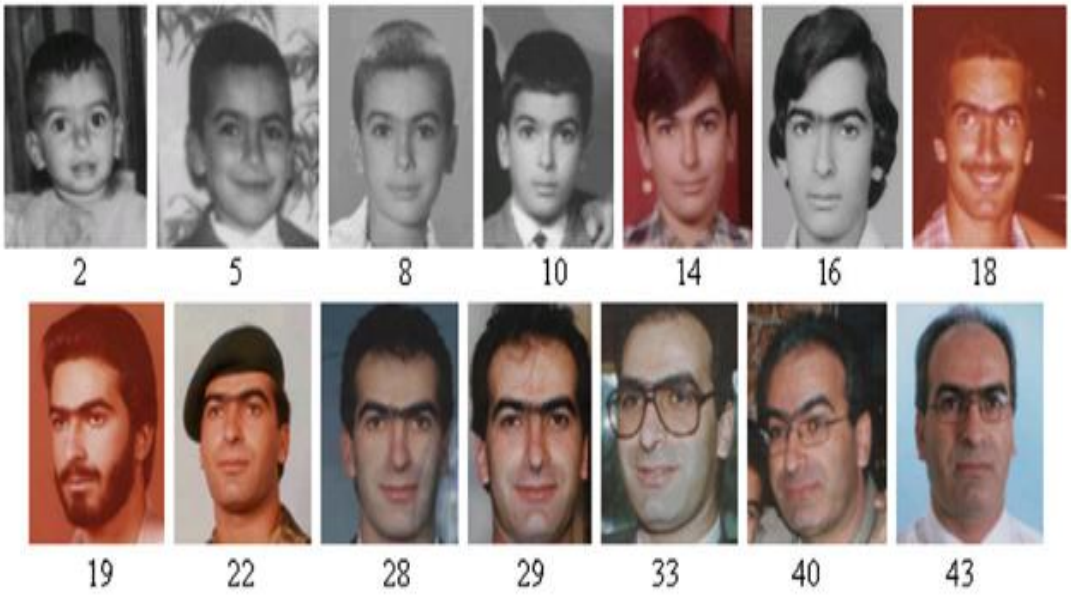
Each image in the dataset was annotated with 68 landmark points located at key positions and also a semantic description of each image was recorded. In particular information about the age, gender, expression, pose, image quality and appearance of occlusions (i.e. moustaches, beards, hats or spectacles) were recorded. The age distribution is shown in Figure 4.4 (a) and the ageing faces example of one subject is shown in Figure 4.4 (b).

Most researchers reporting results using the FG-NET-AD adopted the Leave One Person Out (LOPO) approach where for each of the 82 subjects in the database, an age estimator is trained using images of the remaining 81 subjects and the results are averaged over the 82 trials. Given the small number of images available in the FG-

NET this is the optimum and recommended approach. Table 4.5 summarized the features of the utilized datasets.



(a)



(b)

Figure 4.4: (a) Age Distributions of FG-NET Dataset, (b) Example of Ageing Faces of One Subject

Table 4.5: Summary of the utilized age datasets

Property	FG-NET	Morph-II
Collection Environment	Personal photos	Real-world conditions
Collection Era	Unknown	2003-2007
Digital, Paper Scan	Unknown	Digital & Scan
Image Size	Mostly 400×500	200×400 & 400×480
Compression	JPEG	JPEG
Annotation	68 points	None
Frontal Pose	Uncontrolled	Uncontrolled
Source	Public	Public
Total Used in Experiment	1002	55
Population Trend, age range	>50% are ages 0-13 [0-69]	>92% are ages 16-49 [16-77]
Evaluation Protocol	LOPO	5-fold cross validation
Subset Disjoint	Yes	Yes

4.4 Metrics

The most common metric for accuracy evaluation of the age estimators is the Mean Average Error (MAE) which is also used in this study. MAE computes the average age deviation error in absolute terms as follows:

$$\mathrm{MAE} = \sum_{i=1}^{N} \frac{|\hat{\alpha}_i - \alpha_i|}{N} \tag{4.1}$$

where $\hat{\alpha}_i$ is the computed age of the i subject, α_i is its actual annotated age and N is the total number of test subjects.

Another common metric is the cumulative score (CS) which quantitatively shows the evaluation of age estimation approach by a curve. The CS accuracy at the error ε is computed as follows:

$$CS(\theta) = \frac{N_{\varepsilon \leq \theta}}{N} \times 100\% \tag{4.2}$$

where $N_{\varepsilon \leq \theta}$ is the number of test samples in which their estimated age error θ is not less than ε.

Chapter 5

PROPOSED METHOD I: FEATURE-LEVEL AND SCORE-LEVEL FUSION OF HAND-CRAFTED DESCRIPTORS

5.1 Introduction

According to the success rate of using score-level fusion in multimodal biometric recognition systems, it is believed that accuracy can be improved when the information of two different types of classifiers are consolidated. Due to large inter-class similarity and intra-class variation, we need to perform fusion in two levels with different kinds of feature descriptors: feature-level fusion of local descriptor and score-level fusion of appearance based descriptor and the obtained results from the previous level fusion.

Different feature descriptors were investigated to select the most powerful ones for descriptor selection. Inspired by recent face recognition and texture classification works [64][65][70][71] in the computer vision community, we used Biologically Inspired Features (BIF), Median Robust Extended LBP (MRELBP) and Histogram of Oriented Gradients (HOG) for local feature description and used Kernel Fisher Analysis (KFA) for the appearance based recognition system in which the results of recent research [72] shows that it outperforms LDA and PCA methods.

Figure 5.1 illustrates the overall schematic of the first proposed method. For each input image, two different kinds of features are computed. Feature-level fusion of HOG, BIF and MRELBP is performed by concatenating their feature vectors together. We also compute KFA from the original image and compare the similarity of these two feature vectors with all of the feature vectors in the training set and then select the minimum one for each method. After normalizing these obtained scores, we add them together and make decision by using k-Nearest Neighbour (KNN) classifier.

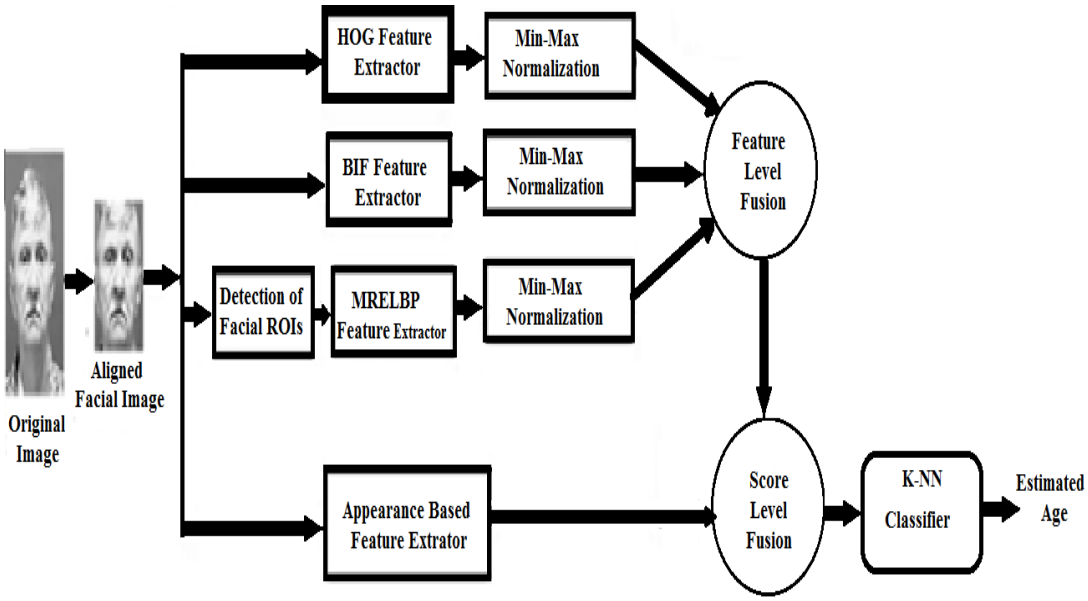


Figure 5.1: Schematic of the First Proposed Method

5.2 Preprocessing

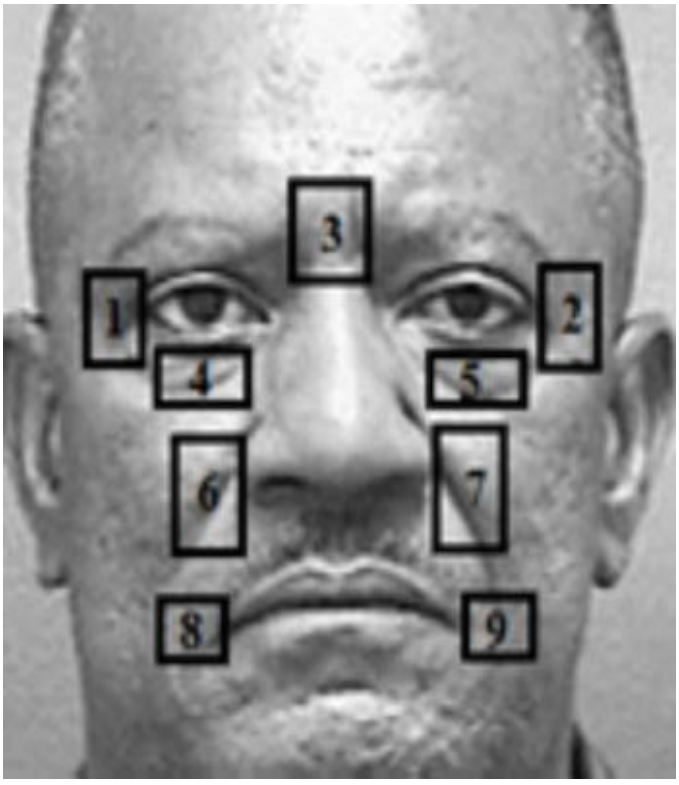
Preprocessing is an essential step in image processing systems in order to enhance the quality of the input images. In this study, face images undergo a series of preprocessing steps in order to extract the region of interest of the facial image that is used for age estimation. We used face detection method described in [73] for the detection of facial images. Since Morph-II contains some tattoo images, we ignore

them in our experiments. After this reduction, the number of face images used as a training set and test set is 55244. After face detection step, facial image will be aligned by using geometric transformation such that the eyes have been symmetrically placed at 25% and 75% of the aligned image. The aligned images are resized to 60 × 60 pixels. Then the fiducial points of face images that are determined by Active Shape Model (ASM) technique [74] and local neighborhood area around these landmarks will be cropped.

5.2.1 Facial Patches

The facial wrinkles provide discriminative information for age estimation, and these facial wrinkles have been considered by many researchers [3][23] [60][61][62]. Facial fiducial points play an important role for robust face recognition systems, especially for uncontrolled environment. In order to build pose robust face descriptors, we should detect precise facial landmarks. By extracting region enclosed to these fiducial points, we can achieve some patches which have the same semantics for different subject images. For each facial image, firstly, we determine some facial fiducial points by using ASM [74] . Figure 5.2 (a) shows the locations of the founded points in a sample image. Afterwards, we crop the face patch by using these landmarks as shown in Figure 5.2 (b) and extract MRELBP features separately from each of these nine regions and concatenate them together. With the help of this process, we obtain the facial patches that include the wrinkles on the facial images. These patches are then used to estimate the age of the person from his or her facial image.

(a)



(b)

Figure 5.2: (a) The Locations of the Landmark Points, (b) Wrinkle Regions Which Are Used for Textural Feature Extraction

5.3 Experimental Settings

In order to measure the effectiveness of the selected visual descriptors and find the optimal experimental setting, we have investigated the effect of different values for each feature detection algorithm parameters.

For HOG algorithm, namely $HOG_{C,B}$, C is the patch size and B is the number of histogram bins. The optimal parameters have been obtained by testing different values for patch size and number of histogram bins with 5-fold cross-validation for Morph-II dataset and LOPO for FG-NET dataset. Figure 5.3 shows the MAE of HOG visual descriptors for Morph-II dataset and the best result is achieved when $C_x = C_y = 19$ and B= 12 in which MAE= 4.16 years. The best result for FG-NET dataset is achieved when $C_x = C_y = 13$ and B= 12 in which MAE= 5.29 years.

	Cx=Cy	number of histogram bins (B)											
		6	7	8	9	10	11	12	13	14	15	16	17
Grid Size (C)	7	5.41	5.15	5.05	4.92	4.90	4.84	4.85	4.83	4.77	4.80	4.82	4.79
	8	5.10	4.90	4.90	4.80	4.74	4.76	4.70	4.68	4.62	4.64	4.60	4.65
	9	4.90	4.72	4.67	4.53	4.55	4.52	4.50	4.45	4.45	4.42	4.43	4.47
	10	4.84	4.65	4.60	4.49	4.50	4.46	4.48	4.43	4.40	4.40	4.40	4.44
	11	4.63	4.53	4.49	4.41	4.40	4.34	4.37	4.31	4.31	4.31	4.37	4.32
	12	4.62	4.51	4.46	4.42	4.39	4.39	4.35	4.30	4.33	4.32	4.34	4.34
	13	4.55	4.44	4.37	4.35	4.37	4.32	4.28	4.26	4.29	4.26	4.31	4.29
	14	4.49	4.39	4.31	4.32	4.34	4.30	4.27	4.25	4.24	4.27	4.29	4.28
	15	4.41	4.29	4.29	4.28	4.27	4.25	4.24	4.19	4.22	4.24	4.28	4.27
	16	4.42	4.31	4.39	4.26	4.33	4.27	4.25	4.23	4.26	4.31	4.31	4.29
	17	4.34	4.29	4.27	4.27	4.26	4.20	4.21	4.21	4.21	4.29	4.28	4.26
	18	4.27	4.22	4.25	4.25	4.22	4.18	4.18	4.20	4.20	4.22	4.21	4.24
	19	4.28	4.20	4.19	4.21	4.19	4.17	**4.16**	4.18	4.18	4.19	4.20	4.23
	20	4.44	4.31	4.27	4.35	4.35	4.34	4.33	4.35	4.34	4.39	4.38	4.42

Figure 5.3: Results for $HOG_{C,B}$ with Varying Patch Size C_x and C_y and Number of Bins B (columns). The Bolded Value Indicates the Optimal Result

The optimal parameters for MRELBP descriptor are found by performing the same procedure. In the case of $MRELBP^{u2}_{R,P}$, the investigation has been performed by testing

different values for the number of sampled neighbors (P) and radius (R), limiting the number of neighbors to either 8 or 16. The best result for Morph-II dataset was achieved by P=8 and R=2 and it was MAE=4.85 years. On the other hand, the best result and for FG-Net dataset is achieved by P=8 and R=1 and it is MAE=5.68 years.

5.4 Feature-level and Score-level Fusion

In order to enhance the system performance and utilize the distinct characteristics of different feature extractors, comprehensive experiments of different fusion level of various pairs of selected visual descriptors have been performed. We tested different type of feature descriptors for both local feature based and appearance based classifiers. We used HOG, CLBP, LBP-HF, Haralick features and MRELBP as local feature descriptors and tested LDA and KFA as appearance-based feature extractors. In each method, we used cross validation approach to find the optimal parameter settings. After feature extraction, in the case of one-descriptor and feature-level fusion, we used a linear SVR for age estimation and in the case of score-level fusion we used k-Nearest Neighbors (k-NN) classifier.

Feature-level fusion causes increasing in the dimensionality of feature space and this expansion causes the curse of dimensionality problem and overfitting. Therefore it is not possible to combine all of the features in this level. Table 5.1 shows the most successful combinations. Feature-level fusion has been obtained by simply concatenating the individually extracted features of separated descriptors. The concatenation of BIF, MRELBP and HOG descriptors achieved the best results in feature level fusion. This combined feature has the advantage of mixing local appearance-based features and textural ones.

In order to solve the aforementioned problem and also to benefit from the other feature descriptors, we performed score-level fusion. According to the success rate of using score-level fusion in multimodal biometric recognition systems, it is believed that accuracy can be improved when the information of two different types of classification systems are consolidated.

The distance between each test sample and its nearest training samples is assumed to be the score of that test sample in the corresponding classification/regression system. These scores are normalized by Min-Max normalization method as follows:

$$x' = \frac{x - Min(x)}{Max(x) - Min(x)} \tag{5.1}$$

where x is the raw score, $Max(x)$ and $Min(x)$ are the maximum and minimum values of the raw scores respectively and x' is the normalized score. After score normalization, the multimodal score vector $\langle x1, x2 \rangle$ is constructed, with $x1$ and $x2$ corresponding to the normalized scores of two different systems. The next step is fusion at the matching score level. The score vector is combined by Sum rule-based fusion method [75] to generate a single scalar score which is then used to make the final decision as follows:

$$fs = w_1 x_1 + w_2 x_2 \tag{5.2}$$

The notation w_i stands for the weight which is assigned to one of the two systems and we decided to use equal weights in all of the experiments in order to give equal chance to each feature extractor.

In order to show that score-level fusion can improve the accuracy of age estimation system, we combined different types of feature descriptor scores. The experimental results show that, in all the cases, the score-level fusion causes meaningful

improvement in MAE. When we perform both feature-level and score-level fusion together, the MAE is reduced to 3.89 years which is better than all the other methods presented in Table 5.1. Therefore, it can be claimed that the performance of our method outperforms the other local and appearance-based methods and all the possible combination pairs of these methods with feature-level and score-level fusion.

5.5 Conclusion

A novel age estimation method based on different level of information fusion is proposed in this chapter. Biologically inspired features (BIF) and texture-based features such as MRELBP and HOG were involved during the first level of information fusion (Feature-level) process and then the obtained concatenated feature vector was fused with appearance-based method of KFA in the second level of information fusion (score-level). Compared with the state-of-the-art methods, our proposed approach obtained comparable MAE on MORPH-II and FG-NET datasets. The experimental results demonstrate that the feature-level and score-level fusion of local features and appearance-based features provide a higher accuracy than the other algorithms.

Table 5.1: Summary of results for different level fusion of various feature extractors that achieve the optimal value (×: feature-level fusion, *: score-level fusion)

Experiment #	Feature Extraction Method				Fusion method	MORPH-II MAE	FG-NET MAE
	MRELBP	HOG	BIF	KFA			
1	×				N/A	4.93	5.68
2		×			N/A	4.47	5.89
3			×		N/A	4.31	4.61
4				×	N/A	5.21	4.84
5	×	×			Feature-level	4.25	5.59
6	×		×		Feature-level	4.34	4.47
7		×	×		Feature-level	4.29	4.38
8	×	×	×		Feature-level	4.09	4.16
9	×		×	×	Feature-level	4.43	5.01
10			×	×	Feature-level	4.37	4.89
11	×	×		×	Feature-level	4.31	4.91
12	×	×	×	×	Feature-level	4.48	5.36
13	×	×			Score-level	4.30	5.33
14	×		×		Score-level	4.36	4.49
15		×	×		Score-level	4.24	4.52
16	×	×	×	×	Score-level	4.88	5.16
17	×			×	Score-level	4.67	5.08
18		×		×	Score-level	4.33	4.89
19			×	×	Score-level	4.21	4.32
20	×	×		*	Feature-level & score-level	4.19	4.28
21	×		×	*	Feature-level & score-level	4.22	4.19
22		×	×	*	Feature-level & score-level	4.11	4.13
23	×	×	×	*	Feature-level & score-level	**3.89**	**4.06**

Chapter 6

PROPOSED METHOD II: MULTI-SCALE LEARNED FEATURES WITH CNN AND HAND-CRAFTED DESCRIPTORS

6.1 Introduction

In this thesis, we propose a new age estimation system with two substructures for estimating the accurate age from facial image by exploiting multi-stage learned features these features with age-related handcrafted features such as wrinkle and skin features from a generic feature extractor, a trained CNN, and combine. Due to large inter-class similarity and intra-class variation, we perform fusion in two levels with different kinds of feature descriptors: feature-level fusion of hand-crafted descriptors and score-level fusion of global learned feature descriptors and perform aggregation on the obtained results of them as the final estimated age.

The overall schematic of the second proposed method is illustrated in Figure 6.1. For each input image, two different kinds of features are computed. The facial wrinkle features, the skin features and the facial component shape features are extracted and combined together by feature-level fusion of Gabor filters, MRELBP and BIF descriptors, respectively. These features are first normalized by Z-score method and then concatenated in order to perform feature-level fusion. BIF features are computed from the whole aligned and cropped face image, while MRELBP and

Gabor filters are computed from several facial regions. Finally the obtained feature vector is used to compute the expected age as explained in Algorithm-1.

We also exploit different learned features from several CNN layers (the last three Convolution + ReLU layers) which are trained on Morph-II or FG-NET datasets for age estimation. These learned features are combined with each other using score-level fusion. In the final step of the second proposed method, namely age aggregation, the obtained ages are combined with each other by weighted average to predict the test sample's age.

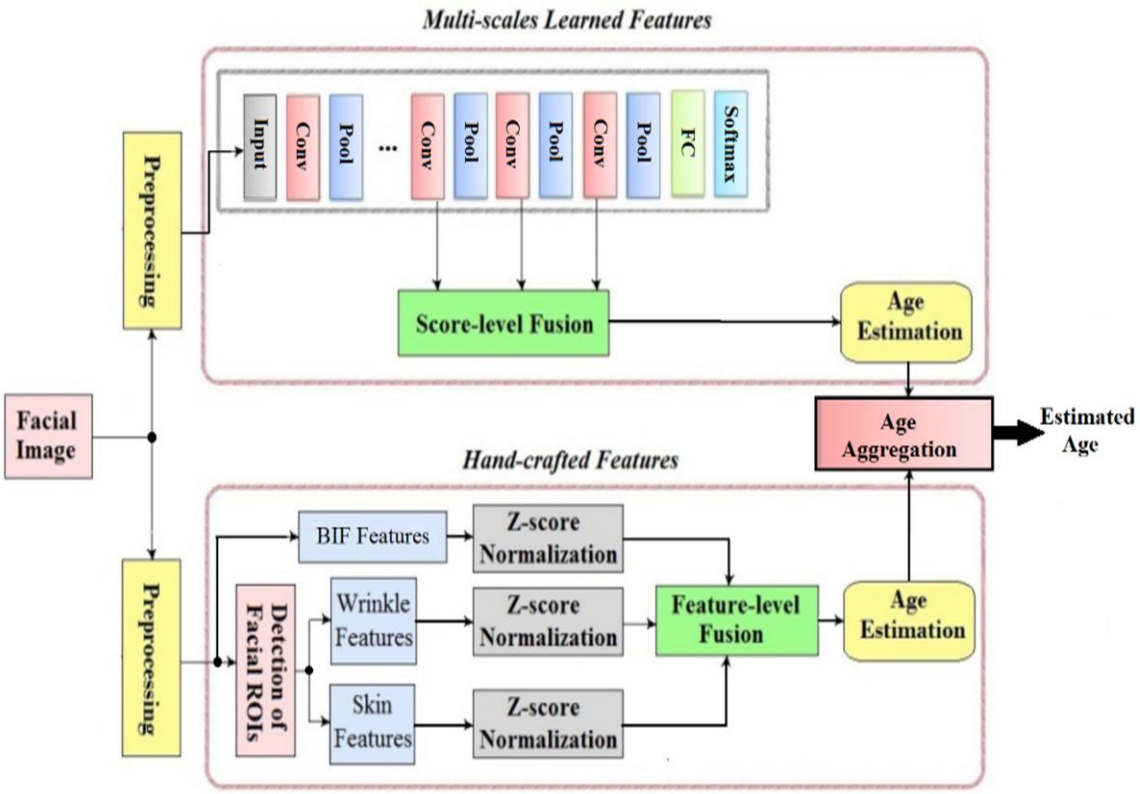


Figure 6.1: Schematic of the Second Proposed Method

6.2 Handcrafted Feature Descriptors

Feature extraction is an important step in an image classification system. The selection of discriminative feature descriptors is a critical decision and it affects the whole system performance. Therefore, we have chosen different types of well-known and efficient local and global feature extractors that have been successfully used in age estimation and the other applications like face recognition with their discrimination ability, computational efficiency, compact feature space size, and robustness to alignment and illumination variance.

In order to build pose robust face descriptors, we should detect precise facial landmarks. By extracting the region enclosed to these fiducial points, we obtain some patches which have the same semantics for different subject images. In order to extract wrinkle features effectively, we computed the strength and quantity of wrinkles in several facial patches by using Gabor filter [63] set on the direction of the facial muscles on these patches.

Facial skin also provides a lot of discriminative information for estimating the age. Unlike wrinkles, skin ageing appears randomly and non-uniformly for each facial part. Ageing of the facial skin mostly appears in the form of freckles and it reduces the amount of collagen that plays a role in reflecting light, and is distributed non-uniformly on the face. As a result, the overall tone of the facial skin becomes non-uniform. Since ageing skin has smooth variations, we selected a feature extractor capable of analyzing microstructures of skin texture. For this purpose, we used MRELBP [64] descriptor which detects very fine details such as edges, lines, spots and flat areas in a computationally efficient manner. Ageing skin texture has many

edges and corner components but young skin has more flat and non-uniform structure. Therefore, the MRELBP descriptor reflects the characteristics of skin ageing and can be used for the skin features.

BIF is one of the common feature descriptor which was successfully used in previous studies in age estimation field [11] [53] [54] [55]. This descriptor tries to model visual processing in the cortex as a stack of increasingly sophisticated layers. The Model consists of two different types of layers: S units' neurons (simple) and C units' neurons (complex). Specifically, S1 layer is constructed by convolving a set of Gabor filters over the grayscale image at four orientations and 16 scales. Then each pair of adjacent S1 unit is combined together to generate 8 bands of units for each direction. In the next layer, this is called C1, the maximum values within local patches and across the scales within a band is computed. Therefore C1 feature includes 8 bands and 4 orientations. The Gabor functions which are used in the S1 units are in the following form:

$$G(x,y) = exp\left(-\frac{(X^2 + \gamma^2 Y^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}X\right) \qquad (6.1)$$

where $X = xcos\theta + ysin\theta$ and $Y = -xsin\theta + ycos\theta$ are the rotations of the Gabor filters with angle θ which varies between 0 and π, σ is the effective width, λ is the wavelength and s is the filter size.

6.3 Experimental Settings and Results

6.3.1 Pre-processing

Preprocessing is an essential step in image processing systems in order to enhance the quality of the input images. In this study, face images undergo a series of preprocessing steps in order to extract the region of interest of the facial image that is used for age estimation. We used face detection method described in [73] and

cropped the detected facial images. Then the location of the eyes and fiducial points of face images are determined by ASM technique [74]. Afterwards, the facial image is aligned by using normalized rotation and geometric transformation such that the eyes have been aligned and placed horizontally from left and right sides of the aligned image by 25% for left and right eyes, respectively. For the handcrafted features, the aligned images are resized to 60 × 60 pixels and converted to grayscale. Then local neighborhood areas around the detected landmarks are used for cropping the facial patches. For the learned features (CNN), all the aligned images are used in RGB format and resized to 256 × 256 and five different cropped size of 227× 227 are fed to the network.

6.3.2 Handcrafted Feature Settings

Handcrafted features consist of wrinkle, skin and shape-based BIF features that are combined using feature-level fusion. The details related to these features are given below.

6.3.2.1 Wrinkle Features

For each facial image, firstly, we determine some facial fiducial points by using ASM [73]. Figure 6.2(a) shows the locations of the detected points in a sample image. Afterwards, we crop the face patches by using these landmarks as shown in Figure 6.2(b) and extract wrinkle features and skin features from all of these nine patches. These discriminative age features are combined with each other in a feature-level fusion manner.

In order to extract wrinkle features from each patch, we followed the method mentioned in [61]. In each wrinkle area, the dominant direction of wrinkles can be estimated, and the Gabor filter sets in each wrinkle area are selected corresponding to this direction. For instance, horizontal filters are assigned to patch 3 of Figure 6.2(b)

(Filters 3,9,15 and 21) while all filters are applied on parts 1 and 2, since the wrinkle around the eyes lies in all orientations. Figure 7.2 illustrates 24 Gabor filters with 4 scales and 6 orientations computed by using (1) with K=6, the scale number of S=4, the lower average frequency of U_l=0.05 and the upper average frequency of U_h=0.4. For more detail refer to [61].

The following filters are applied to the remaining patches: for patch 4 the filters used are: 1, 2, 3, 7, 8, 9, 13, 14, 15, 19, 20 and 21; for patch 5 the applied filters are: 3, 4, 5, 9, 10, 11, 15, 16, 17, 21, 22 and 23; for patches 6 and 8 the employed filters are: 1, 2, 6, 7, 8, 12, 13, 14, 18, 19, 20 and 24; for patch 7 and 9 the used filters are: 4, 5, 6, 10, 11, 12, 16, 17, 18, 22, 23 and 24.

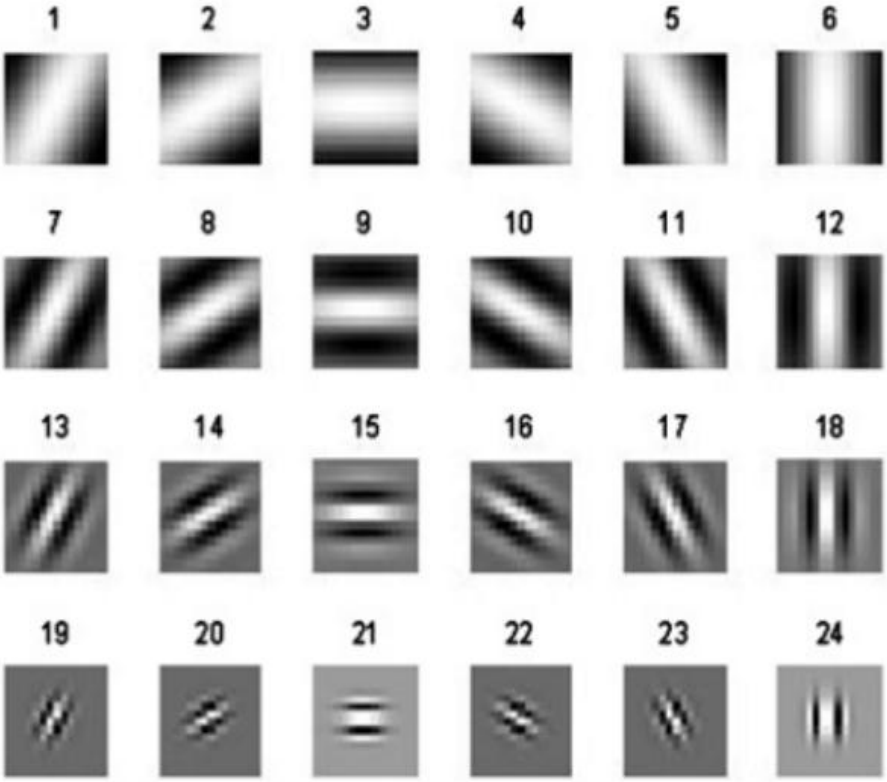


Figure 6.2: Gabor Filter Sets According to the Direction of Facial Wrinkles.

Mean and variance of the magnitude response of the Gabor filter in each wrinkle area are computed, in order to extract the wrinkle features. Therefore the wrinkle feature dimension is 246. These statistical values represent both the strength and quantity of wrinkles. The extracted features are then stored in a feature vector and are used in Algorithm 2 as the input and the MAE is obtained as 4.63 for Morph-II and 4.38 for FG-NET databases.

6.3.2.2 Skin Features

In order to extract skin features, we compute the three components of $\mathrm{MRELBP}_{\mathrm{R,P}}^{\mathrm{riu2}}$, namely,$MRELBP\_CI$, $MRELBP\_NI_{R,P}^{riu2}$ and $MRELBP\_RD_{R,P}^{riu2}$ by using Eq. (3.2) till Eq. (3.4) for all of the nine patches in Figure 5.2(b). Afterwards, the features of these nine patches are concatenated. We use rotation invariant uniform encoding scheme in all cases. The optimal parameter investigation has been performed by cross-validation test of different values for the number of sampled neighbors (P) and radius (R), limiting the number of neighbors to either 8 or 16 and radius to 1, 2 or 3. The best result by using Algorithm-1 for Morph-II was achieved by P=8 and R=2 as MAE=4.85 years and for FG-NET was obtained by P=16 and R=2 as MAE=4.67. The skin feature's dimension is 1377.

6.3.2.3 BIF Features

Unlike the other two handcrafted feature descriptors, the BIF features are extracted from the whole aligned facial image which is obtained by preprocessing step. For BIF features extraction, we extracted C1, S1, C2 and S2 features and tried to find the most discriminative subset of them. We found that the S2 and C2 features cannot work well for age estimation. Therefore we used only the S1 and C1 units. This is consistent with Guo et al. [38] findings. In applying the Gabor filters to S1 units, we found that a smaller size of 5 × 5 can characterize the ageing effects on faces better

than the other sizes. We used the same parameter settings explained in [29], and summarize six parameter details in Table 6.1. In C1 layer, we utilized the maximum operation "MAX" as the pooling filter. For each image, the outputs from the C1 units are concatenated to construct feature vector for each image. This feature vector dimension is 6,976. In [29] the authors showed that the simple PCA method can work well for dimensionality reduction of BIF features without increasing the MAE too much. Since we perform feature-level fusion on handcrafted features, the size of BIF feature is critical. Therefore, we used PCA method to reduce the feature vector's dimension. We tested different sizes for feature dimensions from 100 to 1000 and compute the MAE for each of them. The results showed that the best MAE for feature dimensions equals to 900 for both datasets. The obtained MAE result by using Algorithm-1 for Morph-II is 4.31 years and for FG-NET, it is 4.82 years.

Table 6.1: Parameter settings for BIF feature descriptor

$C1$ Layer			**$S1$ Layer**		
Scale band S	Pool. grid	Overlap Δ_s	filter size s	Gabor σ	Gabor λ
Band 1	6×6	3	5×5 7×7	2.0 2.8	2.5 3.5
Band 2	8×8	4	9×9 11×11	3.6 4.5	4.6 5.6
Band 3	10×10	5	13×13 15×15	5.4 6.3	6.8 7.9
Band 4	12×12	6	17×17 19×19	7.3 8.2	9.1 10.3
Band 5	14×14	7	21×21 23×23	9.2 10.2	11.5 12.7
Band 6	16×16	8	25×25 27×27	11.3 12.3	14.4 15.4
Band 7	18×18	9	29×29 31×31	13.4 14.6	16.8 18.2
Band 8	20×20	10	33×33 35×35	15.8 17	19.7 21.2

Algorithm-1 Expected Age Computation

Input:

Trainset $X_{train} = \{(X_{tr}^i, y_{tr}^i)\}, i = 1, \dots, N_{train}$

Testset $X_{test} = \{(X_{te}^i, y_{te}^i)\}, i = 1, \dots, N_{test}$

FeatureSet FS^m

Output:

$Expected\_age\_vector$, MAE

1: $F\_m^{X\_tr^i} \leftarrow$ Compute FS^m $\forall$ i $= 1, \dots, N_{train}$
2: For j= 1 To N_{test}
3: $F\_m^{X\_te^j} = \langle feat\_m_1^j, \dots, feat\_m_{Sm}^j \rangle$
4: total=0
5: For i = 1 To N_{train}
6: d= *compute_distance*($F_m{}^{X_{tr}{}^i}$, $F_m{}^{X_{te}{}^j}$)
7: S_i = 1/d
8: total=total +S_i
9: Expected_age_vector$_j$= $\sum_{i=1}^{N_{train}} \frac{S_i}{total} \times y\_tr^i$
10: $Error$+= *abs(*Expected_age_vector$_j$- y_te^j)
11: MAE= $Error$ / N_{test}

6.3.3 Feature-level Fusion of Hand-crafted Features

The facial wrinkle features and skin features contain valuable information about the person's age, especially for mature and elders, but these features are not sufficient to accurately estimate the age. Therefore, in order to enhance the system performance and utilize the distinct characteristics of different feature extractors, feature-level fusion of skin, wrinkle and BIF descriptors is performed.

On the other hand, each of the aforementioned extracted features is normalized by using Z-score method as follows:

$$Z = \frac{X - \mu}{\sigma} \tag{6.2}$$

where μ is the mean value of the feature and σ is its standard deviation. After normalization, skin, wrinkle and BIF feature vectors are concatenated in order to perform feature-level fusion. The size of fused feature vector is 246 + 1377 + 900 =

2523. The resulting vector is used as the input FS^m of Algorithm-1 to compute the expected age. For computing the expected age, the similarity between a test sample and all the training set is considered. After applying Algorithm-1 on the obtained feature-level fusion of skin, wrinkle and BIF feature vectors, the resulting MAE for Morph-II is 3.87 and for FG-NET, it is 3.99 years.

6.4 CNN-based Learned Features

CNN-based learned features and score-level fusion of multi-stage CNN learned features are described below in detail.

6.4.1 CNN Architecture for Age Estimation

Our proposed CNN architecture for age estimation is illustrated in Figure 6.3. We assumed that the output of each component (convolution, ReLU, pooling, dropout and fully connected) of the proposed model is treated as a separate layer. Therefore our model has 22 layers. Our CNN model consists of only five convolutional layers and three fully-connected layers. All the input images are in RGB format and resized to 256 × 256 and a cropped size of 227× 227 is fed to the network. For all the five subsequent convolutional layers, the kennel sizes are 3 × 3 with stride set to 1 with padding option so that the output has the same size as the input. Each convolutional layer followed by an S-shaped rectified linear unit (SReLU) [76] and a max pooling layer with kernel size equal to 2 × 2 with strides 2 and a local response normalization layer. Following the 5 convolutional layers, there are 3 fully connected layers containing 512,512 and 54 neurons respectively that each one is followed by a ReLU and a dropout layer. The final fully connected layer output shows the probability for each age class-label. Age estimation can be considered as a discrete classification with multiple discrete value labels. For Morph-II dataset, it is a one dimensional regression problem with the age being sampled from a continuous range

between 16 and 77 and for FG-NET dataset it is between 0 and 70. For computing the expected age from CNN, instead of using simple softmax, we multiply each softmax output probabilities by the corresponding class year label and add them together as follows:

$$Expected\ Age = \sum_{i=L}^{H} p_i \times year\_label_i \tag{6.3}$$

where L, H are lower and upper bound of subject's age, p_i is softmax output probability and $year\_label_i$ is an integer age year value of output neuron i in the last layer of CNN.

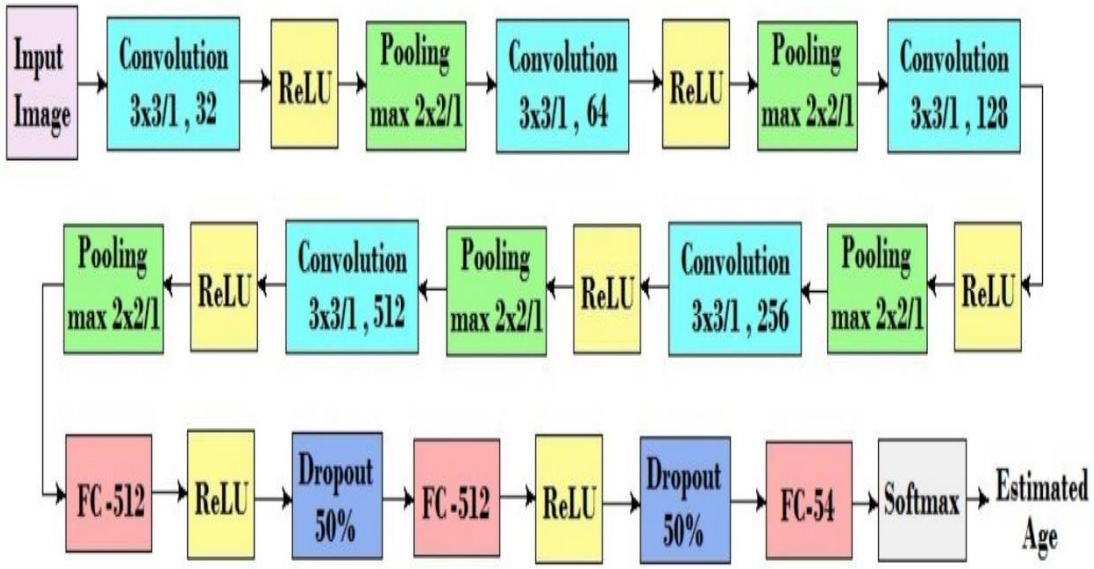


Figure 6.3: CNN Architecture

In Table 6.2, we summarize the details of the CNN model. As mentioned before, in order to be fair in comparison with the state-of-the-art methods which only use the Morph-II dataset images, we did not use pre-trained models for initializing the network and all the weights and biases in all layers are initialized with random values. In order to avoid overfitting, dropout layers with a dropout ratio equal to 0.5

are used. Additionally we performed data augmentation by taking five different cropped segments of size 227×227 pixels from the original 256×256 input image (including the pixels at four corners and the center pixel separately for each segment) and randomly mirror it in each forward-backward training pass.

Table 6.2: The details of CNN architecture

Layers	Type	No. of Neurons	Kernel size
1	Convolution	32×224×224	3×3
3	Max-pooling	32×112×112	2×2
4	Convolution	64×112×112	3×3
6	Max-pooling	64×56×56	2×2
7	Convolution	128×56×56	3×3
9	Max-pooling	128×28×28	2×2
10	Convolution	256×28×28	3×3
12	Max-pooling	256×14×14	2×2
13	Convolution	512×14×14	3×3
15	Max-pooling	512×7×7	2×2
16-19-22	Fully Connected	512×512×54	-

Our proposed CNN architecture was trained through the standard backpropagation technique with a batch size of 32. In order to obtain optimum performance, the other learning parameters are set as follows: to prevent overfitting of training data, the regularization (λ) is set to 0.1, momentum parameters which adjust the speed of learning during training is set to 0.9, and learning rate that control the convergence of the training data is set to 0.001 and linearly changed according to the mean-squared error values in each ten iteration. The training was performed 50 epochs rounds. The CNN MAE for Morph-II is 3.64 and for FG-NET, it is 3.78 years.

6.4.2 Score-level Fusion of Multi-stage CNN Learned Features

In order to show that score-level fusion can improve the accuracy of age estimation system, we combined different learned features from different layers of CNN. Figure 6.4 shows the age estimation accuracy of systems which used different layers'

features. In Figure 6.4, only convolution layers' performance (after applying ReLU on them) are illustrated since the accuracy of each pooling layer, in all cases, is less than the corresponding convolution layers. For the purpose of investigating different combinations of feature layers and finding the best one experimentally, features of the last layer are considered as of necessary and intermediate layer features are added layer-by-layer, one at a time, in a backward fashion until no improvement is observed in age estimation accuracy. This greedy approach for both Morph-II and FG-NET dataset, ignores the features of layers closer to the input layer and as a result, the features of the layers 8, 11 and 14 are selected for score-level fusion.

The experimental results show that combining the intermediate features with last layer features with score-level fusion causes meaningful improvement in MAE. The obtained MAE for Morph-II is 3.34 and for FG-NET, it is 3.57 years.

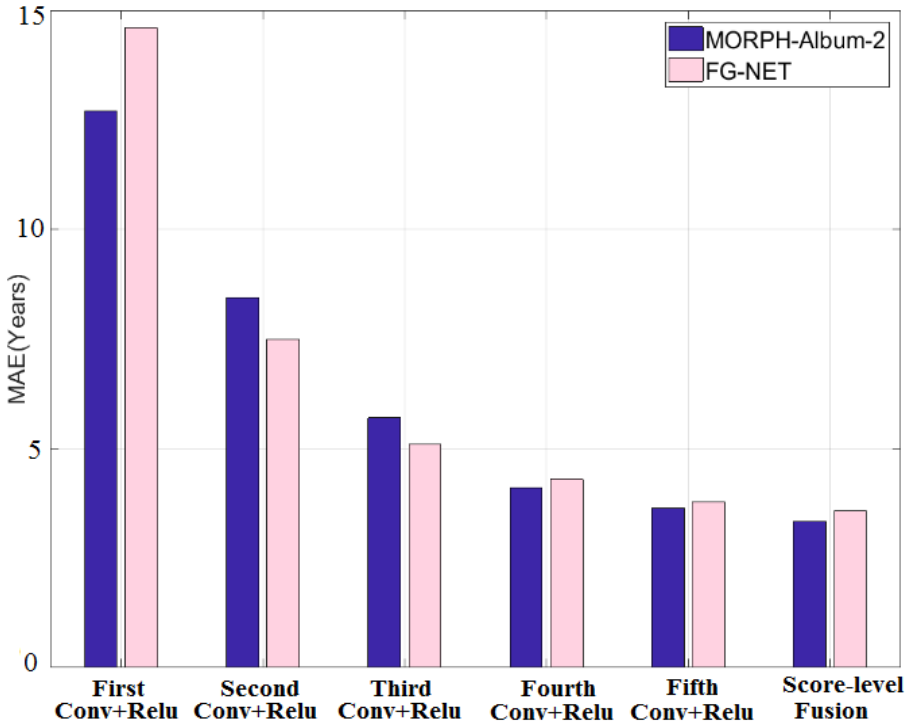


Figure 6.4: Different CNN Layers' Features Performance

6.5 Age Aggregation

In order to utilize both handcrafted and learned features, the estimated ages by these two approaches are combined together to produce the estimated age of the proposed system. This combination is performed by simple weighted average rule as follows:

$$estimated\ age = w_1 \times age_1 + w_2 \times age_2 \quad (6.4)$$

where age_1 is the estimated age by feature-level fusion of handcrafted features, age_2 is the estimated age by score-level fusion of learned features and w_1 , w_2 ϵ [0,1] are weights for each of these two methods ($w_1 + w_2 = 1$). In order to find the optimal weights, we used 5 fold cross-validation method; and for each iteration, we use 30 percent of training data as the validation data. The best settings are $w_1 = 0.24$ and $w_2 = 0.76$ for Morph-II; and $w_1 = 0.37$ and $w_2 = 0.63$ for FG-NET. By using these optimal weights for the corresponding dataset, the final MAE of the second proposed system for Morph-II is 3.17 and for FG-NET, it is 3.29 years which are better than all the other handcrafted-based and learned-based methods presented in Table 6.3. Additionally, in Figure 6.5 and Figure 6.6, we showed the MAE separately for each age in order to investigate the relation between handcrafted and learned features. These results show that for some subjects, the hand-crafted features play a complementary role with learned features. Furthermore, it is clear that generic features extracted from CNN are enhanced by combining them with domain-specific features.

Additionally, The CS curves of the second proposed method compared with state-of-the-art methods on Morph-II and FG-NET datasets are illustrated in Figure 6.7 and Figure 6.8 which outperform all of the other compared methods in all different levels of error.

Table 6.3: The experimental results summary of the 2nd proposed method

Method	Features	MAE	
		Morph-II	FG-NET
Handcrafted Features	Wrinkle features	4.63	4.38
	Skin features	4.85	4.67
	BIF features	4.31	4.82
	Feature-level fusion	3.87	3.99
Learned Features	CNN	3.64	3.78
	Score-level fusion	3.34	3.57
Age Aggregation		3.17	3.29

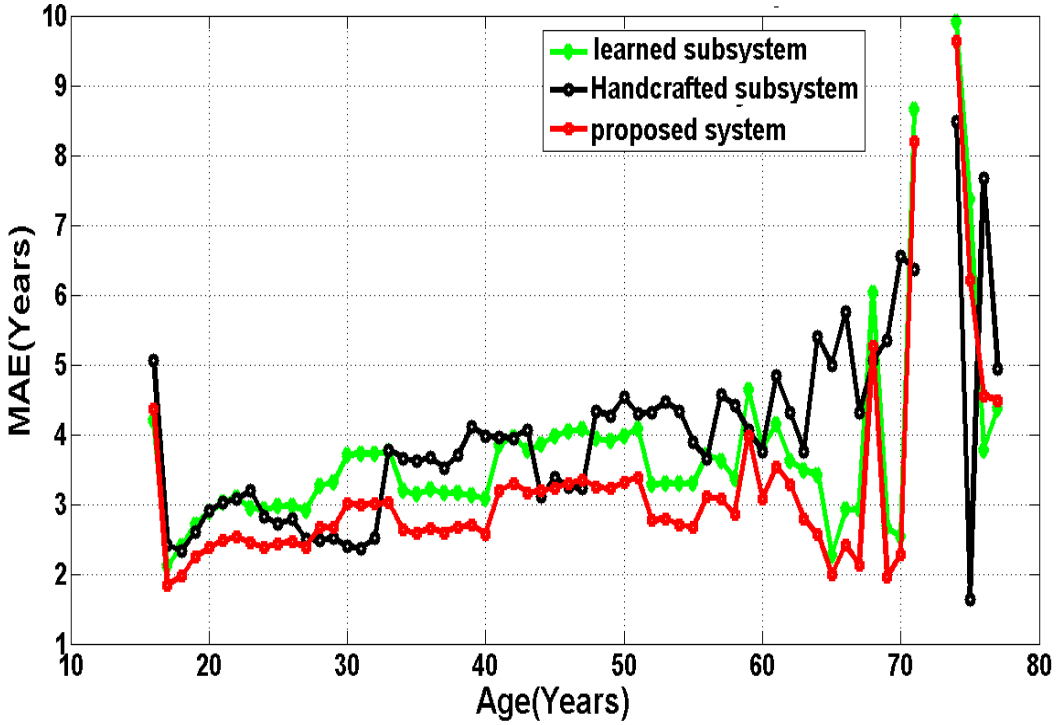


Figure 6.5: MAE of the Second Proposed System and Its Subsystems for Morph-II Dataset

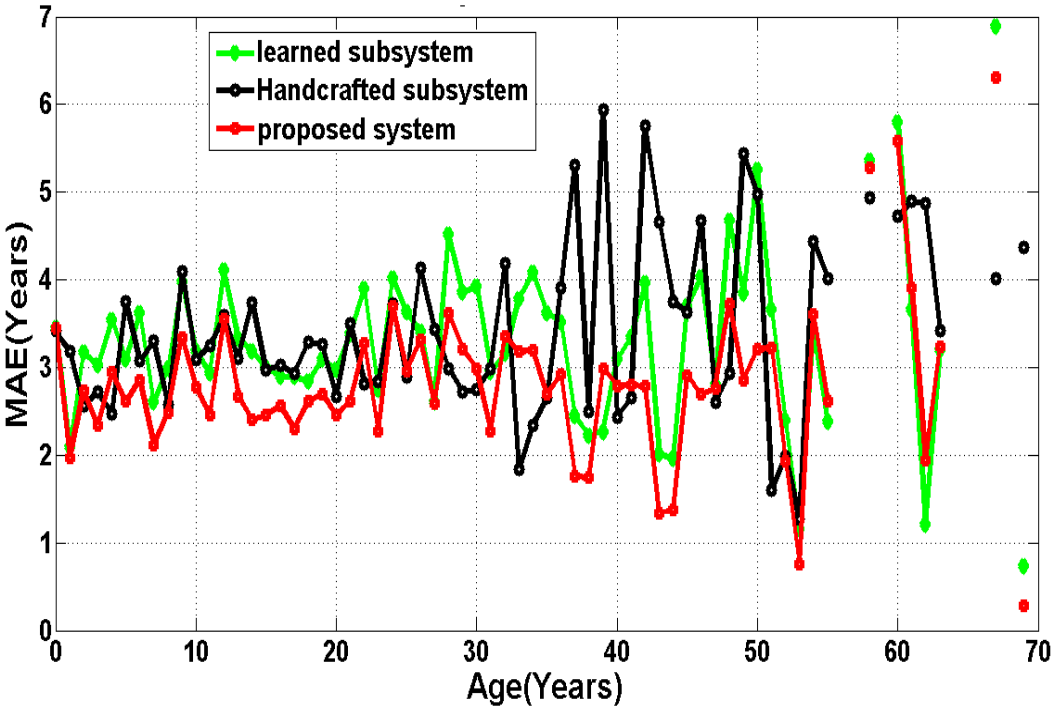


Figure 6.6: MAE of the Second Proposed System and Its Subsystems for FG-NET Dataset

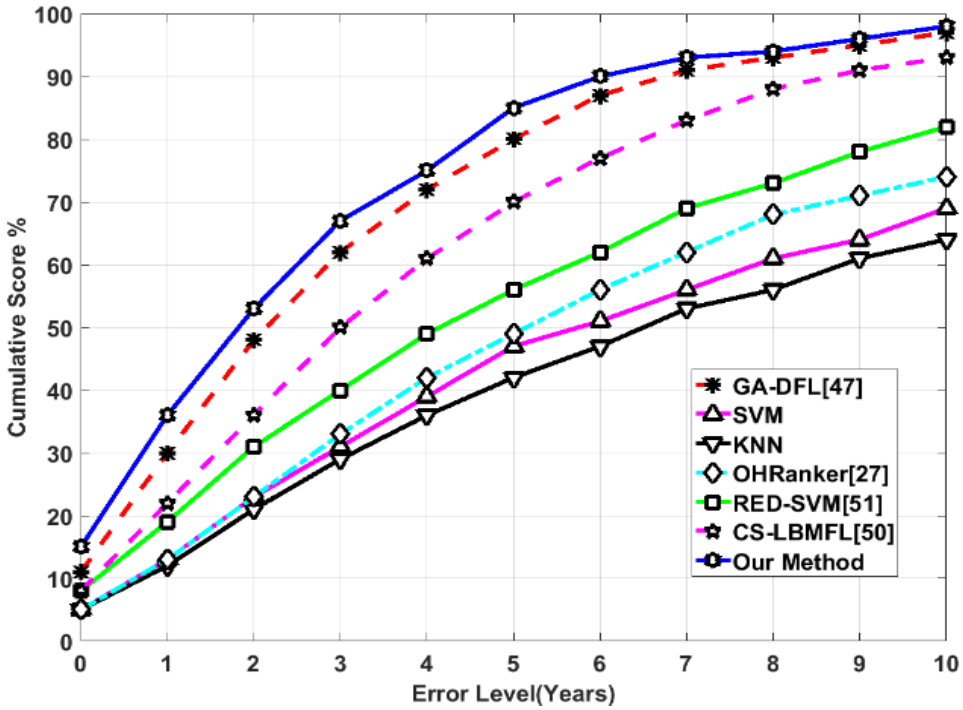


Figure 6.7: The CS Curves of the Second Proposed System Compared with State-of-the-Art Methods on Morph-II Dataset

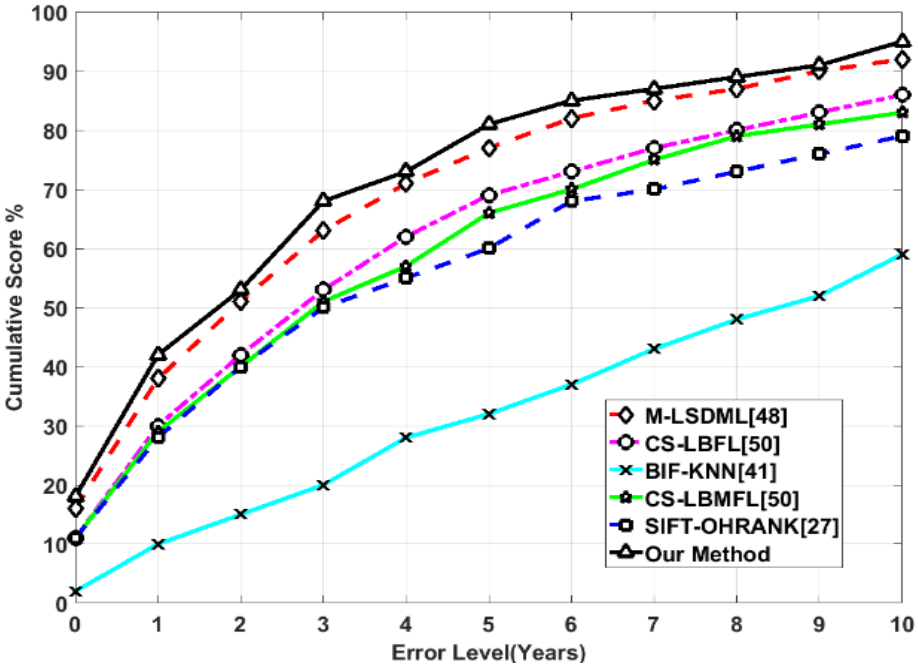


Figure 6.8: The CS Curves of the Second Proposed System Compared with State-of-the-Art Methods on FG-NET Dataset

6.6 Conclusion

In this chapter, we propose a new age estimation system which exploits multi-stage features from a generic feature extractor, a trained convolutional neural network (CNN), and precisely combined these features with a selection of age-related handcrafted features. This method utilizes a decision-level fusion of estimated ages by two different approaches; the first one uses feature-level fusion of different handcrafted local feature descriptors for wrinkle, skin and facial component while the second one uses score-level fusion of different feature layers of a CNN for its age estimation. Experiments on the publicly available Morph-II and FG-NET databases prove the effectiveness of our novel method.

Chapter 7

PROPOSED METHOD III: DAG-CNN AND ITS VARIANTS

7.1 Introduction

The fundamental components of DAG-CNN are given in the following subsections. We shortly introduce VGG-16 and GoogLeNet architectures and explain the method of expected age estimation and score-level fusion for age estimation.

7.2 Directed Acyclic Graph-Convolutional Neural Network

Deep learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural network. Recently, deep artificial neural networks (including recurrent ones) have outperformed numerous state-of-the-art methods in pattern recognition and machine learning. The directed acyclic graph (DAG) networks can represent more complex network architectures compared to simple ones which consist of a linear chain of layers. DAG architecture for neural networks (NNs) has emerged from the idea of recurrent NNs that have some feedback connections from forward layers to backward ones, which give them the ability of capturing dynamic states. The main advantage of DAG-structured networks is that their forward layers can have multiple input parameters from several backward layers. In this way, they can achieve different levels of image representations. A fundamental feature of the deep learning neural networks is the use of connections between their layers, called “skip connection”, that is similar to

DAG-CNNs main idea, and it is shown that these skip connections can improve the accuracy of the classification tasks significantly.

DAG-CNN was proposed by Yang and Ramanan[77] to learn a set of multi-scale image features that are successfully used for classification of three standard scene benchmarks. They showed that the multi-scale model can be implemented as a DAG-structured feed forward CNN. By this approach, it is possible to use an end–to-end gradient-based learning for automatically extracting multi-scale features using generalized back propagation algorithm over the layers that have more than one input. In fact, all the required equations for training the network are standard CNN equations except for the Add and ReLU layers since they have multiple inputs or outputs. Considering the ith ReLU layer in Figure 7.1, let α_i be its input, $\beta_i^{(j)}$ be the output for its jth output branch (its jth child in the DAG), and assume that z is the final output of the softmax layer. The gradient of z with respect to the input of the ith ReLU layer can be computed as in Eq. (7.1):

$$\frac{\partial_z}{\partial_{\alpha_i}} = \sum_{j=1}^{C} \frac{\partial_z}{\alpha\beta_i^{(j)}} \frac{\alpha\beta_i^{(j)}}{\partial\alpha_i} \tag{7.1}$$

where C is the number of output edge of the ith ReLU.

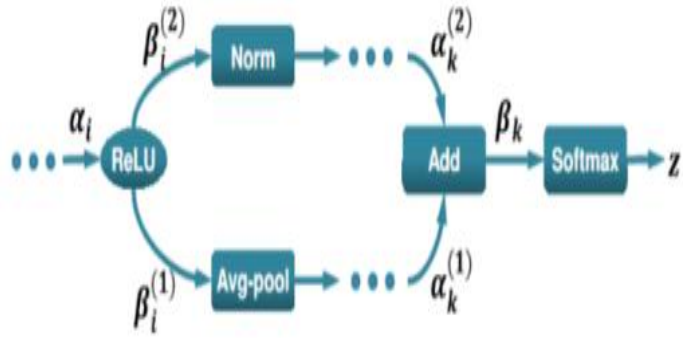


Figure 7.1: Parameter Setup at i-th ReLU [77]

For the Add layer, let $\beta_k = g(\alpha_k^{(1)}, \cdots, \alpha_k^{(N)})$ represents the output of an Add layer with multiple inputs. The gradient along the layer can be computed by applying the chain rule as in Eq. (7.2):

$$\frac{\partial_z}{\partial \alpha_i} = \frac{\partial_z}{\partial \beta_k}\frac{\partial \beta_k}{\partial \alpha_i} = \frac{\partial_z}{\partial \beta_k}\sum_{j=1}^{C}\frac{\partial \beta_k}{\partial \alpha_k^{(j)}}\frac{\partial \alpha_k^{(j)}}{\partial \alpha_i} \tag{7.2}$$

In the convolutional layers, the convolution operation is computed by Eq. (7.3) as follows:

$$X_n = \sum_{k=0}^{N-1} y_k f_{n-k} \tag{7.3}$$

where y and f are the input image and applied filter, respectively and N is the number of elements in the input image. The convolution layer output is represented by vector X. For all layers of the DAG-CNN architecture except ReLU and Add layers, the Eq. (7.4) and Eq. (7.5) are used to update biases and weights as follows:

$$\Delta W_t(t+1) = -\frac{x_\lambda}{r}W_l - \frac{x}{n}\frac{\partial C}{\partial W_l} + m\Delta W_l(t) \tag{7.4}$$

$$\Delta B_l(t+1) = -\frac{x}{n}\frac{\partial C}{\partial B_l} + m\Delta B_l(t) \tag{7.5}$$

where $W, B, l,$ $\lambda, x, n, m, t,$ and C denote the weight, bias, layer number, regularization parameter, learning rate, total number of training samples, momentum, updating step, and cost function, respectively.

In DAG-CNNs, since lower layers are directly connected to the output layer through multi-scale connections, it is guaranteed that these layers' neurons receive a strong gradient signal during learning and do not suffer from the problem of vanishing gradients. In CNNs, the size of the learned features in intermediate layers can be very large and combining these features may cause the curse of dimensionality problem.

In order to overcome this problem, marginal activations by performing average pooling on the learned features of some layers which are used for score-level fusion.

7.3 Baseline Architectures

In this subsection, two different DAG-CNN architectures are proposed to improve the discrimination capability of a deep neural network by allowing its layers to share their learned features and work collaboratively for classification. The proposed multi-scale CNN topologies employ learned features with different level of complexity in order to estimate the subject's age with high precision.

7.3.1 VGG-16 Architecture

VGG-16 architecture [43] had been proposed by the Oxford Visual Geometry Groups' model in ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [43]. VGG-16 is deeper and wider than former CNN structure and it has five batches of convolution operations, each batch consisting of 2 to 3 adjacent convolution layers. Adjacent convolution batches are connected via max-pooling layers. The size of kernels in all convolutional layers is 3×3 convolutional layers and the number of kernels within each batch is the same (increases from 64 in the first group to 512 in the last one). Figure 7.2 illustrates the VGG-16 architecture. This architecture has been used in many researches and it was the first one that outperformed human-level performance on ImageNet.

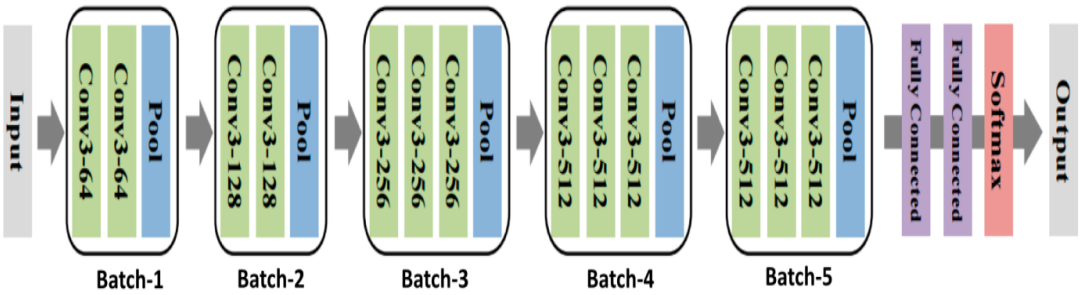


Figure 7.2: VGG-16 Architecture

7.3.2 GoogLeNet Architecture

The GoogLeNet [44] model is the winner of ILSVRC in 2014. A new module named Inception is introduced in [44] which apply various sizes of convolutional kernels to be composed to form more discriminative feature representations. The depth of GoogLeNet reaches to 22 and the number of convolutional layers reaches to 60, so that the errors always vanish with back propagation, and the parameters of low layers may not be optimized sufficiently. To address this issue, two auxiliary classifiers are employed by GoogLeNet to optimize the parameters of low layers. In GoogLeNet, there are 9 Inception modules employed to construct the architecture. The Inception module consists of a few of convolutional kernels with small sizes (such as 1×1, 3×3 and 5×5), which are conducive to limit the scale of parameters and model complexity. To learn efficiently, GoogLeNet introduced 1×1 convolutions for feature dimension reduction. In order to overcome the problems of gradient vanishing and over-fitting, these 9 Inception modules are divided into 3 groups, and three objective functions are added on every 3 Inception modules (Figure 7.3).

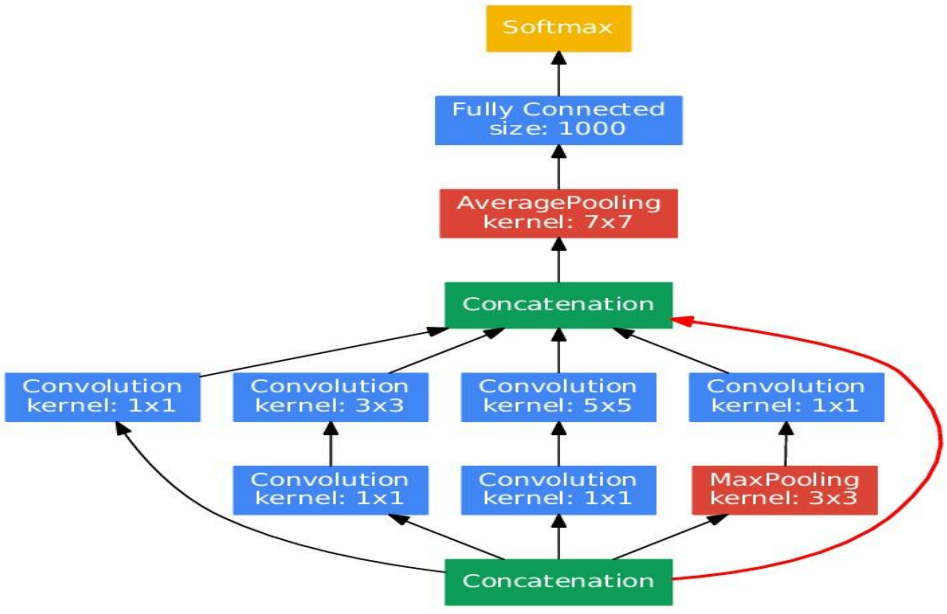


Figure 7.3: Inception Modules of GoogLeNet Architecture

7.4 Expected Age Value

Age estimation can be considered as a discrete classification with multiple discrete value labels. For Morph-II dataset, it is a one dimensional regression problem with the age being sampled from a continuous range between 16 and 77 and for FG-NET dataset it is between 0 and 70. For computing the expected age from CNN, instead of using simple softmax, we multiply each softmax output probabilities by the corresponding class year label and add them together as follows:

$$Expected\ Age = \sum_{i=L}^{H} p_i \times year\_label_i \tag{7.6}$$

where L, H are lower and upper bound of subject's age, p_i is softmax output probability and $year\_label_i$ is an integer age year value of output neuron i in the last layer of CNN.

7.5 Score-level Fusion of Multi-stage CNN Learned Features

CNNs can be used as automatic feature extractors and the cost-free mid-level features extracted from intermediate layers of a CNN can be discriminative for classifying different patterns with varying complexities. There are two ways in order to use these features: feature-level fusion and score-level fusion. In the feature-level fusion approach, different layer features are concatenated to create final feature vector, then it is fed into a classifier or regression. One of the common problems with feature-level fusion is the size of the feature vectors. In CNNs, the size of the learned features in intermediate layers can be very large and combining these features may cause the curse of dimensionality problem. Dimensionality reduction methods such as Principal Component Analysis (PCA) or Discrete Cosine Transform (DCT) can be used to overcome this problem [78]. Another approach for overcoming the dimensionality problem is score-level fusion. In this method, features of each layer

are given to a separate classifier to generate a score vector for the test sample, then these scores are combined together to generate the final decision.

The whole process of score-level fusion is given in Algorithm 1. This algorithm can be expanded for more than two feature sets. In score-level fusion, the distance between each test sample and all the training samples is computed and assumed to be the score of that test sample in the corresponding classification/regression system. These scores are normalized by Min-Max normalization [75] method as follows:

$$x' = \frac{x - Min(x)}{Max(x) - Min(x)} \tag{7.7}$$

where x is the raw score, $Max(x)$ and $Min(x)$ are the maximum and minimum values of the raw scores respectively and x' is the normalized score.

After normalization, the score vectors are combined by Sum rule-based fusion method [75] to generate a single scalar score which is then used to make the final decision.

Algorithm-2 Score-level Fusion of Two Feature Sets(FS^m, FS^n) for Age Estimation

Input:

Trainset $X_{train} = \{(X_{tr}^i, y_{tr}^i)\}, i = 1, \dots, N_{train}$

Testset $X_{test} = \{(X_{te}^i, y_{te}^i)\}, i = 1, \dots, N_{test}$

Output:

Estimated_age_vector, MAE

1: ${F_m}^{{X_{tr}}^i} \leftarrow$ Compute FS^m $\quad \forall$ i $= 1, \dots, N_{train}$

2: ${F_n}^{{X_{tr}}^i} \leftarrow$ Compute FS^n $\quad \forall$ i $= 1, \dots, N_{train}$

3: For j= 1 To N_{test}

4: ${F_m}^{{X_{te}}^j} \leftarrow$ Compute FS^m

5: ${F_n}^{{X_{te}}^j} \leftarrow$ Compute FS^n

6: For i = 1 To N_{train}

7: $score_i^m$= *compute_distance* (${F_m}^{{X_{tr}}^i}$, ${F_m}^{{X_{te}}^j}$)

8: $score_i^n$= *compute_distance* (${F_n}^{{X_{tr}}^i}$, ${F_n}^{{X_{te}}^j}$)

9: For i = 1 To N_{train}

10: normalized $score_i^m$ according to Eq. (7.7)

11: normalized $score_i^n$ according to Eq. (7.7)

12: $fusion_i = score_i^m + score_i^n$

13: $minIndex$=*Find_Min_index*($\boldsymbol{fusion}$)

14: Estimated_age_vector$_i$ = $y_{tr}^{minindex}$

15: $Error$ += *abs (*Estimated_age_vector$_i$ - y_{te}^j)

16: MAE=$Error$/ N_{test}

7.6 DAG-CNN Architecture for Age Estimation

In this section, we propose two new DAG-CNN architectures for estimating the accurate age from facial image by exploiting multi-stage learned features from

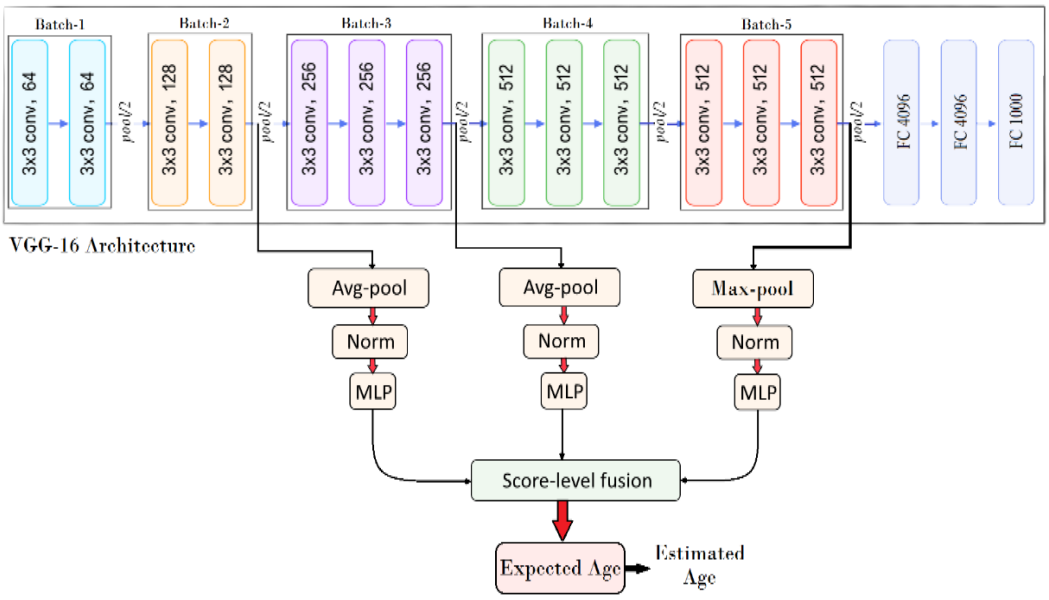
different layers of a VGG-16 CNN and GoogLeNet models. Convolutional neural networks can be used as automatic feature extractors and the learned features can be fed to classifiers like SVMs or NNs to predict the output labels. Mid-level features at intermediate layers of the CNN can be discriminative for classifying different patterns with varying complexities. However, in CNN architectures used in literature so far, these cross-layer heterogeneity features are ignored. It is clear to see that these mid-level features are already computed when the system is trained to extract high-level features, and hence, their usage does not bring any extra computational burden within our proposed model.

According to the success rate of using score-level fusion in face and multimodal biometric recognition systems [3][79][80] it is believed that accuracy can be improved when the information of different types of feature descriptors and classifiers are consolidated. The reason is that different layers of CNN learn different level of information which varies from local and detailed information to more abstract one. In order to test this hypothesis, we construct a DAG-CNN network which automatically performs score-level fusion of different selected layers. Therefore, instead of manually performing feature-level or score-level fusion and feeding the results to a classifier, we propose a multi-scale system by using a CNN with Directed Acyclic Graph (DAG) topology. Our third proposed model can automatically learn different level of features, combine them by score-level fusion method and estimate the final age.

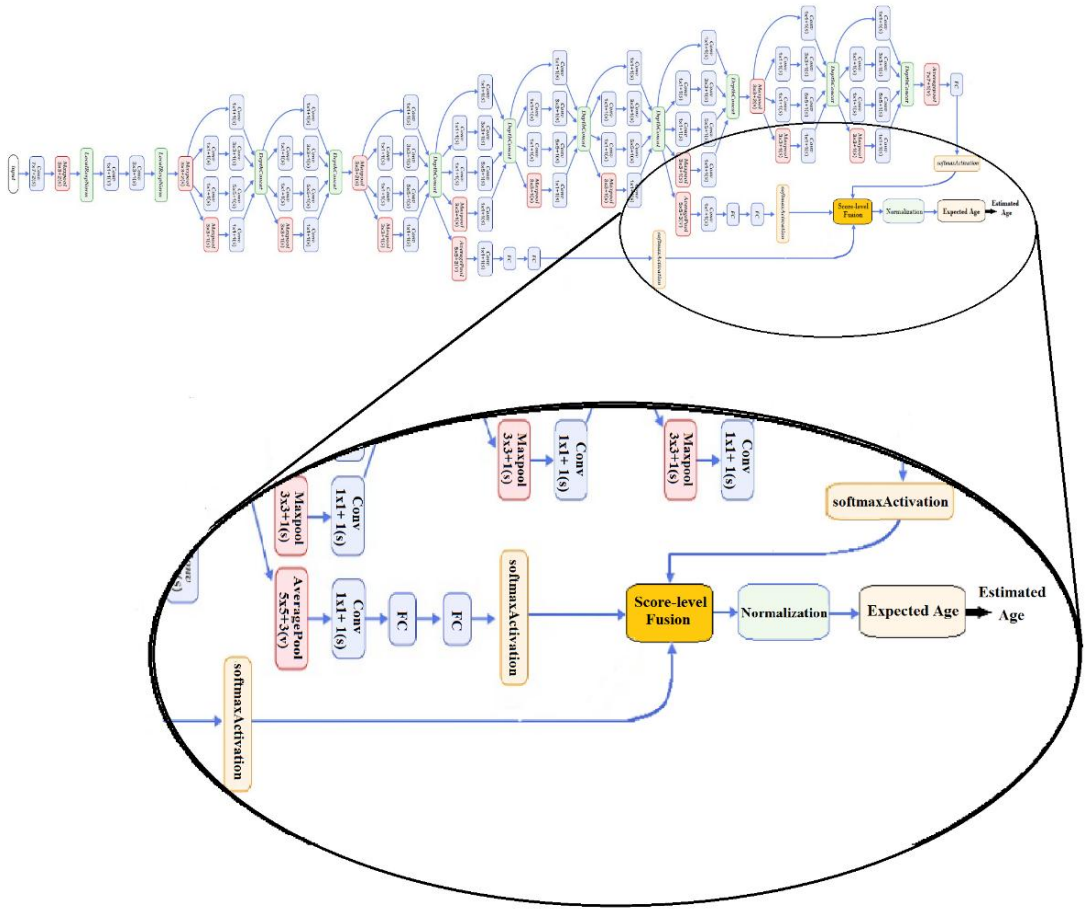
In order to investigate the suitability of DAG-CNN, we propose two different models based on two well-known CNN architectures, namely VGG-16 and GoogLeNet and named these third proposed systems as DAG-VGG16 and DAG-GoogLeNet,

respectively. We use these architectures as the backbone of our proposed DAG topology and employed some branches from their intermediate and last layers. These links are connected to an average pooling layer to reduce their dimensionality, then are normalized and given to a separate fully connected MLP layers. Each of these fully connected layers have the same number of neurons in their last layer and that is equal to number of age classes(54 and 70 different age classes for Morph-II and FG-NET datasets respectively) and generate a score vector for each input image. These score-vectors are added with each other, element by element and the result vector is normalized such that its components' summation becomes 1. Then the normalized vector is fed into the final decision layer in which the subject's age is estimated by Eq. (7.6).

Both VGG-16 and GoogLeNet contain many layers and due to the possible redundancy among these layers' features, it is improper to fuse all of these features. VGG-16 has five batches of convolution operations with 2 or 3 adjacent convolution layers. Adjacent convolution batches are connected via max-pooling layers and these locations are suitable candidate points for multi-stage score-level fusion. Therefore, by considering the trade off between model accuracy and complexity, the VGG-16 is partitioned into five parts and the optimal selection of these parts' features should be considered for the final age estimation. Finally, for DAG-VGG16, we select Batch 2, 3 and 5 as the DAG branch positions by using a greedy algorithm which is explained in Section 5.5. For GoogLeNet based system, in order to select layers for combining in DAG-CNN model in an effective way, with the aid of the auxiliary classifiers defined by GoogLeNet, we select the position of the auxiliary classifiers for getting branches and perform score-level fusion of these three stages.



(a)



(b)

Figure 7.4: Overview of the Third Proposed Methods: (a) DAG-VGG16, (b) DAG-GoogLeNet

In this study, we have shown that combining different level features can improve age estimation accuracy significantly. Particularly, the accuracy is improved when we add features learned by intermediate layers, with the exception of the low-level features of early layers that cause a decrease in estimation accuracy. For the purpose of testing different combinations of feature layers and finding the best one experimentally, features of the last layer are considered as of necessary and intermediate layer features are added layer-by-layer, one at a time, in a backward fashion until no improvement observed in classification accuracy. This greedy approach ignores the features of layers closer to the input layer. Experimental evaluations as illustrated within the next section exhibited that the proposed system's capability of fusion of multi-scale features improves the accuracy of age estimation. The overall schematic of the third proposed methods are illustrated in Figure 7.4(a) and 7.4(b).

7.7 Experimental Settings and Results

7.7.1 Preprocessing

In this study, we used face detection method described in [73] for the detection of facial images. Then the facial image will be aligned by using geometric transformation such that the eyes have been symmetrically placed at 25% and 75% of the aligned image. All the input images are in RGB format and resized to 256 × 256. Five different cropped size of 227× 227 and their flip are fed to the DAG-VGG16 network while the size of cropped input images for DAG-GoogLeNet architecture is 224× 224. The same data augmentation methods are deployed for the offline multi-stage feature fusion systems.

7.7.2 Offline Multi-stage Feature Fusion

In order to investigate the effectiveness of our multi-stage features fusion approach, we manually combined the features from different layers of trained and fine-tuned VGG-16 and GoogLeNet models. In order to implement score-level fusion manually for VGG-16 based system, a pre-trained model on IMDB-WIKI dataset is selected from [51] and fine-tuned on FG-NET and Morph-II datasets. For GoogLeNet based system, we firstly pre-train the model using CASIA-WebFace database. Afterwards, it is fine-tuned on IMDB-WIKI and finally fine-tuned on one of the target datasets, FG-NET and Morph-II. The numbers of neurons in the last layer of the models are changed to output 54 or 70 age labels for Morph-II or FG-NET dataset respectively. Additionally, expected age obtained by Eq. (7.6) is used as the cost functions. In fine-tuning phase for VGG-16 based system, the weights of early layers are frozen and only the Batch-3 to Batch-5 and fully connected layers' weights are updated. These two configurations are considered as the baselines for comparison with offline and online multi-stage score-level fusion approaches. For VGG-16 based system, the baseline MAE is 2.91 for Morph-II and 3.36 for FG-NET databases and for GoogLeNet based system, the baseline MAE are 3.13 and 3.29, for Morph-II and FG-NET databases, respectively.

Afterwards, different layers' features are extracted and are fused together by Algorithm-2 for score-level fusion. In this computation, different layers' features are extracted for all training samples separately. In the test phase, for each test sample and for each separate layer's features, the corresponding feature vector is computed and its distance is compared with all of the feature vectors in the training set and these distances are stored in an array, namely score vector. After computing the score vector for all of the feature sets, the scores are normalized and combined together by

adding them element by element. The training sample's age with minimum distance is considered as the estimated age.

For VGG-16 based system, in order to test different combinations of feature layers and finding the optimal configuration experimentally, the features of the last batch layer (Batch-5) are considered as of necessary and features from different intermediate batch layers are added, one at a time, in a backward fashion until no improvement is observed in age estimation accuracy. When a new batch layer's feature is added, if the accuracy is decreased, we ignore the new batch layer's feature and backtracking and select another batch from the remaining ones. This greedy approach ignores the features of batch layer close to the input layer (Batch-1) and Batch-4 and as a result, the optimal subset of features for score-level fusion is features extracted from batch layers 2, 3, and 5. For GoogLeNet based system, we select the position of the auxiliary classifiers for performing score-level fusion.

The experimental results show that combining the intermediate features with last layer features with score-level fusion causes meaningful improvement in MAE. The MAE of manually multi-stage score-level fusion for VGG-16 based system on Morph-II and FG-NET datasets are improved to 2.86 and 3.22 years old, respectively. The results of this approach shows improvement for GoogLeNet based system too. As shown in Table 7.1, the MAE of GoogLeNet based system on Morph-II and FG-NET datasets are improved to 2.99 and 3.17 years old, respectively.

7.7.3 DAG-CNN Architecture for Age Estimation

DAG-CNN consists of a normal CNN and some branches from its different layer to fuse multi-stage learned features. We selected two publicly available CNN, VGG-16 and GoogLeNet architectures, as the chain-structured or backbone of the DAG-CNN architectures due to their impressive result on the ILSVRC. For DAG-VGG16 system, we started with VGG-16 deep CNN models from [52] which is pre-trained on the IMDB-WIKI dataset and performed the following modifications: instead of rectified linear unit (ReLU), we utilized S-shaped ReLU [76]. Additionally, we use batch normalization between convolution layers to reduce the internal covariate shift. It helps the network to learn how to combine color features in an optimal way. For DAG-GoogLeNet system, we used the aforementioned baseline system as the back-bone structure.

For the purpose of testing different combinations of feature layers and finding the best DAG-CNN architecture experimentally, for both DAG-VGG16 and DAG-GoogLeNet systems, we selected the DAG branch places according to the configuration result obtained in Section 5.5. Each branch is given to an average pooling layer with 5×5 kernel size to reduce its dimensionality and is normalized and finally, is fed into a multi-layer perceptron (MLP) classifier to compute its score. The scores from different MLP are fused together by addition rule [75] . This score-level fusion result is normalized to sum 1 and used to compute the estimated-age.

All of the MLP classifiers contain three layers in which the first and the second layers have 500 neurons and the last fully connected layer of the model to output 54 or 70 classes corresponding to different age labels in Morph-II and FG-NET datasets, respectively. The MLP neuron's weights are initialized to small normally-distributed

numbers. For computing the estimated age, instead of using simple softmax, the estimated age is computed by using Eq. (7.6). Finally, the modified model is employed for fine tuning on the Morph-II dataset. In order to avoid overfitting, we used different learning rate policy for different layers. Therefore we used small learning rate for the feature extraction layers and froze the early layers' weights, but for the fully connected layers we utilized higher learning rate. Additionally, we performed data augmentation by cropping five different regions of 224×224 pixels (227×227 for DAG-GoogLeNet) from the 256×256 input image (four corners and the center one) and their mirror version in the training phase. All the input images are in RGB format.

Our proposed DAG-CNN architecture was fine-tuned through the standard backpropagation technique with a batch size of 32. In order to obtain optimum performance, the other learning parameters are set as follows: to prevent overfitting of training data, the regularization (λ) is set to 0.1, momentum parameters which adjust the speed of learning during training is set to 0.9, and learning rate that control the convergence of the training data are set to 0.001 and linearly changed according to the mean-squared error values in each ten iteration. The training was performed for 50 epochs. The MAE of DAG-VGG16 is 2.81 years for Morph-II and 3.08 years for FG-NET, while the MAE of DAG-GoogLeNet system are 2.81 and 3.08 years for Morph-II and FG-NET datasets, respectively. All of the DAG related results are better than the results of offline multi-stage feature fusion method. The reason is that, in DAG-CNN and during the training phase, different features from different layers are combined and the model learned more discriminative features with respect to the simple CNN model in the baseline and its offline score-level fusion counterpart. All

experiment results are summarized in Table 7.1. Additionally, in Figure 7.5 and Figure 7.6, we showed the MAE separately for each age in order to investigate the effectiveness of each aforementioned system on Morph-II dataset. These results show that in both cases of DAG-CNN and offline score-level fusion, for all subjects' age, the integration of several layers' features caused the improvement in age estimation's accuracy. Furthermore, it is clear that, in most of the age values, the features learned by DAG-CNN have more discriminative power than features obtained by offline score-level fusion.

Table 7.1: The Experimental results summary of the 3rd proposed system

Method	MAE(years)	
	Morph-II	FG-NET
VGG-16 Baseline	2.91	3.36
Offline multi-stage features fusion	2.86	3.22
Proposed DAG-VGG16	2.81	3.08
GoogLeNet Baseline	3.13	3.29
Offline multi-stage features fusion	2.99	3.17
Proposed DAG-GoogLeNet	2.87	3.05

Moreover, The CS curves of the third proposed method compared with state-of-the-art methods on Morph-II and FG-NET datasets are illustrated in Figure 7.7 and Figure 7.8. The cumulative score diagrams of our proposed method outperform all of the other compared methods in all different levels of error. It can be seen that for Morph-II dataset, both DAG-VGG16 and DAG-GoogLeNet systems' are within 1-year error for 40% of samples and within 5-year error for near 80% of them. These

results for FG-NET dataset and in case of DAG-VGG16 system are 34% and 83%, and in case of DAG-GoogLeNet system are 38% and 83%, respectively.

All the experiments are performed on a machine with Intel Xeon E5-2683 - 2.0 GHz processor and 16 GB Ram. All the codes are written with Matlab 2017b platform.

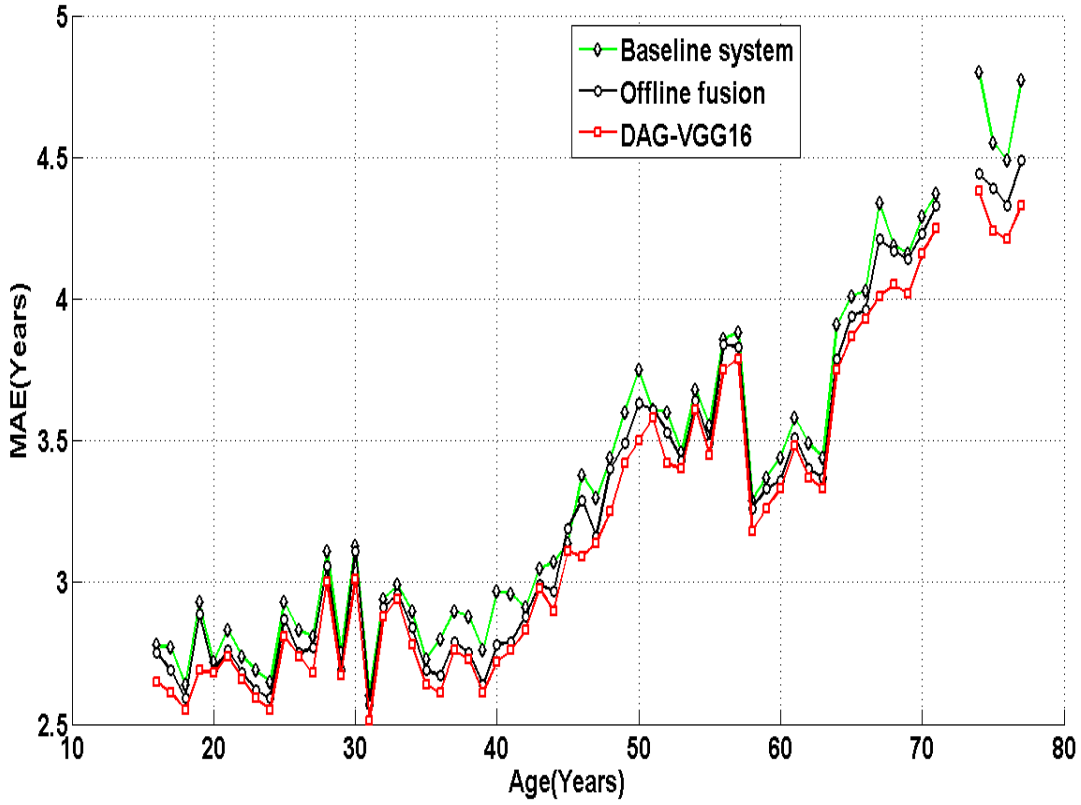


Figure 7.5: MAE of DAG-VGG16 System for Morph-II Dataset

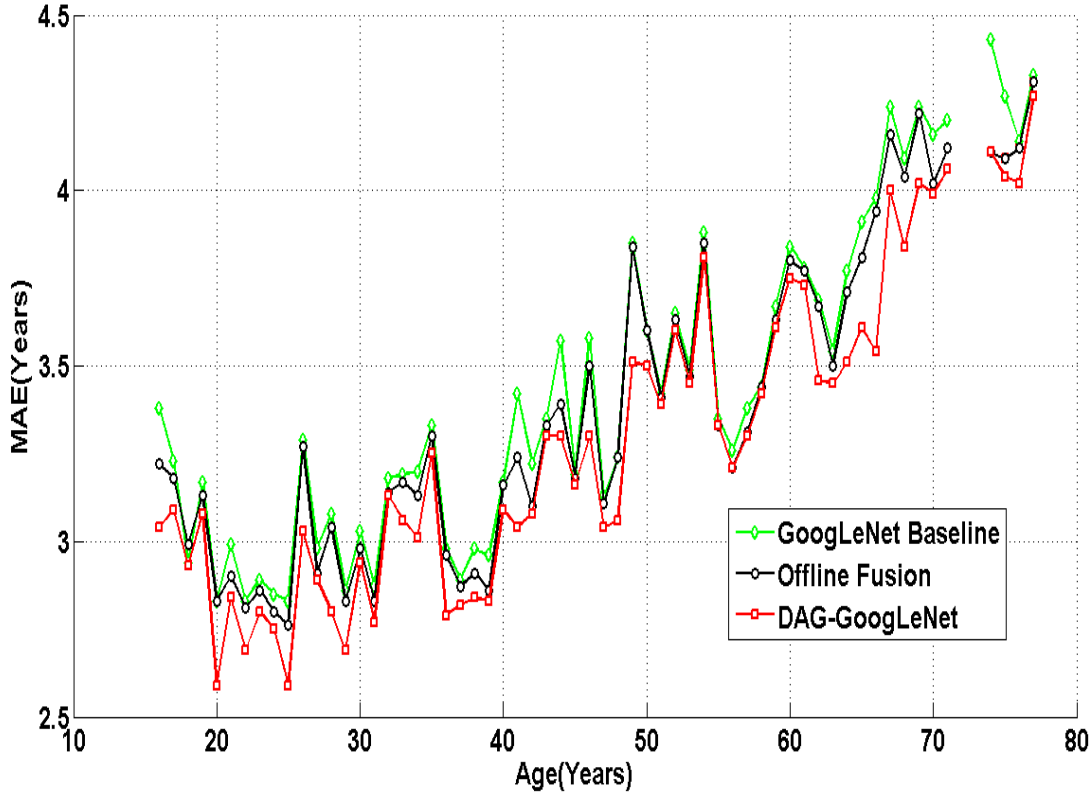


Figure 7.6: MAE of DAG-GoogLeNet System for Morph-II Dataset

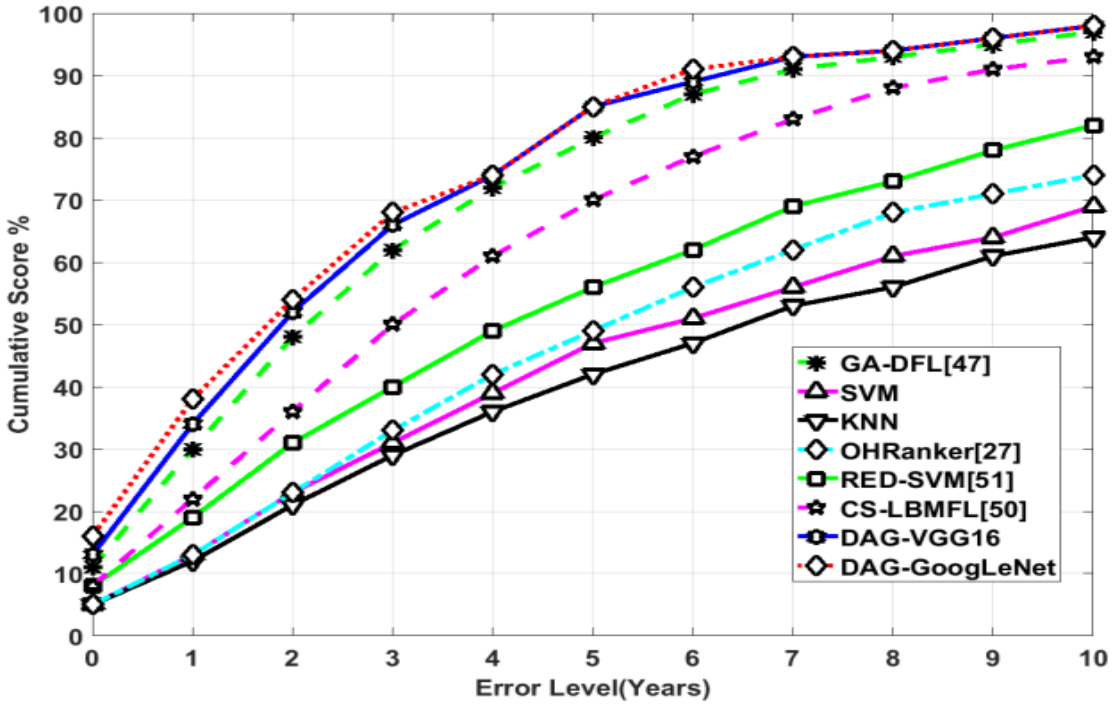


Figure 7.7: The CS Curves of the DAG-CNN Methods Compared with State-of-the-Art Methods on Morph-II Dataset

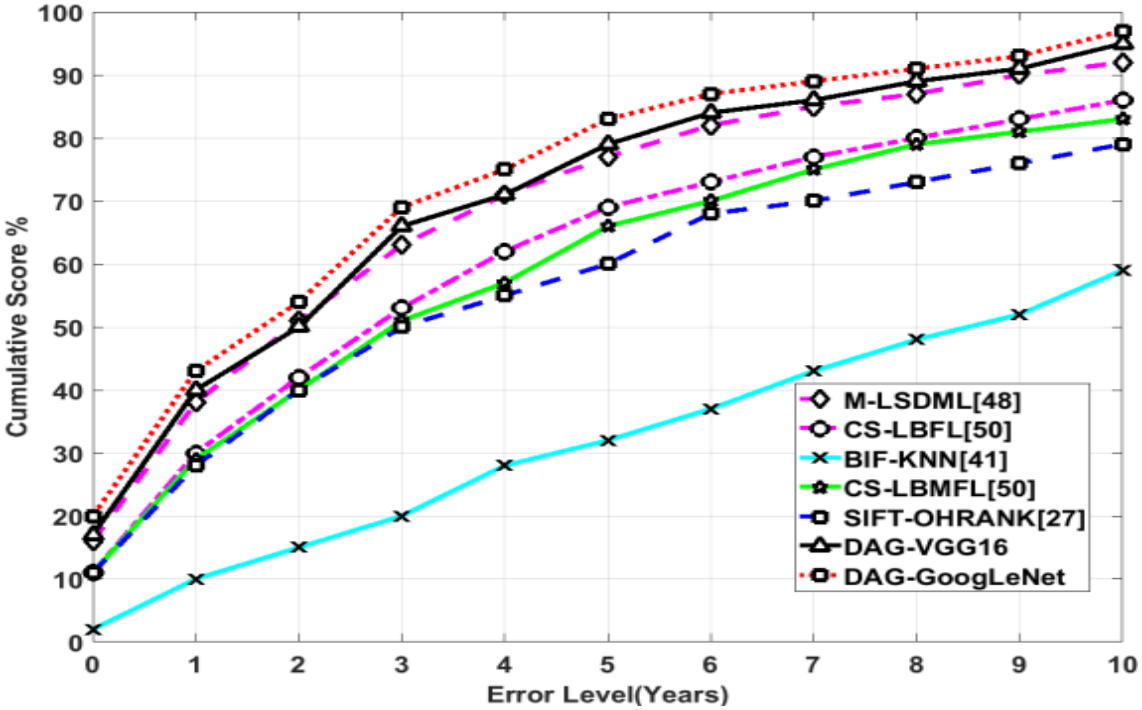


Figure 7.8: The CS Curves of the DAG-CNN Methods Compared with State-of-the-Art Methods on FG-NET Dataset

7.8 Conclusion

In this chapter, we propose a new architecture of deep neural networks namely Directed Acyclic Graph Convolutional Neural Networks (DAG-CNNs) for age estimation which exploits multi-stage features from different layers of a CNN. This system is constructed by adding multi-scale output connections to underlying backbones from VGG-16 and GoogLeNet. DAG-CNNs not only fuse the feature extraction and classification stages of the age estimation into a single automated learning procedure, but also utilized multi-scale features and perform score-level fusion of multiple classifiers automatically. Fine-tuning such models helps to increase the performance and we show that even "off-the-shelf" multi-scale features perform quite well. Experiments on the publicly available Morph-II and FG-NET databases prove the effectiveness of our novel method.

7.9 Comparison with the State-of-the-art Methods

We compare all of the three proposed methods' accuracy with the state-of-the-art methods that presented the results on Morph-II and FG-NET datasets. The robustness and effectiveness of the proposed methods are studied in terms of MAEs in Table 7.2 and Table 7.3. Consequently, the experimental results, the MAE value and CS curve, show that our proposed methods outperform most of the state-of-the-art methods. It is clearly seen that our proposed methods outperform many other methods such as hand-crafted and CNN-based approaches. This improvement is caused by different factors of the proposed methods such as advanced architecture, using additional dataset and transfer learning method, fusion of different layers' features and the expected age formula. The results demonstrate that combining different features by score-level fusion in both offline and DAG-CNN version enhances the performance of the age estimation system.

Table 7.2: Comparison with the state-of-the-art methods on Morph-II dataset

Reference	Method/Feature	MAE
Lanitis et al.,2002[11]	WAZ/AAM+BIF	9.21
Geng et al.,2013 [39]	AAS/AAS+BIF	10.10
Chang et al.,2011 [14]	SVM/AAM	6.49
Chang et al.,2011 [14]	OHRank/AAM	6.07
Chang et al.,2011 [14]	OHRank/AAM+BIF	6.28
Guo and Mu,2011 [19]	PLS/BIF	4.56
Guo and Mu,2011 [19]	kPLS/BIF	4.04
Geng et al.,2013 [39]	IIS-LLD/AAM+BIF	5.67
Geng et al.,2013 [39]	CPNN/AAM+BIF	4.87
Guo and Mu,2013[73]	CCA/BIF	5.37
Guo and Mu,2013 [73]	rCCA/BIF	4.42
Guo and Mu,2013 [29]	kCCA/BIF	3.98
Geng et al.,2013 [39]	MFOR/PCA+LBP+BIF	4.20
Han et al.,2013 [69]	SVM+SVR/BIF+ASM	4.20
Huerta et al.,2015[35]	rCCA/Fusion	4.25
Huerta et al.,2015 [35]	CNN/CNN	3.88
Yang et al.,2015 [52]	Deeprank/Deep Network	3.57
Han et al.,2015 [31]	DIF/Demographic	3.80
Huerta et al.,2014 [35]	Fusion	4.25
Niu et al.,2016 [49]*	OR-CNN/CNN	3.27
Rothe et al.,2016 [51]*	DEX/CNN	3.25
Rothe et al.,2016 [51]*	DEX(IMDB-WIKI)/CNN	2.68
Wang et al.,2015 [81]	DLA/CNN	4.77
Yi et al.,2014 [41] *	CNN	3.63
Duan et al.,2018 [47]	CNN+ELM	3.44
Hu et al.,2017 [48]*	CNN	2.78
Ng et al.,2017[83]*	CNN	3.88
Antipov,2017[84]*	CNN	2.99
1st proposed method	**Two-level fusion of hand-crafted features**	**3.89**
2nd proposed method	**Two-level fusion of CNN and hand-crafted features**	**3.17**
3rd proposed method	**Score-level fusion with DAG-VGG16***	**2.81**
	Score-level fusion with DAG-GoogLeNet*	**2.87**

* denotes that an additional dataset was used

Table 7.3: Comparison with the state-of-the-art methods on FG-NET dataset

Reference	Method/Feature	MAE
El Dib et al., 2010[85]	BIF	3.17
Han et al., 2013[69]	component and holistic BIF	4.6
Geng et al., 2013[39]	label distribution(CPNN)	4.76
Liang et al.,2014[86]	hierarchical framework	4.97
Lu et al., 2015[87]	CS-LBFL	4.43
Lu et al., 2015[87]	CS-LBMFL	4.36
Chang et al., 2015[88]	CS-OHR	4.70
Chen et al., 2013[89]	CA-SVR	4.67
Chen et al, 2016[90]	Cascaded-CNN	3.49
Liu et al., 2018[91]	M-LSDML	3.31
1st proposed method	**Two-level fusion of hand-crafted features**	**4.06**
2nd proposed method	**Two-level fusion of CNN and hand-crafted features**	**3.29**
3rd proposed method	**Score-level fusion with DAG-VGG16**	**3.08**
	Score-level fusion with DAG-GoogLeNet	**3.05**

Chapter 8

CONCLUSION

Age estimation from facial images is an important application of biometrics. In contrast to other facial variations like occlusions, illumination, misalignment and facial expressions, ageing variation is affected by human genes, environment, lifestyle and health which make age estimation a challenging task. In this thesis, three novel age estimation methods are proposed.

In the first proposed system, an integration of different type of feature extraction algorithms is applied on facial images for accurate age estimation. This integration is performed by using two-level fusion of features and scores with the help of feature-level and score-level fusion techniques. In this system, the advantage of using different types of features such as biologically-inspired features, texture-based features and appearance-based features is used. Feature-level fusion of biologically-inspired and texture-based methods is integrated into the proposed method and their combination is fused with an appearance-based method using score-level fusion.

The second proposed age estimation system is based on three different levels of information fusion. The wrinkle features extracted by Gabor filters and texture-based and shape-based features extracted by MRELBP and BIF are involved during the first level of information fusion (feature-level fusion) process. Multi-stage learned features from different layers of a trained CNN for age are combined together by

score-level fusion method. Finally the obtained results of these two approaches are aggregated by using weighted averaging. The aggregation result shows that generic features extracted from CNN are enhanced by combining them with domain-specific features.

The third proposed system deployed a novel CNN architecture for age estimation which is based on automatic multi-stage fusion of information. Multi-stage learned features from different layers of a CNN are automatically combined together by score-level fusion method by using DAG-CNN architecture. We showed that DAG-CNN can improve the discrimination capability of a deep neural network by allowing its layers to share their learned features and work collaboratively for classification.

Compared with the state-of-the-art methods, our proposed approaches obtained significant lower MAE on Morph-II and FG-NET datasets. The experimental results showed that the feature-level and score-level fusion of local handcrafted features and global learned features provide a higher accuracy than the other algorithms. Moreover, the automatic score-level fusion of multi-stage learned features provide the highest accuracy.

8.1 Future Work

There are a number of promising future directions for age estimation. Following are some of future research directions that may see improvement in age estimation performance:

- This study recommends further investigation into new CNN architectures such as ResNet and r-CNN and their DAG versions.

- In the third proposed method, the locations of DAG branches should be determined manually. A further research can be established to design a new CNN architecture, where all of its intermediate layers are combined together by different learnable weights and the system decides which mid-level features should be used in order to improve the accuracy.
- Paying more attention to age estimation algorithms such as combining the classification and regression methods might improve the results further, because choosing an age estimation algorithm is always a critical choice that will increase or decrease the performance notably.
- Databases: A large multi-racial database is needed for effective investigation of ageing in different ethnic groups and gender. Collecting a large database with well distributed age labels is essential. Web image collection is efficient way of achieving this.
- Ethnic based: Faces of subjects from different ethnic groups age differently. Incorporating ethnic parameters as in [166] improves age estimation performance. This approach has not been fully investigated due to lack of large datasets with images from different ethnic groups like African, Asian, and Caucasian.
- Feature enhancement: This study recommends further investigation into enhancement using boosting or feature selection machine learning techniques. Machine learning boosting and feature selection could be applied to fused features or to individual features before fusion.

- People rely on multiple cues to estimate other peoples age such as face, voice, gait and hair. Combining face with one or more other cues for age estimation might remarkably improve the current performance.

REFERENCES

[1] Farkas, L. G. (Ed.). (1994). *Anthropometry of the Head and Face*. Raven Pr.

[2] Angulu, R., Tapamo, J. R., & Adewumi, A. O. (2018). Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1), 42.

[3] Taheri, S., & Toygar, O. (2018). Multi-Stage Age Estimation Using Two Level Fusions of Handcrafted and Learned Features on Facial Images. *IET Biometrics*. DOI: 10.1049/iet-bmt.2018.5141

[4] Taheri, S., & Toygar, Ö. (2019). On the use of DAG-CNN architecture for age estimation with multi-stage features fusion. *Neurocomputing*, 329, 300-310.

[5] Kwon, Y. H., & da Vitoria Lobo, N. (1993, August). Locating facial features for age classification. In *Intelligent Robots and Computer Vision XII: Algorithms and Techniques* (Vol. 2055, pp. 62-73). International Society for Optics and Photonics.

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

[7] Ramanathan, N., & Chellappa, R. (2006). Face verification across age progression. *IEEE Transactions on Image Processing*, 15(11), 3349-3361.

[8] Ramanathan, N., Chellappa, R., & Biswas, S. (2009). Age progression in human faces: A survey. *Journal of Visual Languages and Computing*, 15, 3349-3361.

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

[11]Lanitis, A., Taylor, C. J., & 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.

[12] Luu, K., Dai Bui, T., & Suen, C. Y. (2011, March). Kernel spectral regression of perceived age from hybrid facial features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (pp. 1-6). IEEE.

[13] Kwon, Y. H., & da Vitoria Lobo, N. (1999). Age classification from facial images. *Computer vision and image understanding*, 74(1), 1-21.

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

[15] Geng, X., Zhou, Z. H., & Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 29(12), 2234-2240.

[16] Fu, Y., Xu, Y., & Huang, T. S. (2007, July). Estimating human age by manifold analysis of face pictures and regression on aging features. In *Multimedia and Expo, 2007 IEEE International Conference on* (pp. 1383-1386). IEEE.

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

[18] Cai, D., He, X., Han, J., & Zhang, H. J. (2006). Orthogonal laplacian faces for face recognition. *IEEE transactions on image processing*, 15(11), 3608-3614.

[19] Guo, G., & Mu, G. (2011, June). Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on* (pp. 657-664). IEEE.

[20] Wu, T., Turaga, P., & Chellappa, R. (2012). Age estimation and face verification across aging using landmarks. *IEEE Transactions on Information Forensics and Security*, 7(6), 1780-1788.

[21] Kwon, Y. H., & da Vitoria Lobo, N. (1999). Age classification from facial images. *Computer vision and image understanding*, 74(1), 1-21.

[22] Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International journal of computer vision*, 1(4), 321-331.

[23] Hayashi, J. I., Yasumoto, M., Ito, H., & Koshimizu, H. (2002). Age and gender estimation based on wrinkle texture and color of facial images. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on* (Vol. 1, pp. 405-408). IEEE.

[24] Ahonen, T., Hadid, A., & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12), 2037-2041.

[25] Gunay, A., & Nabiyev, V. V. (2008, October). Automatic age classification with LBP. In *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on* (pp. 1-4). IEEE.

[26] Ylioinas, J., Hadid, A., & Pietikäinen, M. (2012, November). Age classification in unconstrained conditions using LBP variants. In *Pattern recognition (icpr), 2012 21st international conference on* (pp. 1257-1260). IEEE.

[27] Liu, C., & 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.

[28] Gao, F., & Ai, H. (2009, June). Face age classification on consumer images with Gabor feature and fuzzy LDA method. In *International Conference on Biometrics* (pp. 132-141). Springer, Berlin, Heidelberg.

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

[30] Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11), 1019.

[31] Han, H., Otto, C., Liu, X., & Jain, A. K. (2015). Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 1148-1161.

[32] Guo, G., & Mu, G. (2014). A framework for joint estimation of age, gender and ethnicity on a large database. *Image and Vision Computing*, 32(10), 761-770.

[33] Fu, Y., Guo, G., & Huang, T. S. (2010). Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32(11), 1955-1976.

[34] Weng, R., Lu, J., Yang, G., & Tan, Y. P. (2013, April). Multi-feature ordinal ranking for facial age estimation. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1-6). IEEE.

[35] Huerta, I., Fernández, C., & Prati, A. (2014, September). Facial age estimation through the fusion of texture and local appearance descriptors. In *European Conference on Computer Vision* (pp. 667-681). Springer, Cham.

[36] Makihara, Y., Okumura, M., Iwama, H., & Yagi, Y. (2011, October). Gait-based age estimation using a whole-generation gait database. In *Biometrics (IJCB), 2011 International Joint Conference on* (pp. 1-6). IEEE.

[37] Xia, B., Amor, B. B., Daoudi, M., & Drira, H. (2014, January). Can 3D shape of the face reveal your age?. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on* (Vol. 2, pp. 5-13). IEEE.

[38] Lanitis, A., Draganova, C., & Christodoulou, C. (2004). Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1), 621-628.

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

[40] Yang, M., Zhu, S., Lv, F., & Yu, K. (2011, June). Correspondence driven adaptation for human profile recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 505-512). IEEE.

[41] Yi, D., Lei, Z., & Li, S. Z. (2014, November). Age estimation by multi-scale convolutional network. In *Asian conference on computer vision* (pp. 144-158). Springer, Cham.

[42] Yan, C., Lang, C., Wang, T., Du, X., & Zhang, C. (2014, December). Age estimation based on convolutional neural network. In *Pacific Rim Conference on Multimedia* (pp. 211-220). Springer, Cham.

[43] Ueki, K., Hayashida, T., & Kobayashi, T. (2006, April). Subspace-based age-group classification using facial images under various lighting conditions. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on* (pp. 6-12). IEEE.

[44] Fu, Y., & Huang, T. S. (2008). Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4), 578-584.

[45] Guo, G., Fu, Y., Dyer, C. R., & Huang, T. S. (2008, June). A probabilistic fusion approach to human age prediction. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on* (pp. 1-6). IEEE.

[46] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[47] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

[48] Hu, Z., Wen, Y., Wang, J., Wang, M., Hong, R., & Yan, S. (2017). Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7), 3087-3097.

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

[50] Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017, May). An all-in-one convolutional neural network for face analysis. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on* (pp. 17-24). IEEE.

[51]Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4), 144-157.

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

[53] Yoo, B., Kwak, Y., Kim, Y., Choi, C., & Kim, J. (2018). Deep Facial Age Estimation Using Conditional Multitask Learning With Weak Label Expansion. *IEEE Signal Processing Letters*, 25(6), 808-812.

[54] Liu, H., Lu, J., Feng, J., & Zhou, J. (2017). Ordinal Deep Learning for Facial Age Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*. doi: 10.1109/TCSVT.2017.2782709

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

[56] Shu, X., Tang, J., Lai, H., Liu, L., & Yan, S. (2015). Personalized age progression with aging dictionary. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3970-3978).

[57] Duong, C. N., Quach, K. G., Luu, K., Le, T. H. N., & Savvides, M. (2017, October). Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *Computer Vision (ICCV), 2017 IEEE International Conference on* (pp. 3755-3763). IEEE.

[58] Tang, P., Wang, H., & Kwong, S. (2017). G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing*, 225, 188-197.

[59] Farrajota, M., Rodrigues, J. M., & du Buf, J. H. (2016, December). Using Multi-Stage Features in Fast R-CNN for Pedestrian Detection. In *Proceedings of the 7th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion* (pp. 400-407). ACM.

[60] Batool, N., & Chellappa, R. (2015). Fast detection of facial wrinkles based on Gabor features using image morphology and geometric constraints. *Pattern Recognition*, 48(3), 642-658.

[61] Choi, S. E., Lee, Y. J., Lee, S. J., Park, K. R., & Kim, J. (2011). Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition*, 44(6), 1262-1281.

[62] Izadpanahi, S., & Toygar, Ö. (2014). Human age classification with optimal geometric ratios and wrinkle analysis. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(02), 1-16.

[63] Manjunath, B. S., & Ma, W. Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8), 837-842.

[64] Liu, L., Lao, S., Fieguth, P. W., Guo, Y., Wang, X., & Pietikäinen, M. (2016). Median robust extended local binary pattern for texture classification. *IEEE Transactions on Image Processing*, 25(3), 1368-1381.

[65] Liu, L., Fieguth, P., Guo, Y., Wang, X., & Pietikäinen, M. (2017). Local binary features for texture classification: taxonomy and experimental study. *Pattern Recognition*, 62, 135-160.

[66] Ricanek, K., & Tesafaye, T. (2006, April). Morph: A longitudinal image database of normal adult age-progression. *In Automatic Face and Gesture*

Recognition, 2006. FGR 2006. 7th International Conference on (pp. 341-345). IEEE

[67] Gehrig, T., Steiner, M., & Ekenel, H. K. (2011). Draft: evaluation guidelines for gender classification and age estimation. *Technical report, Karlsruhe Institute of Technology*.

[68] Crowley, J. L., & Cootes, T. (2009). FG-NET: Face and Gesture Recognition Working Group (2002). http://www-prima.inrialpes.fr/FGnet/. Accessed 10 Apr 2018

[69] Panis, G., Lanitis, A., Tsapatsoulis, N., & Cootes, T. F. (2016). Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics*, 5(2), 37-46.

[70] Theriault, C., Thome, N., & Cord, M. (2011, September). HMAX-S: deep scale representation for biologically inspired image categorization. In *Image Processing (ICIP), 2011 18th IEEE International Conference on* (pp. 1261-1264). IEEE.

[71] Shan, C. (2010, October). Learning local features for age estimation on real-life faces. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis* (pp. 23-28). ACM.

[72] Liu, C. (2006). Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5), 725-737.

[73] Guo, G., & Mu, G. (2013, April). 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* (pp. 1-6). IEEE.

[74] Hsu, T. C., Huang, Y. S., & Cheng, F. H. (2010, September). A novel ASM-based two-stage facial landmark detection method. In *Pacific-Rim Conference on Multimedia* (pp. 526-537). Springer, Berlin, Heidelberg.

[75] He, M., Horng, S. J., Fan, P., Run, R. S., Chen, R. J., Lai, J. L., ... & Sentosa, K. O. (2010). Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognition*, 43(5), 1789-1800.

[76] Jin, X., Xu, C., Feng, J., Wei, Y., Xiong, J., & Yan, S. (2016, February). Deep Learning with S-Shaped Rectified Linear Activation Units. In *AAAI* (Vol. 3, No. 2, pp. 3-2).

[77] Yang, S., & Ramanan, D. (2015). Multi-scale recognition with DAG-CNNs. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1215-1223).

[78] Nanni, L., Ghidoni, S., & Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, 71, 158-172.

[79] Eskandari, M., & Toygar, Ö. (2014). Fusion of face and iris biometrics using local and global feature extraction methods. *Signal, image and video processing*, 8(6), 995-1006.

[80] Sim, H. M., Asmuni, H., Hassan, R., & Othman, R. M. (2014). Multimodal biometrics: Weighted score level fusion based on non-ideal iris and face images. *Expert Systems with Applications*, 41(11), 5390-5404.

[81] Wang, X., Guo, R., & Kambhamettu, C. (2015, January). Deeply-learned feature for age estimation. In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 534-541). IEEE.

[82] Duan, M., Li, K., Yang, C., & Li, K. (2018). A hybrid deep learning CNN–ELM for age and gender classification. *Neurocomputing*, 275, 448-461.

[83] Ng, C. C., Cheng, Y. T., Hsu, G. S., & Yap, M. H. (2017, May). Multi-layer age regression for face age estimation. In *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on* (pp. 294-297). IEEE.

[84] Antipov, G., Baccouche, M., Berrani, S. A., & Dugelay, J. L. (2016). Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 96-104).

[85] El Dib, M. Y., & El-Saban, M. (2010, September). Human age estimation using enhanced bio-inspired features (EBIF). In *Image Processing (ICIP), 2010 17th IEEE International Conference on* (pp. 1589-1592). IEEE.

[86] Liang, Y., Wang, X., Zhang, L., & Wang, Z. (2014). A hierarchical framework for facial age estimation. *Mathematical Problems in Engineering*, 2014.

[87] Lu, J., Liong, V. E., & Zhou, J. (2015). Cost-sensitive local binary feature learning for facial age estimation. *IEEE Transactions on Image Processing*, 24(12), 5356-5368.

[88] Chang, K. Y., & Chen, C. S. (2015). A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing*, 24(3), 785-798.

[89] Chen, K., Gong, S., Xiang, T., & Change Loy, C. (2013). Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2467-2474).

[90] Chen, J. C., Kumar, A., Ranjan, R., Patel, V. M., Alavi, A., & Chellappa, R. (2016, September). A cascaded convolutional neural network for age estimation of unconstrained faces. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on* (pp. 1-8). IEEE.

[91] Liu, H., Lu, J., Feng, J., & Zhou, J. (2018). Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security*, 13(2), 292-305.