

Towards Automation and Human
Assessment of Objective Skin
Quantification

JHAN SAAD A ALARIFI

PhD 2021

Towards Automation and Human
Assessment of Objective Skin
Quantification

JHAN SAAD A ALARIFI

A thesis submitted in partial fulfilment of
the requirements of
Manchester Metropolitan University
for the degree of Doctor of Philosophy

Visual Computing Group
Department of Computing and
Mathematics
Manchester Metropolitan University


2021

Declaration of Authorship

I, Jhan Saad A ALARIFI, declare that this thesis titled, "Towards Automation and Human Assessment of Objective Skin Quantification " and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:



Abstract

The goal of this study is to provide an objective criterion for computerised skin quality assessment. Humans have been impacted by a variety of face features. Utilising eye-tracking technology assists to get a better understanding of human visual behaviour, this research examined the influence of face characteristics on the quantification of skin evaluation and age estimation. The results revealed that when facial features are apparent, individuals do well in age estimation. Also, this research attempts to examine the performance and perception of machine learning algorithms for various skin attributes. Comparison of the traditional machine learning technique to deep learning approaches. Support Vector Machine (SVM) and Convolutional Neural Networks (CNNs) were used to evaluate classification algorithms, with CNNs outperforming SVM. The primary difficulty in training deep learning algorithms is the need of large-scale dataset. This thesis proposed two high-resolution face datasets to address the requirement of face images for research community to study face and skin quality. Additionally, the study of machine-generated skin patches using Generative Adversarial Networks (GANs) is conducted. Dermatologists confirmed the machine-generated images by evaluating the fake and real images. Only 38% accurately predicted the real from fake correctly. Lastly, the performance of human perception and machine algorithm is compared using the heat-map from the eye-tracking experiment and the machine learning prediction on age estimation. The finding indicates that both humans and machines predict in a similar manner.

Acknowledgements

A large number of individuals have contributed to my knowledge and ideas over the course of putting together this thesis. Through this study, I would like to acknowledge and express my heartfelt gratitude to my three supervisors: Prof. Moi Hoon Yap, Prof. Darren Dancey, and Dr. John Fry. They have provided me with invaluable guidance, valuable time, technical assistance, and friendly dealings throughout the course of my research project. I would not have been able to complete my thesis without their continuous encouragement, advice, and attention. I would like to especially thank and express my gratitude for Dr. Moi Hoon Yap, without whom I would not have made it through my research journey. For I am grateful for her continuous support and guidance, and the invaluable encouragement throughout this study. Your unwavering support, and belief in me from the beginning made my research possible. I would like to express my gratitude to my family and friends for their unwavering encouragement, patience, and support. The direct, and indirect assistance and support I have received from a variety of individuals have inspired and allowed me to undertake this investigation. Many thanks to Mr Jirah Jam, Mr. Guido Ascenso, Dr. Connah Kendrick and all the members of the Visual Computing Lab for their assistance and insightful remarks. In addition, I would like to express my heartfelt gratitude to the SciEng Research Degrees and IT Services teams for their helpful assistance during my studies. Finally, I would like to express my deepest appreciation to Saudi Arabia Ministry of Education for sponsoring a PhD studentship for this research.

Contents

Declaration of Authorship	i
Abstract	ii
Acknowledgements	iii
1 Introduction	1
1.1 Introduction	1
1.2 Motivation	4
1.3 Aim and Objectives	5
1.4 Contributions	6
1.5 Thesis Organisation	6
2 Literature Review	8
2.1 Skin Assessment Applications	8
2.2 Human skin assessment	9
2.2.1 Skin Assessment	11
2.2.2 Skin Lesions	11
2.2.3 Facial Skin	12
2.3 Skin Ageing	13
2.3.1 Natural Aging Process	13
2.4 Facial Aging Factor	15
2.4.1 Intrinsic Factor	16
2.4.2 Extrinsic Factor	17

2.4.3	Perceived age	18
2.5	Skin Quality	18
2.6	Human perception	19
2.7	Eye-tracking overview	20
2.7.1	Eye-tracking for human perception	21
2.8	Facial Skin Assessment	22
2.8.1	Skin Cancer Assessment	25
2.9	A review on Face Datasets	28
2.10	Summary	44
3	Preliminary Research and Datasets	45
3.1	Introduction	45
3.2	Social Habits Dataset	46
3.3	Make-up Dataset	47
3.3.1	Standardization of combining both datasets	49
3.4	Experiment I: Automated Wrinkles Annotator	49
3.4.1	Proposed Method	50
3.4.2	Results and Discussion	52
3.5	Experiment II: Automated Acne Detection	53
3.5.1	Proposed Method	54
3.5.2	Results and Discussion	55
3.6	Summary	56
4	Facial Skin Classification using Machine Learning and Deep Learning Techniques	58
4.1	Introduction	58
4.2	Brief Overview of Deep Learning	60
4.3	Classification evaluation	62
4.4	Facial skin classification	64
4.5	Proposed Method	64

4.6	Results and Discussion	65
4.7	Summary	66
5	Human perception: Face Age Estimation and Skin Quality	67
5.1	Introduction	67
5.2	Hypotheses	68
5.3	Eye-tracking Experiments	69
5.4	Statistical Analysis	70
5.4.1	Age estimation	72
5.4.2	Skin Quality	76
5.5	Eye-tracking results	79
5.5.1	Age estimation	80
5.5.2	Fixation count	81
5.5.3	First fixation	82
5.5.4	Revisits fixation count	83
5.5.5	Skin Quality	85
5.5.6	Fixation count	85
5.5.7	First fixation duration	86
5.5.8	Revisits fixation count	87
5.6	Summary	88
6	Machine Perception of Skin Quality	90
6.1	Introduction	90
6.2	Skin Quality Evaluation	91
6.3	Skin quality prediction	95
6.4	Machine generated skin lesions	95
6.5	Dermatologists evaluation	97
6.6	Summary	99
7	Human Perception and Machine Prediction	100

7.1	Introduction	100
7.2	Human perception of face age	101
7.2.1	Computer methods for face age estimation	102
7.3	Method	103
7.3.1	Experiment I: Human perception	103
7.3.2	Experiment II: CNN model	105
	Data preprocessing	105
	Network architecture	105
	Machine Specification	106
7.4	Results	106
7.4.1	Experiment 1: eye-tracking	106
7.4.2	Experiment 2: CNN visualisation	108
7.5	Discussion	108
7.6	Summary	111
8	Conclusion and Future Work	112
8.1	Introduction	112
8.2	Research Findings	112
8.3	Future Work	116
8.3.1	Data collection	116
8.3.2	Human perception	117
8.3.3	Objective skin assessment	118
8.4	Summary	118
A	Publications	120
B	Data Collection	121
C	The Eye-tracking raw data	127

List of Figures

2.1	The structures and layers of the skin (Skin-remedies.com, 2017).	10
2.2	Image showing different colours of various skin lesion with melanin storage.	12
2.3	Barata, Marques, and Emre Celebi, 2019	28
3.1	The facial expressions captured	47
3.2	The angles at which photos were taken.	47
3.3	Cameras location and experiment setup.	48
3.4	The make-up data collection stages.	48
3.5	The lighting variation at 90°	48
3.6	Wrinkle Annotator.	51
3.7	Manual annotation.	52
3.8	Comparison of the two filters, where (a) is the original image, (b) is the filter proposed in (Ng et al., 2014) and (c) is the proposed filter	55
3.9	Example of the results: where (a) is the original image and (b) is the detection results (red correct detection and green is the missed detection)	55
3.10	Manual annotation.	56
4.1	Feedforward Neural Network.	61
4.2	Convolutional Neural Network.	62
4.3	Sample skin patch from each three classes.	65

5.1	Random stimuli example. (a) ROI mask overview, (b) Frontal face image, (c): is featureless face after applying the mask. . . .	69
5.2	The experimental setting for the Experiment II.	70
5.3	Eye-tracking sections on the face.	80
5.4	Average fixation count for experts and non-experts on the face parts for mask and full face.	82
5.5	Average time fixation in ms for experts and non-experts on the face parts for mask and full face.	83
5.6	Average count of fixation revisits for experts and non-experts on the face parts for mask and full face.	84
5.7	Average fixation count for experts and non-experts on the face parts for mask and full face.	86
5.8	Average time fixation in ms for experts and non-experts on the face parts for mask and full face.	87
5.9	Average count of fixation revisits for experts and non-experts on the face parts for mask and full face.	88
6.1	Comparison of skin quality score from experts for faces and masks for age group (21-30).	92
6.2	Comparison of skin quality score from experts for faces and masks for age group (31-40).	92
6.3	Comparison of skin quality score from experts for faces and masks for age group (41-50).	93
6.4	Comparison of skin quality score from experts for faces and masks for age group (51-60).	94
6.5	Comparison of skin quality score from experts for faces and masks for age group (61-70).	94
6.6	Examples of the fake and real images.	98

6.7	Comparison of generated skin lesions using SRGAN Ledig et al., 2017 (Left) compared to ground-truth images (Right) . . .	99
7.1	The data collected from eye-tracking. Left panel: visualisation about fixation. The numbers indicate the number of ms that the gaze was focused on that area. Also, a longer gaze on an area corresponded to a larger red circle around that area. Right panel: image data converted into heat-maps to visualise the areas where the participants focused on the longest (the longer a participant's eyes focused on an area, the more red the heat-map)	104
7.2	Comparison of human eye gaze pattern with computer model activation map. The first row is for age group (21-30), second row (31-40), third (41-50), fourth (51-60), and fifth (61-70). The number indicates the age group (leftmost two columns). Average heat-map from eye-tracking results and Grad-CAM (rightmost two columns).	107

List of Tables

2.1	Summary of existing face datasets	30
4.1	Confusion matrix	62
4.2	Performance measures	63
4.3	SVM results.	66
4.4	GoogLeNet results.	66
5.1	This is the legend for Kendall's correlation coefficient	72
5.2	Fleiss' Kappa Statistics - Absolute Agreement Summary for Age Estimation	72
5.3	Kendall's Correlation Coefficient for age estimation.	72
5.4	Kendall's Correlation Coefficient for Skin Quality	76
5.5	Fleiss' Kappa Statistics - Absolute Agreement Summary for Skin Quality	76
7.1	Age and gender information of 44806 samples from Album2 of MORPH dataset	105

Dedicated to my parents

1 Introduction

This chapter introduces skin assessment as well as the terminology that will be used throughout the thesis. It discusses the motivation for this research and expands on the aim and objectives. It also discusses the gaps found in the literature. This chapter also includes the structure and organisation of the thesis.

1.1 Introduction

Facial Skin Analysis is a vital yet difficult task essential to numerous fields, one of which dermatologists use to examine skin (Korotkov and Garcia, 2012). Facial skin is affected by both internal and external factors, internal factors are unalterable and include genetics and ageing which have a major impact on facial skin appearance, and external factors which include environmental effects such as sun exposure, air pollution, nutritional (Puizina-Ivic, 2008) and social habits such as poor diet, smoking and alcohol use (Puizina-Ivic, 2008; Osman et al., 2017). Ageing is a natural process resulting in decreasing facial skin elasticity (Puizina-Ivic, 2008), and can be exacerbated by external factors.

Several studies explain the importance of facial skin texture, due to its vital influence on the perception of attractiveness, health and age (Perrett et al., 1998; Rhodes, Sumich, and Byatt, 1999). Facial skin appearance is important since it affects many aspects of people's lives, for example, their well-being, employability chances, and interpersonal relationships (Samson, Fink, and

Matts, 2010). Skin texture provides meaningful information about skin surface and its geometry (Fink, Grammer, and Thornhill, 2001).

As we grow older, the facial skin loses firmness, begins to wrinkle, and develops discoloration and uneven pigmentation, which is exacerbated by photoaging in particular. As a result, the frequency of lines and wrinkles, as well as dyschromia and a decrease in bulk light reflection, are all factors that influence the estimation of face skin age in humans. Although the latter have been shown to alter female age perceptions by up to 20 years, as well as affecting health judgments when skin subsurface cues are absent (Merinville et al., 2018), it remains to be seen whether skin topography has a similar effect on perception and, indeed, how skin colour and topography relate to one another in this regard.

At the time of writing, we are only aware of one study that investigated how skin texture cues (lines and wrinkles) affect the female's perceived age in facial composites. (Fink and Matts, 2008) illustrated that skin colour distribution and skin surface topography cues have a substantial impact on the perception of age and health in female faces, as well as on the perception of male faces. Partner selection is influenced by skin colour, and paler skinned women are often selected by men (Fokuo, 2009). Conversely, females with darker skin were rated as more attractive than light-skinned females (Watson, Thornton, and Engelland, 2010). Skin quality properties are frequently assembled and evaluated by a well-trained expert who assigns visible skin samples, either live or from photographs, to a recognised quality grade on a predefined grading scale. However, using a machine vision approach to assess skin quality properties can provide an objective analysis. Because a professional's experience and knowledge are subjective and can differ between graders, this can help to avoid problems with repeatability and reproducibility. This could result in a lower cost and more effective analysis, as well as a

more consistent assessment of skin quality (Quer et al., 2017). This will, however, come at a cost because the task will necessitate a large dataset of high-resolution images. Furthermore, data cannot be gathered without taking into account facial cosmetics such as make-up, cosmetic facial skin surgery, or non-cosmetic facial surgery. For example, someone who has had Botox will have a different age estimation than someone of the same age group who has not had Botox. All relevant questions in the questionnaire must be carefully considered so that they do not influence the results negatively or positively. Also, different ethnic groups age differently due to different conditions and nutrition standards. Another aspect of the model will be based on the algorithm designer, which can be put to research as many algorithms will be proposed and the state-of-the-art established and used as a benchmark for other models.

Human perception differs from machine perception, in that while machines heavily depend on dataset, which may vary, the human perception remains consistently invariant and remains the best. However, human and machine perception can be comparable in that meaningful clinical information can be removed and the region of the skin during assessment seen to be more suspicious. For information to be integrated in human perception, processing of sensory information requires time whereby the brain must gather further information to rectify ambiguities (Frolov et al., 2019) (Maksimenko et al., 2020). Perception in humans can be influenced in several ways which may deem it a disadvantage, relative to machine perception. Pre-existing knowledge, which aids in interpretation of information can be dependent on memory, which can be biased or inaccurate in interpreting visual input (Appelle and Countryman, 1986). Furthermore, focus of attention impacts human perception, as stated by the hypothesis of focused attention (LaBerge and Samuels, 1974) whereby what is seen by the observer is decided by what

the observer attends to. Therefore, perception can be selective; simultaneous processing of information is not possible and is limited to a specific area of space at a given time in humans. In machines, there is a possibility where altering size of attention focus is possible according to setting of images to be process and its size, thus altering form of focus is possible.

There are many ambiguities in the field of machine perception, a field concerning machines that are able to sense and interpret their environments (Nevatia, 1982; Makino et al., 2020). It is necessary to establish perceptual awareness, which is the process of processing and condensing various sensory information before interpreting it, resulting in automated perception of objects. It is difficult to design an artificial model of human-like machine perception because the varying functional systems in humans are difficult to replicate in machine applications. This also applies to representing information and processing in such systems, as well as combining and consolidating data from various systems. Furthermore, there is no explicit, single model of human perception, but rather a plethora of contradictory theories and blind spots within them. Thus, using the human perceptual system as a biological archetype to derive a machine model is difficult. More research on human perception versus machine perception is needed to potentially characterise and mitigate this effect in order to better understand the differences between machines and humans.

1.2 Motivation

Skin appearance and properties are changing over time, it is a challenging task to provide objective quantification for skin assessment. Human judgment is subjective and often inconsistent. It would be crucial to characterise the progressive but subtle variations in facial appearance when people

age using machine learning methods, because it has many important consequences, including the following biases:

- i. Human perception which includes dermatologists, whom might be biased towards certain facial features. The eyes and nose provide more information for age estimation than any of the other face components (forehead, eyebrows, mouth, and shape) (Nkengne et al., 2008; Han, Otto, and Jain, 2013).
- ii. Designing a machine learning method that has human-like performance, is important but difficult.

1.3 Aim and Objectives

The research aim is to propose algorithms that can be used to compare the performance of the dermatologist with machine learning methods.

To achieve the aim, the following objectives have been established:

1. To establish high-resolution datasets for facial skin assessment.
2. To propose new computer methods for objective quantification of facial skin assessment.
3. To investigate human perception of skin quality and face age estimation with and without facial features
4. To establish an objective quantification technique on skin quality assessment for various age groups.
5. To conduct a comparative study on the performance of human and machine in skin assessment.

1.4 Contributions

This section presents a summary of contributions in respect to the objectives, outlined as follows:

1. Two new high resolution face datasets for skin assessment were proposed (Objective 1, Publication [P01]).
2. A computer method for objective quantification of acne were proposed (Objective 2).
3. The first to propose facial skin classification using Convolutional Neural Networks (Objective 2, Publication [P01]).
4. To investigate human perception on skin quality, It is hypothesised that facial features influence both the accuracy of age prediction and the quality of skin assessment as perceived by humans. Understanding the impact of facial characteristics on human perception and assessment accuracy (Objective 3).
5. An expert investigation of skin quality assessment on five age groups. Dermatologists are unable to distinguish between images of real skin lesions and images generated by a computer (Objective 4).
6. A new study compares the similarities and differences between human judgement and machine prediction on face age estimation, where eye regions are important cues for both (Objective 5, Publication: [P03]).

1.5 Thesis Organisation

The following chapters are organized as follows:

- **chapter 2** includes an overview and related work sections accompany each contribution being described in each section.

-
- **chapter 3** includes description of proposed high resolution datasets (Contribution 1) and proposed computer methods for objective quantification and detection of wrinkles and acne (Contribution 2).
 - **chapter 4** introduces computer methods for objective quantification of facial skin assessment based on classification techniques using conventional machine learning and the state-of-the-art CNN models to classify the facial skin patches of three types i.e. normal, spot and wrinkles (Contribution 3).
 - **chapter 5** undertakes an investigation of human judgment by using an eye-tracking device to recognise the potential factors that might affect this. It will be carried out with participants to understand how they observe and score faces (Contribution 4).
 - **chapter 6** conducts similar experiments as in Chapter 5, but the participants are from dermatologists (Contribution 5). It compares the perception of experts (dermatologists) vs non-expert.
 - **chapter 7** compares human and machine performance on two folds: 1) A comparative study of human eye-tracking results with machine prediction on face age estimation; and 2) Machine generated skin lesions on human prediction (Contribution 6).
 - **chapter 8** concludes the thesis by presenting a summary of the achievements, limitations and challenges, and provide new insights for potential future works.

2 Literature Review

This chapter discusses the aspects of skin appearance and structure. The chapter is divided into three main sections: human skin; normal and diseased skin along with their assessment methods and human perception on skin.

2.1 Skin Assessment Applications

Over the years, many skin care applications have been developed with a high influx of high-end applications developed with the introduction of deep learning models (Goodfellow et al., 2014). The availability of tele-dermatology services, along with self-care applications are increasing. Some of those application were created for medical practice, aid medical students, dermatologists, surgeons, nurses and other medical professional with their role.

Dermatology Atlas (*An Ongoing Commitment to Equity in Medicine*) is a valuable information source for skin problems with dermatology. It provides a comprehensive way to better understand any given skin condition. Many skin analysis applications gained popularity in recent years.

Cureskin (*Acne, Pimples, Skin Hairfall Treatment: CureSkin - Apps on Google Play*) was developed to aid the population by developing an AI tool that is able to diagnose six types of the most common skin conditions in India, as a result of a lack of dermatologists. These six types includes pimples, acne,

scars, dark spots, pigmentation, and dark circles, after the diagnosis, the system would recommend a skincare regimen.

Skin Vision (*Skin Cancer Melanoma Detection App 2021*) is an app that evaluates track moles and evaluates the risk of developing skin cancer by keeping a track of the moles. It contained more than 27k cases of skin cancer.

Molemap (*Home: MoleMap New Zealand*) is another common application for diagnosing the skin cancer moles, they have a wild range of selection/types of skin.

2.2 Human skin assessment

Skin is the largest organ of the human body (Kanitakis, 2001). It consists of three main layers as illustrated in Figure 2.1. The epidermis, the dermis and the hypodermis. They will each be discussed briefly in that order. The outer layer is the epidermis, this is a thin layer that works on the outer part of the skin to mainly protect the functions of the skin: protection, sensation, thermoregulation and metabolic function (Kanitakis, 2001). The cells in this layer divide in the basal layer, the cells change on a regular basis, within a 2-4-week period. The epidermis is divided into sub-layers, the second main layer of the skin is the dermis, which is right under the basal cell layer of the epidermis. The dermis layer is the thickest inner layer. It is an important layer for sensation purposes as well as protecting the thermoregulation function. This layer consists of fibroblast, sweat glands, nerves, blood supply and hair. The deepest and final layer of the skin is the hypodermis layer which carries the metabolic function of the skin, it also contains sweat glands and adipose tissue.

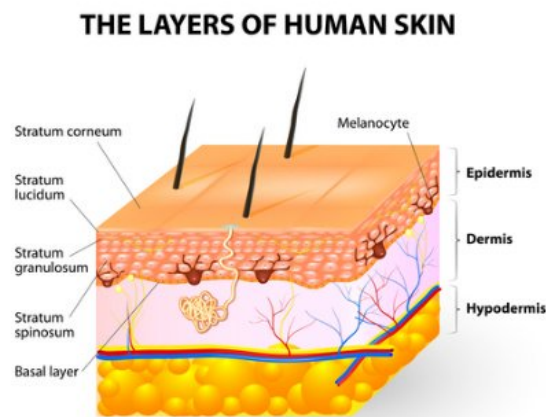


FIGURE 2.1: The structures and layers of the skin (Skin-remedies.com, 2017).

There are a number of studies that considered the human perception of age in order to develop a computer method for age estimation (Han, Otto, and Jain, 2013; Lanitis, Taylor, and Cootes, 2002; Geng, Zhou, and Smith-Miles, 2007). These studies considered the whole face known as global features. (Osman et al., 2017; Ng et al., 2015a) used part of the face only by applying a mask to the face to estimate age, which are known as local features. However, the study is limited to age estimation only by analysing wrinkles as a sign of ageing. They did not include other facial attributes such as spots and pigmentation which are related to skin ageing. (Nkengne et al., 2008) studied the influence of facial skin attributes on the perceived age of Caucasian women. The study had two objectives: to understand the influence of facial features when estimating age and to understand the influence of gender in terms of judgement that they make. The study consisted of 173 stimuli images of Caucasian women with an age range of 20 to 74 years, and the participants of the study consisted of 20 men and 28 women who were asked to estimate the age group of the woman in each image and classify it as young (less than 35 years), middle aged (35-50 years) or senior (older than 50 years). In addition, in terms of the grading perception, females estimated age more

accurately than males. As a result of including same gender bias, (Wright and Sladden, 2003) explained that same gender are more accurate in recognising and identifying each other. The study concluded that different skin attributes such as eyes and mouth influenced the process of age estimation. The study highlighted the importance of examining the effect of facial features when judging skin along with considering the age factor. In addition, analysing skin according to age is crucial. Although there are numerous factors associated with skin appearance, when it comes to non-clinical aspects of skin texture such as cosmetology, a study suggested to consider age when analysing skin (Matts, 2008), in addition, emphasised the importance of the perceived age and appearance in women. Thus, we examine the behaviour of human judgement using eye-tracking to learn more about human perception while assessing skin quality and age.

2.2.1 Skin Assessment

Human faces contain too many complex pieces of information for the observer to process. A dermatologist normally assesses facial skin using the naked eye who allocates noticeable skin samples, both live or from photographs, to a recognised quality grade on a predefined grading scale (Ng et al., 2014), which brings in subjectivity to the process. Another limitation is the effect of facial features when it comes to skin assessment (Ng et al., 2015b; Nkengne et al., 2008).

2.2.2 Skin Lesions

Pigmented skin lesions can be described in terms of their colour or colours: red, white, yellow, light brown, dark brown, grey, blue, or black (Cortez et

al., 2020). The pigment may be due to melanin, keratin, blood, or exogenous pigment (e.g. a tattoo) (Caumes, 2020). Epidermal proliferative skin lesions are often classified as melanocytic or non-melanocytic (also called keratinocytic) (Kawahara, BenTaieb, and Hamarneh, 2016). Melanocytic lesions are composed of benign or malignant proliferations of melanocytes and keratinocytic lesions are composed of proliferations of keratinocytes, which may store melanin as shown in Figure 2.2.

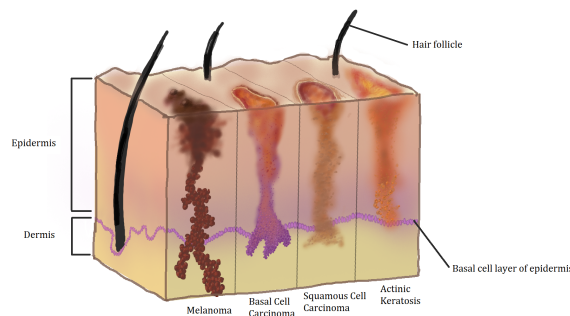


FIGURE 2.2: Image showing different colours of various skin lesion with melanin storage.

2.2.3 Facial Skin

Facial skin analysis is an important part of many disciplines, including dermatology, cosmetology and computer vision (igarashi2007appearance). In the field of computer vision, many applications (Igarashi, Nishino, Nayar, et al., 2007) have been developed to achieve an objective measure of diagnosing skin texture using state-of-the-art machine learning approaches, which are useful in providing an objective analysis (Ng et al., 2014; Ng et al., 2015b). This can avoid problems with repeatability and reproducibility, since a professional's experience and knowledge is subjective and can differ amongst graders. Moreover, it can potentially result in reduced cost and more effective analysis while providing a consistent assessment of skin quality (Prats-Montalbán et al., 2009). There is great importance in providing objectivity to the dermatologist's visual evaluation of skin in order to efficiently develop

effective pharmaceutical treatments (McKenzie et al., 2011). Recently, several skin assessment methods have been established. For instance, analysis of the skin appearance around pores on the face (Mizukoshi and Takahashi, 2014), evaluation of facial wrinkle improvements over time (Luebberding, Krueger, and Kersch, 2014), measuring facial wrinkles using quantification methods and automatic detection (Cula et al., 2013). Most of these assessments were subjective and revolved around a clinical perspective and a professional's opinion rather than an objective assessment. Further research is required to understand the definition of skin quality based on human perception.

2.3 Skin Ageing

The continuous process of ageing occurs with advancing age, and can be accelerated by external factors. This natural process occurs due to reduction in collagen and fibroblasts, as well as dermal mast cells, with time revealing the appearance of dry, wrinkles, and inelastic skin (Landau, 2007). Skin ageing is driven by intrinsic and extrinsic factors, which impact the rise in reactive oxygen species (ROS) with age influences ageing whereby the body becomes increasingly inefficient in removing ROS, therefore degrading collagen and activating collagenases affecting skin structural integrity (Naidoo and Birch-Machin, 2017) Obvious signs of skin ageing can be seen by the appearance of dull and dry skin, age spots, fine lines, and wrinkles (Taylor, 2000).

2.3.1 Natural Aging Process

Changes in skin condition are indicators of the irreversible ageing processes that occurs in the human body throughout a person's life. As we age, the texture and appearance of our skin gradually shift from a smooth, uniform youthfulness to a rougher, more unevenly pigmented skin surface. Facial

wrinkling, sagging, and pore enlargement are marks of the accumulation of photo-damage over long periods of time, and are irreversible without therapeutic intervention. An explanation of the biomechanical and physiological phenomena related to ageing is beyond the scope of this study, and can be found in (Gartstein and Shaya, 1996). Instead, this research investigates facial ageing from a different, more mathematical perspective. Assuming that faces are points in a high-dimensional space (Valentine, 1991), we define the distance between any two faces as their distinctiveness. Negatively correlated with distinctiveness is attractiveness. This is because, as argued by (Rhodes and Tremewan, 1996), less distinctive and more typical faces are considered more attractive. Following this theory and concepts, (Deffenbacher et al., 1998) found that face ageing is related to attractiveness and distinctiveness. Their experiments are based on a 3D face model built on Caucasian male and female young adults. By applying facial caricatures to images of the participants' faces, they found that young faces are more attractive and less distinctive than old faces. When the degree of facial caricature was increased, the faces looked older, the distinctiveness increased, and the attractiveness decreased; all the variable changes observed were linear.

Some authors, including (Braun et al., 2001), believe that the merging of several faces can remove unpleasant asymmetries, irregularities, wrinkles, and pimples so that the skin looks perfectly smooth, clear, and younger. Moreover, the filter makes faces look younger and can enhance attractiveness. This indicates that facial components such as wrinkles and pigments affect the perceived age of a face. (Rexbye and Povlsen, 2007) explored how signs of ageing are read and interpreted in consumer culture. The main indicators of age are biological (skin, eyes, and hair colour), but they are supplemented by vigour, style, and grooming. Indeed, they found that activity and style of clothing are key factors in the interpretation of visual ageing signs. In 2010,

(Aznar-Casanova, Torro-Alves, and Fukusima, 2010) studied the influence of wrinkles on face age estimation. They analysed the qualitative (type of wrinkle) and quantitative (density of wrinkles and depth of furrows) distributions of wrinkles. Their results indicated that the greater the number of wrinkles and the depth of furrows, the older a face was rated. In addition, the quantitative component (density of wrinkles) had a stronger effect on facial age estimation than the qualitative component (type of wrinkle). These findings are also reported in (Mark et al., 1980), where faces with deep furrows were considered older than faces with shallow furrows. Moreover, their experiments showed that the perceived facial age was more strongly influenced by the gender of the face (wrinkles contributed to making male faces appear older than female faces), and that the youngest participants (undergraduate students) attributed, on average, lower ages to the faces shown to them than the more senior participants did.

2.4 Facial Aging Factor

Integral to the ageing process, notions as defined by (Fu, Guo, and Huang, 2010) are to be recognised to further understand and differentiate information about human face as follows:

- i. Perceived age: Using visual appearance to gauge individual age of human subjects.
- ii. Actual age: The individuals true age where it is the cumulated years after birth.
- iii. Estimated age: Using a machine to deduct individual age of human subject from visual appearance.

Factors which influence them relate to skin firmness, controlled by production of collagen. As well as this, the skin's elasticity which is reduced in ageing due to slowing production of elastin, and appearance of sagging skin as a result of reduction of fat cells. Moisture retention, small muscle contractions, slowing of skin shedding process and reduced skin cell turnover are additional contributing factors. Varying factors hold great impact on facial aging factor which are to be taken into account, of which include culture and environment, genetics, ancestry, and trauma or disease. Thus, the suggested characterization of face ageing effects should ideally take into consideration a variety of variables (Ramanathan and Chellappa, 2008).

2.4.1 Intrinsic Factor

Skin ageing is a continuous change in the dermal and epidermal layers caused by extrinsic (sun exposure, polluted air, smoking, and poor nutrition) and intrinsic factors (thin, dry skin, fine wrinkles). With continuous efforts to identify the ageing sequence, progressive patterns have been observed over time, making succession predictable to a degree. Hypervascularity has the effect of causing the epidermis to atrophize, giving the appearance of aged skin. Although the ageing process is unpredictable, it has been observed (Albert, Ricanek Jr, and Patterson, 2007) that noticeable age-related skin changes occur more gradually in the age group 41-50 than in the age group 20-30, possibly as a result of changes in skin texture and tissue elasticity decline. With skin maturation, wrinkling, elastosis, hypervascularity, irregular or blotchy pigmentation, coarseness, laxity, atrophy, dryness and itching are observed. Detected alterations in skin texture, fat atrophy, volume and elasticity loss are a result of age progression, in addition to soft and hard tissue shape and size alterations (Sveikata, Balciuniene, Tutkuvienė, et al., 2011). While soft tissue

changes are apparent in the human face as it ages, bone alterations or remodelling have also been observed (Behrents, 1985), there is evidence for craniofacial bone form alterations, including increase in head circumference, head length, breadth between cheekbones, and face height. Certain alterations in the dentoalveolar region and an increase in anterior facial height have been shown to alter the visual appearance of the lower face. Age-related changes such as sex, disease, sun exposure, weight, drug use, and their associated effects influence computer-based models of ageing.

2.4.2 Extrinsic Factor

Skin ageing likelihood has been shown to be directly affected by immoderate sun exposure, heavy smoking, and obesity seen significantly in periorbital region (Suppa et al., 2011). Such exogenous factors can stimulate ROS production, resulting in damage to skin structural integrity due to oxidative stress. Exposure to Solar Ultraviolet Radiation (UVR) largely drives extrinsic ageing due to oxidative stress and is responsible for > 80 percent of extrinsic facial ageing (Flament et al., 2013). Photo-ageing caused by UVR accelerates ageing and premature ageing, due to collagen synthesis disruption (Rinnerthaler et al., 2015), and was shown to be protective by frequent use of high factor sun protection (Ekiz et al., 2012). UVR exposure can alter skin texture, whereby skin appears thickened and furrowed, also encouraging squinting that results in wrinkles surrounding the eyes (Shaw Jr et al., 2010). Extrinsic phenotypic skin ageing features include telangiectasia, deep wrinkles, coarse skin elastosis, actinic keratoses, and irregular pigmentation (Šitum et al., 2010). Some are observed to a greater extent in fairer skinned individuals relative to darker skin partly due to increase stimulation of melanocyte proliferation (Tobin, 2017). External stressors such as smoking aid in macroscopic skin ageing and wrinkling (Leung and Harvey, 2002) as seen in elderly

participant; this data indicates that skin ageing does not offer a reliable objective measure of cumulative UVR exposure, and care should be taken before it is utilised in this manner. It was discovered that a component of cigarette smoke has a major function in protein carbonylation (Avezov, Reznick, and Aizenbud, 2014) which measures biological age, and is promoted by ROS. Additionally, exposure to factors such as drug usage and stress-related behaviours is thought to contribute to face ageing (Taister, Holliday, and Borrmann, 2000), as well as wind and dehydration.

2.4.3 Perceived age

Make-up application has an effect on the appearance of the face. Despite the fact that applying makeup is a common practise with significant social implications, the mechanisms by which cosmetics affects social perception remain largely unexplored. (Russell et al., 2019) studied the effect of makeup on age perception, they discovered that cosmetics has a variety of effects on the appearance of ageing. When women in their 40s and 50s used cosmetics, they looked to be years younger. However, whether using makeup or not, 30-year-old women seemed to be the same age, while 20-year-old women appeared to be older when wearing full face makeup. Wearing too much makeup will affect the skin and will change the perceived age. Makeup is not a true representation perceived age.

2.5 Skin Quality

As skin quality remains undefined, how it is measured is undetermined; the judgement of skin quality can differ due to observed changes in human skin. Therefore, it is vital to measure skin quality in relation to and based on age, as it is an integral factor. While it is challenging to establish an absolute ageing pattern that may be utilised to measure a certain age, due to changes in skin

texture, skin smoothness, appearances of wrinkles and blemishes, it would affect the measure of skin quality. These can be influenced by idiosyncratic features, gender, genetics, and trauma that influence ageing characteristics and rates.

2.6 Human perception

The face is a vital bodily component due to its many functions in identification and communication. Commonly, dividing the face into an upper, middle, and lower third conveniently assesses morphological effects of ageing (Coleman and Grover, 2006). (Berry and McArthur, 1986) reported that age-related differences in craniofacial development have a significant influence in social perception; if face features are usually indicative of psychological qualities, they may influence perceptions. (Liao et al., 2020) Findings indicated that when casual observers were asked to estimate age, their visual attention moved downward to cover more of the lower face relative to those untasked. With expressive wrinkles versus wrinkles due to age, interpreting human skin can be influenced, it is as well important to account for difference in ageing patterns seen in males and females, as they can also influence human perception.

Facial appearance is a composite of both basic morphologic characteristics and emotional expression and is largely influenced by human perception, repeated facial expressions result in hyperfunctional facial lines, and their appearance may communicate false emotions or incorrect personality characteristics (Cox and Finn, 2005). Moreover, the way individuals tasked to look at faces, it's observed that they tend to draw focus on the eyes more than skin, which can skew human perception and the way they make judgments. Currently, in the area of perception there are difficulties and limitations in computer-assisted automatic age synthesis and estimation. In an

attempt to decipher human perception, analysis shows the perception of a face is at most the sum of its parts (Gold, Mundy, and Tjan, 2012).

Generally, there are a number of studies on the perception of female facial skin appearance. These studies include the investigation conducted by (Fink et al., 2012; Matts and Fink, 2010; Samson, Fink, and Matts, 2010), which illustrates that the skin surface colour and texture of ones face has an influence on the judgment of their age and attractiveness by others. However, (Nkengne et al., 2008) indicate that facial features have an effect on the overall human judgment of the face, especially in females. Human behaviour can be understood by using eye-tracking (John et al., 2017). (Dreiseitl, Pivec, and Binder, 2012) used eye-tracking to assess the difference in the characteristics of pigmented skin areas. It involved sixteen participants of different expertise diagnosing twenty-eight digital dermoscopic images of pigmented skin regions. The study concluded that there is no major difference among people of different expertise in assessing the lesions. In addition, (Krupinski et al., 2014) used eye-tracking to understand dermatologists' assessment before and after online training, in order to test the impact of online training on dermatologist's accuracy and performance. Their study concluded that there are significant differences before and after training. Therefore, this research will adopt the same approach.

2.7 Eye-tracking overview

Eye tracking is becoming an increasingly significant technique in a variety of fields, including human-computer interaction, psychology, computer vision, and medical diagnostics, among others. Humans make snap judgments about others based on their everyday observations, and their impression of others' ages influences their interactions with them (Angulu, Tapamo, and Adewumi, 2018). In a similar vein, human behaviour may be deduced by

other people simply by looking at it with their eyes. Ageing is a personified inevitable process that offers both a dynamic and practical problem in computer vision. As a result, eye-tracking can be employed with computer vision algorithms to understand ageing and human behaviour, albeit it is still a difficult process to complete. Eye tracking devices have been offered as a solution to this problem since they are capable of handling this difficult task. Eye tracking devices are real-time digital image processors that monitor the centre of the observer's pupils and measure the size of the pupils from an infrared video picture of the observer's eyes captured by a camera (Brunyé et al., 2019). That is, the eye tracker takes into consideration vector positions of the observer's eyes and computes the vectors that link the eye position to perceived world locations.

2.7.1 Eye-tracking for human perception

It is a powerful tool in many disciplines, such as diagnostics. Various eye movement characteristics, including smooth saccade pursuit, inhibitory gaze, fixation time, pupil diameter, saccade length, scan route, and Region Of Interest (ROI), (Li et al., 2020) have been examined using a variety of visual stimuli to determine their effects. Nevertheless, in this thesis, skin quality and age estimation were determined using fixation time, fixation count, and length of gaze, all of which were measured. It is possible to use fixation information to determine how much attention individuals have devoted to various stimuli.

Individuals' fixation information may be utilised to determine how much attention they have given to various stimuli. Fixation duration and fixation count are the two measures of attention allocation that are most often employed in clinical settings (Wang et al., 2014). Fixation time (fixation duration) is the length of time spent fixating on anything might indicate how

deeply one has dug into obtaining information. A longer period of fixation suggests difficulties in obtaining information from the item, or it shows that the object is more interesting in some manner than the previous one. Fixation counts are occurrences in which the eye is drawn to a certain region of the screen (Doherty, O'Brien, and Carl, 2010). Based on the object of research, fixations are determined by two parameters: (i) pixel radius, and (ii) minimum duration in milliseconds. But the surroundings will vary based on the subject under investigation, which is why it is such an important component of this thesis. Length of gaze (gaze estimate) is a method that attempts to determine the intentions and interests of users (Tsukada et al., 2011). The link between picture data and gaze direction is addressed by (Hansen and Ji, 2009), and gaze estimation methods are concerned with this relationship. Several eye characteristics (such as pupil size and corneal reflection) derived from image data gathered from single or multiple cameras are proposed by (Chennamma and Yuan, 2013) for the purpose of predicting gaze orientations.

2.8 Facial Skin Assessment

The use of computer-assisted techniques for skin examination is a current research issue that has been ongoing for more than two decades. In clinical practice, it is critical for the early detection of skin cancer and other skin disorders (acne, pigmentation). It is common practice to collect and analyse data on skin quality attributes under the supervision of a well-trained professional who assigns visible skin samples, either in person or from pictures, to a recognised quality grade on a specified scale of quality. Computer vision approaches to assessing skin quality characteristics, on the other hand, are beneficial in giving an objective analysis (Ng et al., 2014; Ng et al., 2015b). Given that a professional's expertise and knowledge are susceptible to interpretation and might differ across graders, this can help to prevent difficulties

with repeatability and reproducibility. Moreover, it has the potential to result in lower costs and more effective analysis, as well as a more uniform evaluation of skin quality (Prats-Montalbán et al., 2009). It is critical to provide objectivity to the dermatologist's visual examination of the skin in order to design pharmacological therapies that are both efficient and successful. Many skin assessment methods have been developed in recent years, including the study of the skin appearance surrounding pores on the face (Mizukoshi and Takahashi, 2014), the evaluation of facial wrinkle improvements over time (Luebberding, Krueger, and Kerscher, 2014), the measurement of facial wrinkles using quantification methods, and automatic detection (Cula et al., 2013). As opposed to objective assessment, most of these evaluations were subjective and focused on the clinical perspective and the professional's judgement rather than on the data itself.

To completely grasp the idea of skin quality as seen by people, more research is required. Computer-aided machine learning approaches have been widely employed in a variety of pattern identification tasks, including the evaluation of skin's overall quality. A brief discussion of the computer-assisted skin cancer assessment approaches is given due to the limited amount of work in face skin evaluation currently available. For the identification of skin problems such as acne (Shen et al., 2018a), computer-based techniques have been developed. Many classification problems were successfully completed using these standard machine learning approaches. They do, however, have certain unintended effects. For example, the number of hidden layers, hidden nodes, and learning rates can all have an impact on the performance of an ANN (Artificial Neural Network). Another downside is that in order to attain ideal performance, the network must be extensively trained, which is why the Support Vector Machine (SVM) (Schmidhuber, 2015) was selected for this experiment as a more appropriate alternative. Over the recent decade, SVM

has become increasingly popular. SVM was used to categorise skin texture in an early melanoma detection technique, as well as to categorise skin colour in a skin colour categorization method (Yuan et al., 2006; Khan et al., 2012). In the picture classification domain, CNNs, on the other hand, outperformed all other methods (Schmidhuber, 2015).

Deep learning algorithms have recently outperformed conventional machine learning algorithms in tasks such as classification, facial recognition, and face tracking. It focuses on comprehending the hierarchical representations of data through the use of a deep architectural model (Wang and Sng, 2015), as demonstrated by the use of a deep CNN by (Krizhevsky, Sutskever, and Hinton, 2012), to categorise high-resolution pictures in the ImageNet LSVRC-2010 competition. There were a total of 1.2 million pictures and 1000 distinct classes used in the training of the network, with error rates of 39.7 percent for the top 1 class and 18.9 percent for the top 5. That is an example of one of the benefits of using this strategy. On the other hand, the data that was employed in that method had nothing to do with skin characteristics. The researchers (Esteva et al., 2017) brought a successful deep learning method to skin cancer to the level of dermatologists, evaluating the network performance against the performance of 21 doctors. Despite this, the study was focused on therapeutic applications. As a result, we will examine the performance of CNNs in the categorization of non-clinical skin characteristics such as spots and wrinkles.

As a result, multiple Conventional Machine Learning (CML) techniques and CNNs will be tested using a variety of skin characteristics in this dissertation. Spots, wrinkles, and normal skin patches were all classified using the parameters and settings we supplied. Following that, in Chapter 3, we compare the performance of SVM (Wang, 2005) and GoogLeNet (Szegedy et al., 2015) algorithms side by side. A brief discussion of the computer-assisted

skin cancer assessment approaches is given due to the limited amount of work in face skin evaluation currently available.

2.8.1 Skin Cancer Assessment

The use of Deep Convolutional Neural Networks on skin lesions has recently been tested, and the results have shown that they can outperform dermatologists in terms of identifying skin lesions, particularly skin cancer. The accuracy of 44 medical physicians, medical students, and dermatologists was tested by the authors (Cho et al., 2019) in comparison to the DCNN. Because of the limitations of the model that was employed in the study, the Area Under the Curve (AUC) values were quite similar amongst each other. It is essential to note, however, that the number of dermatologists participating in this trial is strictly limited to 18. Another research, carried out by (Maron et al., 2019), compared the performance of 112 dermatologists with a computerised approach in the detection of skin cancer in a different setting. A total of 11444 pictures were used in the study, with 6390 of those photos being biopsy confirmed. In addition, the findings indicate that DCNN methods are effective in distinguishing between malignant and benign skin lesions, and that they outperform skin experts in identifying five different skin lesions, which are as follows: i) solar keratosis, ii) basal cell carcinoma, iii) benign keratosis, iv) melanocytic nevi, and v) melanoma. Another machine learning technique is to use features that have been handcrafted. Using residual learning techniques to overcome the limitations of deep networks is a promising approach (Carcagnì et al., 2019).

(Carcagnì et al., 2019) and colleagues utilised an ensemble approach to classify several categories. The suggested technique is capable of categorising

seven different types of skin lesions. Initially, the model is based on a multi-level baseline of DenseNet-121, which employs two transition layers in addition to the first two dense blocks, with the remainder of the dense layers being reduced from the original implementation. Classification was conducted out using the SVM classifier on the ISIC 2018 data (Tschandl, Rosendahl, and Kittler, 2018). The unbalanced data, on the other hand, was dealt with by employing certain data augmentation methods on the classes with little sample data, such as rotation, flipping, and affine transformations. After that, there is a centre cropping as a preprocessing for the network. Because the discriminating feature could not be seen, the softmax and the centre loss were combined to produce a loss that was visible. The suggested model was evaluated by comparing its accuracy, recall, and F-score to the original DenseNet-121 model. Other approaches, such as colour constancy to reduce data unevenness, segmentation phase to increase network performance, and more labelled data are proposed as a result of the research.

Another approach examined utilising segmentation as a guide to enhance the network and obtain more prominent characteristics in order to increase recognition accuracy (Premaladha and Ravichandran, 2016; Yu et al., 2016). In addition, (Yan, Kawahara, and Hamarneh, 2019) designed AttenMel-CNN, which is an end-to-end trainable attention model for melanoma detection. They propose that past knowledge be used more efficiently by regularising the attention map with the region of interest. As a result, the performance of the models is enhanced. The model structure is based on a modified version of VGG16. Despite this, all of the dense layers have been eliminated and replaced with a few new ones. Two attention models and three vectors were used, and they were concatenated with the classification layer to get the final result. The datasets from ISIC 2017 by (Codella et al., 2018) and ISIC 2018 by (Tschandl, Rosendahl, and Kittler, 2018) were used; however, due to

the variability in the data, the investigation was limited to melanoma rather than all other cancer types. In order to deal with imbalanced datasets, the focused loss (a modified version of cross entropy loss) was developed. In addition, data augmentation with random rotation, flipping, and cropping during training, centre cropping was utilised during the preprocessing step. When comparing the findings with state-of-the-art approaches, it was necessary to determine the average precision and area under the curve. Despite the fact that the findings are the same, their model is able to localise the characteristics of the lesion. The drawback of their study is that they only classified two types of skin cancer, while there are more malignant skin diseases other than melanoma in the skin cancer genus.

In clinical, there is a desire for computer-aided solutions that are self-explanatory. (Barata, Marques, and Emre Celebi, 2019) used clinical experts' knowledge to improve the performance of lesion diagnosing systems. This was accomplished by developing a hierarchical organisation of skin lesions that was defined by dermatologists; and guiding the classification decision by using attention modules that identify relevant parts of the region. It is composed of three major blocks: a decoder that receives the picture and performs feature extraction, a decoder that classifies the class region on a regular basis, and a classifier that classifies the class region on a random basis. The attention module is used to guide the decoder to various areas in order to increase the transparency of the network Figure 2.3 and to visualise the network structure. The investigation made use of two datasets; ISIC 2017 and ISIC 2019. Lesions were first categorised as either melanocytic or non-melanocytic in nature. Then there is the degree of malignancy, which determines whether the lesion is benign or malignant. Finally, the determination of the diagnosis (e.g., melanoma, nevi, basal cell carcinoma or vascular lesion). In addition, their model generates an attention map for each class that is used. The

model combines CNNs with Long Short-Term Memory Network (LSTM) and attention modules to create a more complex model. Furthermore, ensemble techniques were employed without relying on any additional external data. Colour constancy was performed as a preprocessing step in order to normalise the colour variance. In order to increase the performance of the model, the following data augmentation techniques were used: random cut, random flip, and random colour transformation. The model was assessed

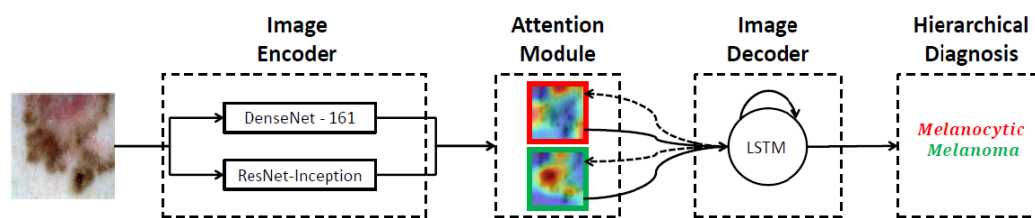


FIGURE 2.3: Barata, Marques, and Emre Celebi, 2019

using the following criteria: sensitivity, specificity, and overall balanced accuracy (81 percent), as well as the area under the curve for each class. The model's attention was first drawn to the surrounding skin, but when it came to conducting the final classification, the model was more particular and localised. The study demonstrates the model's capacity to identify clinically relevant areas to diagnoses while also providing helpful information about different regions for different illnesses.

2.9 A review on Face Datasets

This section discusses the different types of databases that have been used by researchers over the years for facial analysis (wrinkle detection, face age estimation, and facial wrinkle analysis). A face constitutes a unique identity for an individual such as age, gender and ethnic background. There is ongoing research in facial analysis application such as face recognition, gender classification, facial skin assessment, face age estimation and facial expression

recognition (Meyers and Wolf, 2008) and (Han, Otto, and Jain, 2013). Cosmetic and facial analysis applications are now recently being researched to facilitate cosmetic surgery, facial modification (i.e. cosmetic make-up) (Bertacchi and Silveira, 2019). This is because people are more and more focused on their appearance (Arakawa, 2004) and facial beauty (Ohchi, Sumi, and Arakawa, 2010; Batool and Chellappa, 2014). The performance of AI algorithms on facial analysis applications depends on the information provided. Customers using virtual try-on technology for cosmetic or medical facial lift will be more satisfied. However, some of the databases used do not provide accurate age estimation before make-up and after make-up information. An overall comparison of these datasets has been displayed in Table 2.1.

The **FG-NET** model (Fu et al., 2014) used an aging database released by (Panis and Lanitis, 2014). The database was created to support research in understanding the changes in facial appearance caused by ageing (Lanitis, 2008). This database has been mostly used in carrying out research in age estimation, age-invariant analysis, face recognition and age progression. FG-NET is made up of 1002 images from 82 different subjects. The age range of the database is from newborns to 69 year old, with 40 years being the most populated age group of the database. The images of this database were collected from scanned photographs of personal collections of the various subjects on the dataset (Panis and Lanitis, 2014). This leaves the image quality on the photographer, the imaging equipment and the overall condition during the photography. The images on the FG-NET vary in resolution, sharpness and illumination combined with the face view angle and facial expression. Some images also have occlusions in the form of facial hair, hats, and spectacles.

Bosphorus Dataset The Bosphorus dataset is made up of 106 subjects of high resolution 2D colour images with varied sizes of forehead wrinkles which are

TABLE 2.1: Summary of existing face datasets

Dataset Name	Total Images	Total Subjects	Age Range	Resolution	Expression	Makeup	Gender	Colour Checker	Year
FG-NET	1002	82	0-69	-	Yes	-	M/F	-	2004
Bosphorus	-	105	-	512x768	Yes	Yes	M/F	-	2012
FERET	14,126	1199	10-70	-	Yes	-	M/F	-	2000
MORPH	55134	13000	16-77	400x500	Yes	-	M/F	-	2003
IFDB	3600	616	2-30	480x640	Yes	No	M	No	2007
KFDB	52	1920	20-59	640x480	Yes	-	M/F	-	2004
IMPA-	534	38	20-50	Yes	640x480	-	-	-	2008
FACE3D Caltech	450	27	-	896x592	-	-	M/F	-	-
Faces FEI PUT	2800	200	19-40	360x260	Yes	-	M/F	-	2005
Face Database CAISIA-	10000	100	20	2024x1536	No	No	M/F	No	-
WebFace LFW	202,599	10,177	-	-	Yes	-	M/F	-	2017
CelebFaces	13,233	5,749	-	250x250	Yes	-	M/F	-	2007
CelebA	87,628	5436	-	-	-	-	M/F	No	-
PAL	202,599	10,177	-	-	-	-	M/F	No	-
CARC	>1000	576	19-93	-	Neutral	-	M/F	-	2004
CACD	160000	2000	16-62	-	Yes	-	M/F	-	2014
CAS-	163,446	2000	16-62	-	Yes	-	M/F	-	2014
PEAL CMU-	99,594	1040	-	360x480	Yes	-	M/F	-	2005
PIE	41,368	68	-	640x480	Neutral	-	M/F	-	2007
YGA	8000	1600	0-93	60x60	Yes	Yes	M/F	-	2007
WIT	26222	5500	3-85	32x32	Yes	-	M/F	-	2010
HOIP	306600	300	15-64	640x480	Neutral	-	M/F	-	-
LHI	50000	8000	9-89	120x160	-	-	M/F	-	-
YMU	604	151	-	-	No	Yes	F	No	-
VMU	204	51	-	-	No	Yes	F	No	-
MIW	154	125	-	-	No	Yes	F	No	-
MIFS	642	107	-	-	No	Yes	F	No	-
AI&R	34,17	22-61	640x480	-	-	-	M/F	-	-
BCFD	-	147	20-62	531x704	-	-	M	-	1995
PubFig	60000	200	-	-	Yes	-	M/F	-	2009
MIT	>2500	154	-	480x640	-	-	M/F	-	-

cropped for the face region. The Bosphorus dataset, introduced by (Savran, Sankur, and Bilge, 2012), is a 3D and 2D dataset made up of images showing facial expressions and systematic variations in poses with different occlusions. The images were captured with a structured lighting system for 3D and normal lighting system for 2D respectively. Due to the lighting system, the subjects on the database portrayed specular reflections on the face which does not affect the image texture but could cause noise in the 3D data. (Savran, Sankur, and Bilge, 2012) prevented this by applying a special powder on the face which did not change the skin colour of the subjects face. The process of enabling a range of facial analysis task on a 3D database from expression to recognition depend on the following: i) Facial Action Unit Coding System action portraying ground-truth for both single and compound. ii) Emotional expressions. iii) Ground-truthed poses. iv) Spectacles, hand and hair occlusions. The dataset is made up of 105 subjects with 31-54 samples per subject with three occlusions (hair, hand, spectacles), and they comprise 34 facial expressions of action and six emotions displaying systematic head poses (13 yaw, pitch and cross rotation) (Savran, Sankur, and Bilge, 2012). A total of 18 subjects on the database have beard/moustache and 15 other subjects have short facial hair. The dataset is made up of 60 men and 45 women with an age range of between 25 and 35 with most of the subjects caucasian. With up to 54 face scans available per subject and 34 of these having 31 scans due to fewer number of expressions. Also, included in the dataset are faces of 27 professional actors/actresses. The total number of face scans is 4666 with each manually labelled for 24 facial landmark points subject to their visibility under the given scan. The 2D face images of this dataset according to (Ng et al., 2015a) were acquired with an ordinary camera under good illumination conditions. Each subject of the dataset is comprised of several images of different poses and facial expression. (Ng et al., 2015a) carried out their

experiment based the manually repeated cropped images following a finding based on (Batoool and Chellappa, 2014) ground truth by hand labelling proving to be more reliable for obvious observation in forehead wrinkles for most facial images. The original sizes of the images of the Bosphorus dataset were 512×768 pixels.

FERET Database The Facial Recognition Technology (FERET) database is a collection by the George Mason University and US army research laboratory that began in September 1993 (Phillips et al., 2000b). In 1996, the collection of the database was a total of 1564 sets of images consisting of 14126 total images. There are 1199 subjects of which 365 are duplicate images. With 9-20 different poses, made up of 2 facial expressions, 2 different illuminations at 2 different times each. A number of images of subjects between 508 to 980 were collected following the poses; right and left profile, right and left quarter profile and right and left half profile. The age range on this database is between 10 and 70. The FERET database consist of six age groups (ng2015will) of (10-19, 20-29, 30-39, 40-49, 50-59, 60-70) with the total images for each group equal to 38, 950, 629, 474, 209, and 66 respectively. In total, there are 14051 gray scale images which are 256×384 pixels in size. The color images on this database have been standardized by the National Institute of Standard and Technology to an original gray-scale image of high resolution (517×768). Ng et al., 2015a in their study "Will wrinkle estimate the face age" used the FERET database in their search for age estimation using wrinkle detection. They mentioned that, wrinkle-based features (skin textures) can be an effective feature for face representation due to spatial location and orientation selectivity. The images of the FERET database have high resolution compared to the FG-NET, hence the reason they used the FERET database for their findings. They proposed the Hybrid Hessian Filter (HHF) to detect

wrinkles on the forehead of images. Facial wrinkles according to this study are repeated facial muscular movements and expressions with unique characteristics depending on gravity and frequency of movement. Their study presents a method for estimating age through the localization and counting of wrinkles, which is unique because this varies by individual. They used a multi-scale filter to assess geometrical structures of the skin due to the variations in sizes of the wrinkles. Their analysis was based on multi-level Hessian eigenvalues to emphasize the local behaviour of an image thus identifying the wrinkle. Their proposed method uses a Face++ detector suggested (*Face Detection*) for face detection and feature extraction. This was followed by a face mask with ten pre-defined fixed coordinates for eyes and mouth and wrinkle region for normalization. The mask is cropped and scaled to the original size of the image on the FERET database images.

MORPH Database: The MORPH database is a publicly available database that contains multiple individuals at different ages across time. It was initially created in 2003 by the Face aging group at the University of North Carolina Wilmington and expanded up till 2008 by (Albert and Ricanek Jr, 2008). The database contains two albums. Album 1 contains 1690 images from 515 individuals (men and women) of various ancestry groups and album 2 contains 15204 images from more than 4000 individuals. The age range of these individuals is between 15 to 68 years and the images organized by "decade of life" categories. Album 1 has images of size 400×500 pixels that reveal the head and the face. This album(1) has 1253 images containing individuals of African American descent, 434 images containing individuals of European American descent and 3 other. The total number of men are 1405 and women are 285 with the majority of the individuals at the age between 18-29.

The minimum age is 15 years and the maximum age is 68 years. Each individual on the database has an initial image described as the youngest-age image and additional images as age progresses are added to their file on the database. Album 2 has 15204 images(12984 male images and 2220 female images) with the majority of the individuals between the ages 40 and 50 years. The metadata(age, sex,ancestry,height and weight) have been recorded on the database. In a later version, the MORPH database is described by ("[Preliminary Studies on a Large Face Database MORPH-II](#)") to contain 55134 images with more than 13000 individuals spanning over the years, the author describes it as the MORPH II and contains 77% black faces, 19% white face and the remaining 4% other for a total of over 55000 images.

Iranian Face Database(IFDB) The Iranian Face database (IFDB) is a large database of Iranian subjects at ages between 2 and 85 year old. It was collected between January and February 2007 by the department of Engineering,Islamic Azad University of Karaj. (Bastanfard, Nik, and Dehshibi, [2007](#)) It contains over 3600 colour images corresponding to 487 men and 129 women totaling 616 of peoples faces. It is a large dataset that can support age classification system. The database contains images with no restrictions to occlusions such as hair, spectacles, make-up and cloths to participants. Ground-truth information of ID,age,kind of pose or expression and spectacle is provided per subject. The subjects were photographed with fine resolution digital camera in normal light. The images are colour images of 480×640 pixels in resolution and 24 bit depth and 40K bytes size in JPG format. The subjects were photographed on a background without any flashes in daylight as enough luminosity for wrinkle detection and facial feature extraction without shadows is needed.

Korean Face Database(KFDB) Korean Face Database(KFDB) is a large database of 1920 subjects Roh and Lee, 2007 of Korean faces that include not only images but also ground truth information and description of the files. It is made up of 52 images under different conditions such as illumination(21 images), Expression(10 images), pose(21 images). There are 100 subjects above the age of 50, 300 between 40-49, 300 between 30-39, 300 between 20-29 and 100 below the age of 19. The image sizes are 640×480 pixels and 24 bit colour depth stored in both BMP and JPEG format. The description(age, gender, birth-place, image conditions and date) and ground truth files are stored in ASCII text. The images on the database were taken in a studio consisting of an LCD monitor, an octagonal prism frame with 7 cameras(NTSC, CCD camera) and 16 lights. The subjects can view their faces on the LCD monitor in the studio.

IMPA-FACE3D Database: was created by (Mena-Chalco, Cesar-Jr, and Velho, 2008). It is made up of 38 subjects with 534 static images with different facial expressions(neutral frontal, joy, anger and fear). The database is made up of 22 male and 16 female with ages between 20-50 with image sizes of 640×480 pixels.

Caltech Faces Database:The frontal faces dataset is collected by Markus Weber at California institute of Technology. It contains 450 face images . 896×592 pixels in JPEG format. 27 different people under different lighting conditions, expressions and background.

FEI subset of frontal faces is a Brazilian face database that contains a set of face images taken between June 2005 and March 2006 at the artificial intelligence laboratory. It contains 14 images for each of 200 individuals, a total

of 2800 images. The Subset of the face dataset is cropped to size 360×260 pixels with 400 full frontal faces of neutral or non-smiling expression. The subjects on this database are of ages 19-40 years with an equal number of male and female. All the images were taken on a homogeneous background in an upright frontal face pose and a profile rotation of 180 degrees.

Radboud Faces Database(RaFD) is a database collection by (Langner et al., 2010) consists of 67 models displaying 8 emotional expression. The models were trained according to the Facial Action Coding System(FACS) to show different facial expressions. This was photographed simultaneously from five different camera angles in a highly controlled environment. The dataset contains images of both adult and child models with the same stimulus and under the same level of technical control. The images on the database were cropped to the size 1024×681 pixels. They used five Nikon cameras (models D200,D2x, and D300) with resolution between 10 and 12 Mpx. Three 500W flashes were used for illumination.

PUT Face Database The PUT database is a colour database developed by the Poznan University of Technology in Poland. It contains about 10000 images of pixel size 2048×1536 of 100 subjects taken under controlled illumination conditions. It contains coloured images and is publicly available for research purposes (Kasinski, Florek, and Schmidt, 2008). The subjects in the database consist of 22 face orientations under the same lighting conditions. Nearly all the images are of high resolution and of white males in their early 20s with neutral expressions and without glasses.

The large **Age-Gap Face dataset (LAG database)** has 3828 images of 1010 celebrities designed in 2017 from LFW. It contains at least one child/young image and one adult/old image. The images are aligned with the eyes in a

horizontal position and scaled with a fixed distance between the eyes and cropped to 200×200 pixels (**bianco2017large**). The images are publicly available for download in (Bianco, 2017).

CAISIA-WebFaces Database This dataset according to (Yi et al., 2014), this is currently the largest dataset for face verification and identification. The dataset contains 494, 414 images of 10575 celebrities both male and female. Each subject has approximately 47 images with no attribute labels.

Labelled Faces in the Wild (LFW) (Learned-Miller et al., 2016) introduced a dataset that contains images of celebrities spanning a range of conditions of everyday life. It is more natural with variation in factors such as illumination, pose, race, expression, occlusions and background. This dataset was designed in 2007 and contains 13,233 images of 5,749 subjects, with the number of images per subject ranging from 1-80. It was designed to study face recognition in unconstrained environments (Huang et al., 2008), and the majority of the images per subject in the range of 2-5. The images are 250×250 pixels in size, 1680 of the people have two or more distinct photos in the dataset with an individual description. It is publicly available in (Huang et al., 2012).

CelebA CelebFaces Attributes Dataset (CelebA) is a collection of 202,599 facial images of celebrities (Liu et al., 2015), each with 40 binary attribute annotations. The images are cropped to size 178×218 pixels as of 2015. There is large pose variation and background clutter associated with the database alongside a large diversity and rich annotations. This makes it an appropriate test set for facial image synthesis (Shen et al., 2018b).

PAL One of the largest datasets is the PAL dataset with a wide age range

of adulthood with a maximum age of 93 years. The images on this database are made up of African-American, Caucasian and other ethnic background and put together by Minear and Park(2004) from PAL. The dataset contains 576 individuals with over 1000 colour images (Ebner, Riediger, and Lindenberger, 2010). In this database, there are 218 faces of adults aged 18-29, 76 faces of ages between 30-49, 123 faces of adults ages between 50-69 and 158 faces older than 70. The faces are labeled for each image and classified according to gender with the approximate age as of 2014.

CARC The Cross-Age Reference Coding(CARC) is a large-scale image dataset available as a reference set on the internet. The database is a retrieval of images across age from the Cross-Age CElebrity dataset(CACD). It contains more than 160000 images of 2000 celebrities with an age range of between 16-62. It was created in 2014 by (Chen, Chen, and Hsu, 2014)for Age-invariant Face recognition and Retrieval.

CACD Cross-Age Celebrity dataset(CACD) is face dataset of celebrities collected from the internet with more than 163,446 images from 2000 celebrities. The use of search engine with key words such as name, year(2013-2014) was used during the collection of the dataset. The ages of the celebrity is just an estimate as it was calculated by a simple subtraction of the birth year from the year when the photo was taken. The age range is from 16-62. The dataset is publicly available and it has images with facial expression of celebrities (Hoon Yap et al., 2018).

CAS-PEAL Database CAS-PEAL is a database designed by the Joint Research and Development Laboratory for advanced computer and communication technology(JDL) of Chinese Academy of Sciences(CAS). (Gao et al.,

2007) This database was designed for researchers of face recognition technology and contains images with factors such as Pose, Expression, Accessories and Lighting (PEAL). It contains 99,594 images of 1040 subjects (595 male and 445 females). Each subject was captured simultaneously by 9 cameras setup in a semi-circular shelf across different poses in a shot. A further 18 images were captured in other two shots and all of the images captured were of size 250×187 s. Five kinds of expressions (surprise, smile, close eyes, frown, open mouth) were captured. Six kinds of accessories (3 glasses and 3 caps) with variations in background alongside 15 lighting conditions, distance to the camera and age variations were also taken into consideration. Only 30900 images out of the 99594 are on current release in gray-scale of size 360×480 s.

CMU-PIE Database (Sim, Baker, and Bsat, 2002) from Carnegie Mellon University-PIE (Pose, illumination, Expression) database was constructed with samples varying in facial expressions under good illumination conditions and a large number of poses. It contains 41,368 images obtained from 68 individuals imaged in a 3D room with a set of 13 high-quality synchronized colour cameras with 21 flashes. The image sizes are 640×480 s in size of which display neutral face, smile and close eyes to simulate a blink. In addition, 60 frames of recording was captured per subjects using three cameras (frontal, 3/4 and profile view).

YGA This is an in-house (internal database) by University of Illinois at Urbana-Champaign (UIUC) age database containing 8000 images of high-resolution. It holds an equal number of male and female (800) of Asian origin between the ages of 0 to 93. The face images were cropped and resized in gray-scale format to a size of 60×60 s. The description of each image has ground truth age information with five near frontal images per subject. The subjects have

significant variation in illumination, facial expression and make-up due to the images photographed in a street environment.

WIT-DB Waseda human-computer Interaction Technology Database developed with subjects of Japanese origin of an actual age-group 3-85. The age labels are divided into 11 nonoverlap groups. The database contains 26,222 face images(14,214 male face images and 12,008 female face images) (Fu, Guo, and Huang, 2010). There are 5,500 Japanese subjects(3000 males and 2,500 females) with each subject containing 1-14 samples.The images were photographed at different illumination backgrounds. The faces of the subjects are without occlusions with normal facial expressions and smiles in some images. The sizes cropped to the facial region to 32×32 s.

Human and Object Interaction Processing(HOIP) (Escalera et al., 2016) shows this database contains 306,600 images of 300 equal male and female subjects. The age varies between 15-64 with a five year interval gap. It contains facial images of neutral expression of 640×480 s and 24 bit colour depth.

Lotus Hill Research Institute(LHI) Suo et al., 2007; Suo et al., 2008) This database is an adult database of Asian face made up of 50000 mid-resolution images of subjects at various age groups. The images were taken with little illumination and pose variations. About 8000 colour images of 120×160 s of this database has images of subjects in the age range 9-89 with approximately 100 images per subject.

AI&R This database is an internal Asian face database(Institute of Artificial Intelligence and Robotics Xi'an Jiaotong University,China) with four subsets AI&R V1.0 (Facial expression), AI&R V2.0(Aging), AI&R V3.0(View) and

AI&R V4.0(Illumination). (Fu and Zheng, 2006) The images are 640×480 s stored in JPEG with 24bit color depth. AI&R V1.0 contains 300 images of 20 subjects taken under same illumination condition. AI&R V2.0 contains 17 subjects between the age of 22-61 with 34 frontal images. AI&R V3.0 160 subjects with images taken at 11 different views which makes it 1,760 images. AI&R V4.0 has 104 images of 13 frontal faces in eight variant illumination conditions.

Burt's Caucasian Face Database (Burt and Perrett, 1995). This database contains images with neutral expression of 147 Caucasian male of European descent, aged between 20 and 62. The subjects on the database are without occlusions(facial hair and glasses) and without any make-up. The images are of resolution 531×704 taken under the same illumination condition and are 24 bit color in depth.

PubFig Databse This is a dataset containing images of public figures(Kumar et al., 2009). It was created in 2009 inspired by the LFW dataset which was created in 2007. The dataset contains 60000 images of 200 subjects with approximately 294 images per subject. No information regarding the age gap, and other imaging conditions have been given. Its large number of images per subject gives the advantage over LFW for the construction of subsets of images with different facial expressions and illumination.

Age-cGAN (Antipov, Baccouche, and Dugelay, 2017) is a dataset of about 120,000 images which is a subset of the **IMDB-Wiki dataset** (Rothe, Timofte, and Van Gool, 2015). It has different age categories that range from 0-60+ years old. The datasets are available for research purposes. Some of these datasets are publicly available and can be downloaded from the internet.

Other datasets are private and can be accessed by request via relevant authorities.

Current Datasets have various limitations that make them not suitable for face analysis. Table 2.1 gives a brief overview of these datasets. In order to perform a test of a skin analysis algorithm, it is important to use a dataset free of makeup. Although other conditions such as illumination may influence the results, special care (using a colour checker) should be considered in designing datasets used for face analysis in future. Most of the datasets in this research have existed for a while and were used for face detection and recognition. Some of them do not tell if the subjects had make-up or no make-up or whether a colour checker was used during its construction.

Several challenges still exist with using existing datasets as they were originally designed for other facial analysis applications due to the illumination conditions of the photographed images (Batool and Chellappa, 2016; Mathew, 2016). In other research (Batool and Chellappa, 2014), images have been downloaded from the internet with no information with regards to makeup. Therefore, it is important to have full knowledge of a database before its use on algorithms implemented for the analysis of face detection and its applications. The information with regards to make-up is vital for the performance and analysis of the results. Makeup datasets are considered as a result of its application effect on skin appearance, changes the colour, contrast and texture of the face (Dantcheva, Chen, and Ross, 2012).

YMU (Dantcheva, Chen, and Ross, 2012) consists of 151 Caucasian female faces that was collected from YouTube makeup tutorials. Every female was captured four times, two before and two after the makeup application. However, MIFS (Chen et al., 2017) is a dataset that was obtained the same way

but was not limited to Caucasian female, the data collected from random makeup tutorials with total of 107.

MIW (Chen, Dantcheva, and Ross, 2013) used FRGC (Phillips et al., 2000a) dataset and applied synthetic make-up on total of 51 Caucasian female. However, these face data is associated with several limitations such as lighting console, camera position and facial expressions as the images were captured from online tutorials. To date there is lack of high resolution female faces for non-medical approaches, therefore make-up dataset was created in [section 3.3](#).

Currently, there are limited datasets available for the analysis of skin conditions. An available dataset called DermNet consists of a over 23,000 images of various skin diseases. However, this dataset has two limitations. One limitation was that the data collection was not under a controlled environment, which has caused inconsistencies in the images and affected their integrity as well as their accuracy. Another limitation was that the images were not only of facial skin conditions, but also of different diseased body parts, which are unsuitable for this experiment focusing on the classification of common facial skin conditions. To address these limitations, we proposed an ongoing collection of consistent, high-quality images of faces from a wide demographic and from participants who engage in different social habits (see [section 3.2](#)).

Hence these datasets are not reliable for the testing of the algorithms that are relevant to this research. Therefore, designing a suitable dataset is necessary for this research as it will determine the performance of the algorithm. To address these limitations, this thesis introduces two high-quality images of faces from a wide demographic and from participants who engage in different social habits (see [section 3.2](#) and [section 3.3](#)).

2.10 Summary

With relation to skin quality, but not exclusively, this chapter gives an overview and comprehension of age estimate and analysis, as well as changes in facial appearance associated with ageing. In addition, the cosmetic applications that are used to evaluate the skin's quality are highlighted. According to the findings of this study, there aren't nearly enough automated computerised techniques for evaluating skin. As a result of this evaluation, our research is focused on understanding human perception of skin with and without face characteristics, with the goal of better understanding dermatologists' judgments. In addition, we investigated datasets that are utilised for facial analysis applications in this work. As a result, we discovered that the candidates were not excellent enough for the aim put forth for this thesis. Therefore, we gathered an image dataset for use in our facial analysis, which will be appropriate for the purposes of this study. The current face datasets that were used in this research will be presented in 3. The following are examples of study directions:

- Due to limitations in high-resolution datasets, we will be collecting images in a much better controlled environment.
- Better understanding of human age estimation.
- The definition of skin quality in relation to age from humans perspective.
- Better understanding of the similarity of machine vision versus human perception.

3 Preliminary Research and Datasets

This chapter will introduce two new high-resolution face datasets: social habits and make-up dataset. This is due to a lack of high-resolution face datasets, as mentioned in [section 2.9](#); however, both datasets are useful for algorithm development and validation. Furthermore, two preliminary experiments were carried out, in which an automated wrinkle annotator and an automated acne detection system were proposed and tested on these datasets.

3.1 Introduction

Over the years, researchers have proposed different approaches for skin analysis. Despite the development of face algorithms using existing datasets, the resolution, data availability and controlled setting in which the datasets were captured pose limitations. In order to proceed with research in this area, access to high quality images plays a key role. Therefore, it is important to understand which datasets can be used for face detection and analysis.

[section 2.9](#) examined some of the limitations of face datasets, which are unsuitable for testing the algorithms that are the subject of this research. As a result, developing a good dataset was a critical need, as it is from data that an algorithm's performance comes. To overcome these restrictions, this thesis presents two datasets of high-quality photographs of faces drawn from

a diverse demographic and from subjects who exhibit a range of social behaviours, facial expressions, and cosmetics application. These datasets are described in this section. To benchmark the new datasets, this thesis proposed two computer methods: automated wrinkle annotator [section 3.4](#) and automated acne detection [section 3.5](#).

3.2 Social Habits Dataset

The first dataset, Social Habits Dataset, focuses on participants who engage in social habits that can include, but are not limited to, smoking and alcohol consumption. The dataset received approval from the Manchester Metropolitan Research Governance and Ethics Committee. The dataset consists of 164 images of participants with a mean age of 48.43 (standard deviation (SD): 21.44, ages between 18 and 92). There are 25 different self-reported ethnicities in the dataset, including Caucasian, African, Arabic, Chinese and Malaysian. The ethnic group with most participants are white British, with 119 images. The main reported gender is females (107), followed by males (56) and one transgender.

To understand how certain habits can affect a person's facial skin properties, participants were asked to complete a questionnaire asking if they consumed alcohol or smoked. Overall, 67 participants reported they did not drink alcohol, 88 reported they were habitual drinkers and 8 used to but stopped. As for smoking, 85 people were non-smokers, 21 habitually smoked tobacco in some form, 1 uses an electronic cigarette only, 6 had partaken in smoking a few times in their lives and 51 used to smoke but stopped. The images were taken with a Nikon D5300 at a resolution of 4496×3000 to capture as much detail as possible of participants faces.

Firstly, five expressionless images of each participant were captured at different angles to allow for a full view of the face and its profiles Figure 3.2. Next, participants were asked to pose in six different facial expressions based on Ekman's (Ekman, 1999) universal facial expressions: happiness, sadness, surprise, disgust, anger, and fear, shown in Figure 3.1 Replicating these expressions allows the dataset to include some variation in the way the facial skin of participants deforms due to natural expressions. Being able to differentiate between an actual wrinkle and ridges caused by expression lines would be extremely useful when analysing facial conditions in the future and will allow to distinguish between natural expression-caused deformities and any others that were caused for other reasons, i.e. aging and social habits.



FIGURE 3.1: The facial expressions captured



FIGURE 3.2: The angles at which photos were taken.

3.3 Make-up Dataset

The second dataset is called Make-up Dataset; it too received approval from the Manchester Metropolitan Research Governance and Ethics Committee. To maximise the appearance of facial skin in the dataset, four high resolution cameras were located in four different angles (0, 20, 45, 90, as shown in Figure 3.3).



FIGURE 3.3: Cameras location and experiment setup.

Figure 3.5 displays the staging of the face data; the image on the left shows the natural face without makeup, the image in the middle shows the same face with foundation on, and the image on the right shows the face with full make-up.

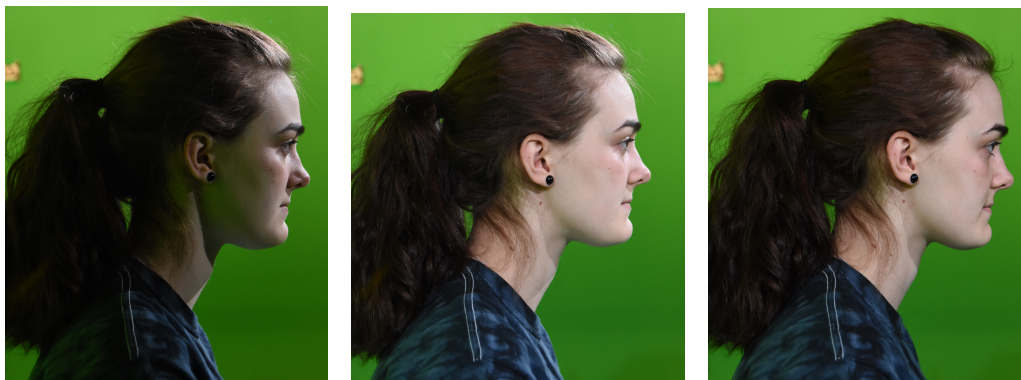


(a) Natural

(b) Foundation

(c) Full make-up

FIGURE 3.4: The make-up data collection stages.



(a) Light off

(b) Light on

(c) Both sides on

FIGURE 3.5: The lighting variation at 90° .

At each stage of the make-up dataset collection, the facial expressions also were captured.

3.3.1 Standardization of combining both datasets

The datasets described in the previous sections have many similarities (ID naming convention; angles from which photos are taken; participant age, gender, and social habits information), making their combination into one, larger dataset straightforward.

3.4 Experiment I: Automated Wrinkles Annotator

The appearance of wrinkles is affected by many factors. Even though wrinkles are highly associated with ageing, it is observed that some individuals have less wrinkles than others. The patterns and rates with which wrinkles grow are still not well understood. Wrinkle detection has gained popularity recently and many automated computerised methods were proposed to localise wrinkles (Batool and Chellappa, 2012; Cula et al., 2013; Batool and Chellappa, 2014; Ng et al., 2014; Ng et al., 2015b). Although some algorithms report good reliability (Ng et al., 2014; Ng et al., 2015b; Batool and Chellappa, 2012; Batool and Chellappa, 2015), there are still a few limitations to previous works:

- The majority of them were only validated on forehead datasets.
- Due to the majority of wrinkles being horizontal (Albert, Ricanek Jr, and Patterson, 2007), some algorithms work only on horizontal wrinkles (Ng et al., 2014; Ng et al., 2015b)
- Wrinkle detection was evaluated based on wrinkle lines (line segment after thinning process), not wrinkle regions.

- Most algorithms are unable to separate coarse wrinkles from fine wrinkles.
- The existing algorithms were not able to provide a clear definition of wrinkle depth.

To address the issues above and the limitations of human annotation, an automated facial wrinkles annotator for full face high resolution images is proposed here. The proposed algorithm represents wrinkles as regions (as opposed to lines), and is able to differentiate between fine and coarse wrinkles. The key contributions of this algorithm are:

- Automated annotation of coarse and fine wrinkles regions on high resolution face images.
- Generation of Probabilistic Wrinkle Map alongside with the wrinkle regions to provide wrinkle depth information.
- Demonstration of the robustness of the algorithm on the two datasets described above.

3.4.1 Proposed Method

Only high resolution images were used to demonstrate the capability of the proposed method, particularly in detecting fine wrinkles. To generate the ground truth from human annotation, it was impracticable to manually annotate a large-scale dataset, as it is too time consuming a task. Therefore, a subset of 20 full-face images (10 images from the Social Habits dataset, 10 images from the FERET dataset (Phillips et al., 2000b)) was used instead, as a case study. These images were manually annotated to provide a ground truth for the algorithm to be developed.

The proposed algorithms works as follows. Given an input image I , the directional gradients ($\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial y}$) of I are computed, where $\frac{\partial I}{\partial y}$ (denoted as \mathcal{I})

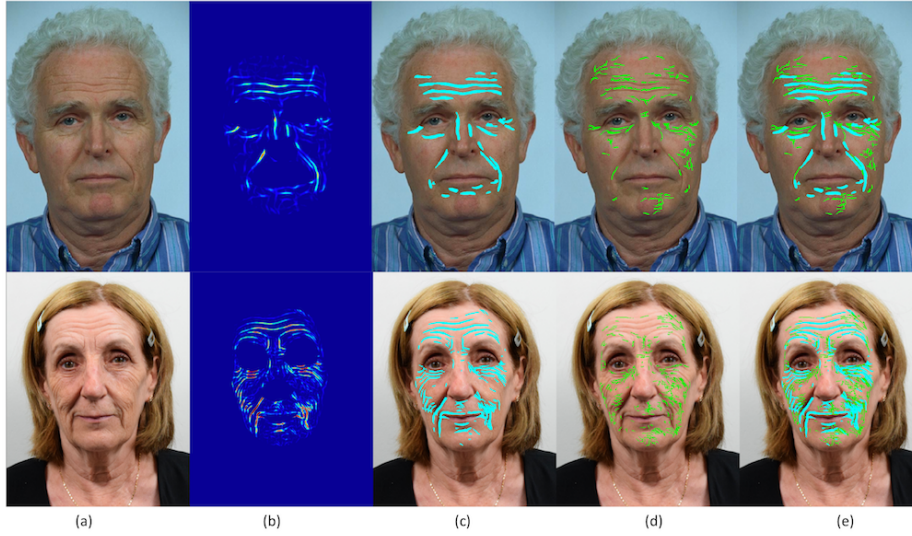


FIGURE 3.6: Visual illustration of automated extraction of coarse wrinkles and fine wrinkles regions: (a) Original image, (b) Probabilistic wrinkles map, (c) Coarse wrinkles, (d) fine wrinkles, and (e) Combined wrinkles regions.

emphasizes a horizontal variation, which was proposed by Ng et al., 2014; Ng et al., 2018 to detect horizontal wrinkles, and $\frac{\partial I}{\partial x}$ (denoted as \mathcal{V}) emphasizes a vertical variation, which the proposed algorithm uses to detect vertical wrinkles regions. Both gradients are used as inputs to construct a Hessian filter \mathcal{H} (Ng et al., 2018) at location (x, y) as:

$$\mathcal{H}_\sigma(x, y) = \begin{bmatrix} \mathcal{H}_{a,\sigma}(x, y) & \mathcal{H}_{b,\sigma}(x, y) \\ \mathcal{H}_{b,\sigma}(x, y) & \mathcal{H}_{c,\sigma}(x, y) \end{bmatrix} \quad (3.1)$$

where σ is the filter scale, and $H_{a,\sigma}$, $H_{b,\sigma}$ and $H_{c,\sigma}$ are the second derivatives of \mathcal{I} along the horizontal, diagonal, and vertical directions, respectively. In this work, the value of σ was changed depending on whether the model had to focus on coarse (σ_c) or fine (σ_f) wrinkles.

It was found empirically that for images from the FERET dataset it was best to use $\sigma_f = 2$ and $\sigma_c = 4$, while for images from the Social Habits dataset it was best to use $\sigma_f = 3$ and $\sigma_c = 6$. This is because the images in the two datasets have two different resolutions (512×768 for FERET; 1000×1300 for

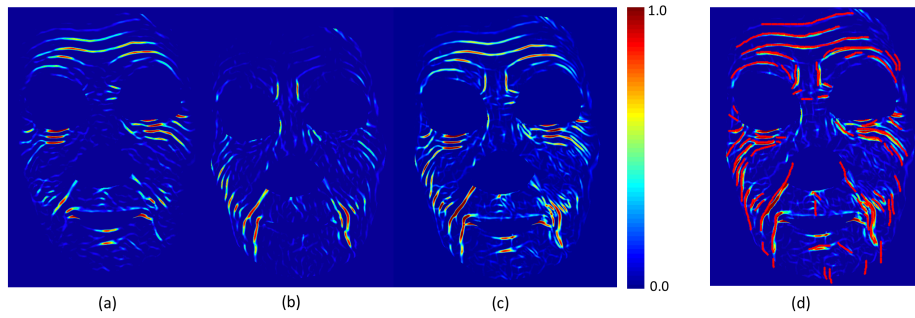


FIGURE 3.7: Probabilistic map on: a) Horizontal wrinkles; b) Vertical wrinkles; c) Probabilistic map by our proposed method; and d) Manual annotation of Wrinkle lines (RED) overlay on our probabilistic map.

Social Habits) and, consequently, the wrinkles have different resolutions and levels of detail.

The filter σ_f detects fine wrinkles but also coarse ones. To separate the two types, a post-processing step was performed. Let R_f be the regions detected as fine wrinkles and R_c be the regions detected as coarse wrinkles. Coarse wrinkles are removed from R_f by imposing $R'_f = \{\forall x \in R_f \mid x \notin R_c\}$. Figure 3.7 shows an example of the output of the proposed model, which is presented as a probabilistic wrinkle map in which high values indicate deep wrinkles and low values indicate shallow ones.

3.4.2 Results and Discussion

Figure 3.6 illustrates the step-by-step results of the proposed method on coarse and fine wrinkle annotation. Compared to the manually annotated ground truths, the proposed method achieved a *Jaccard Similarity Index (JSI)* of 80%. This work was published in (“Automated Facial Wrinkles Annotator.”), [P02] on the publication list.

3.5 Experiment II: Automated Acne Detection

There are different acne types: blackhead, whitehead, cyst, and pustule (Ramli et al., 2012). Dermatologists typically assess acne by visually assessing the area and making an estimate based on observation. Such a method, however, is inherently subjective. Some objective acne assessment methods (Shen et al., 2018a; Nguyen, Thai, and Le, 2021) have been developed as a result of recent advances in computerised algorithms, as opposed to dermatologists' subjective approach. However, such methods are less accurate than dermatologists. As a result, a novel method for detecting acne based on the Hessian filter algorithm is presented here.

The majority of acne detection methods are based on the color descriptor approach or on some kind of clustering. For example, (Malik et al., 2014) used a modified k -means clustering and SVM to classify different types of acne, reaching a sensitivity of 90% and a specificity of 97.2%, using only 50 facial images.

The most recent related work was proposed by (Amini et al., 2018). The authors implemented an automated facial acne assessment algorithm from smartphone images. Their algorithm, which first detected the face from the image using facial landmarks, and then detected the regions of interest (i.e. regions with acne) with an accuracy of 92%. Then, the algorithm classified the detected regions into different categories of acne by converting the input RGB image to a CIE $L^*a^*b^*$ color code and applying a Gaussian filter to the resulting image. Lastly, Otsu thresholding was applied to achieve an accuracy of 98%. The method used 60 digital images for training and 10 real face images for validation.

Some authors (notably, (Shen et al., 2018a)) also tried using Convolutional Neural Networks to automatically diagnose facial acne vulgaris and achieved

an accuracy of 81%. However, using such methods requires a lot of data.

3.5.1 Proposed Method

A new method for acne detection is proposed here, using skin patches from the dataset developed in [section 3.2](#), which includes images of resolution 100×100 pixels of various acne types. The model is then validated using 10 high resolution images from the same dataset. The proposed algorithm was inspired by (Frangi, 2001), who developed a mathematical filter originally designed to work on images of blood vessels. The Frangi filter has been successfully applied to various applications such as wrinkle detection (Ng et al., 2014) and retinal vessel detection (Sofka and Stewart, 2006). [Figure 3.8](#) demonstrates an example of an acne image (a), comparison between the method from (Ng et al., 2014) (b) and the proposed filter (c). In particular, the Frangi filter is suitable for wrinkle detection because both blood vessels (for which the filter was designed) and wrinkles appear in images as lines. However, acne presents itself as a blob-like structure. To adapt the Frangi filter to acne detection, then, some pre and post-processing steps were implemented. First, given a coloured face image $I(x,y)$, it is converted to greyscale. The directional gradients $\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial y}$ are then computed and combined into a 2×2 Hessian matrix. Then, the eigenvalues of the Hessian matrix are computed to find the direction of variability. Finally, two similarity measures are computed, R_β and S , where S is a measure of the sensitivity to blob-like structures and R_β (calculated by multiplying the eigenvalues, as shown in [3.1](#)) is a measure of the area of detected regions. For more detailed steps, refer to (Frangi, 2001). Lastly, an Otsu threshold is applied to the outputs.

$$R_\beta = \left| \lambda_1 \right| * \left| \lambda_2 \right| \quad (3.2)$$

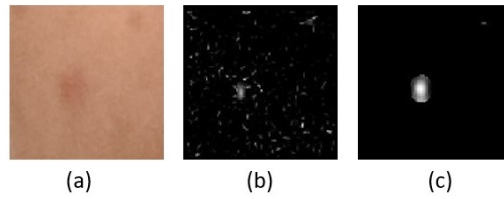


FIGURE 3.8: Comparison of the two filters, where (a) is the original image, (b) is the filter proposed in (Ng et al., 2014) and (c) is the proposed filter

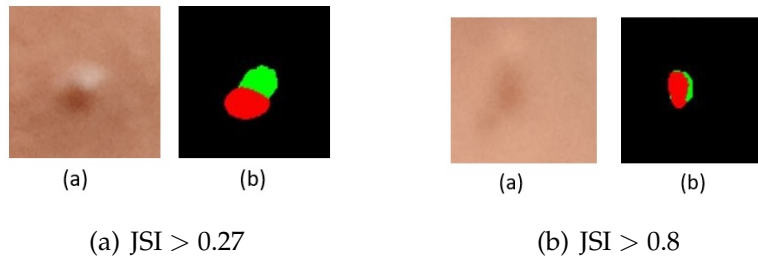


FIGURE 3.9: Example of the results: where (a) is the original image and (b) is the detection results (red correct detection and green is the missed detection)

$$S = \sqrt{\lambda_1^2 + \lambda_2^2} \quad (3.3)$$

3.5.2 Results and Discussion

The Jaccard Similarity Index (JSI), reported in equation (4.6), was used to measure the similarity between the outputs of the proposed method and the ground truths, which were obtained using MATLAB under a controlled environment.

$$J(x, y) = \frac{|x \cap y|}{|x \cup y|} \quad (3.4)$$

When the JSI is > 0.1 , the method employed indicates a hit-or-miss situation, as illustrated in Figure 3.9 (a). When JSI is > 0.2 , it shows segmentation, and the percentage is 64.5%, which is greater than the standard (threshold) value at JSI > 0.4 (45.5%).

In this experiment, the reliability of the manual annotations was measured amongst three coders. Each coder was asked to manually label a sample of 25

images from the main dataset. Considering a JSI threshold equal to or greater than to 40% (as used in (Ng et al., 2014), for coders A and B, the JSI of 99% of the labels was above the threshold, with a standard deviation (STD) of 14.27; for coders A and C, the JSI of 92% of the labels was above the threshold, with an STD of 16.77; and for coders B and C, the JSI of 88% of the labels was above the threshold, with an STD of 18.61. The average reliability of manual annotation between coders A, B and C was therefore 93%. However, the annotation is a challenging task and has inconsistency shown in Figure 3.10.

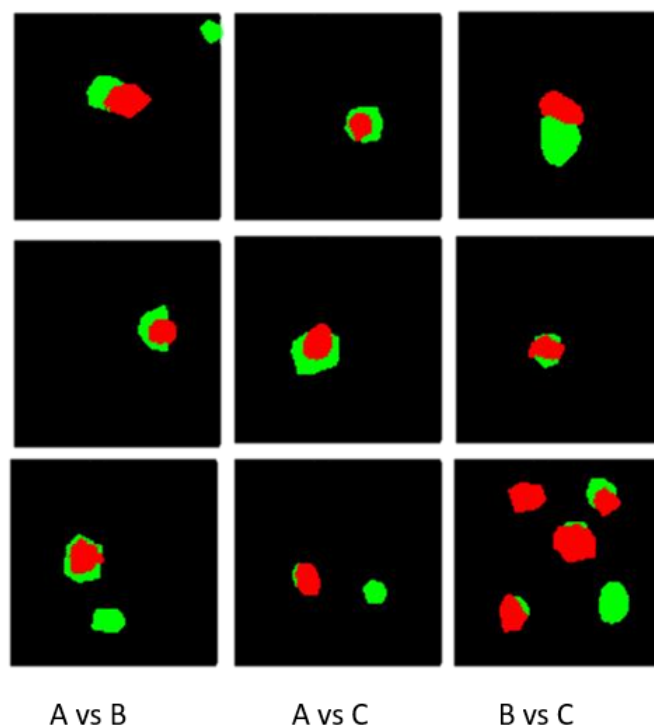


FIGURE 3.10: annotation among three coders.

3.6 Summary

In this chapter, two new high-resolution datasets and two computer methods for objective quantification of skin assessment were proposed.

In Experiment I, a new fully automated wrinkle annotator for coarse wrinkle and fine wrinkle labeling was proposed. With the rapid growth of the use

of deep learning in computer vision, automated annotators for data labeling play an important role for ground truth preparation. Since it is impractical to manually annotate fine wrinkles, the method proposed here will likely benefit data driven approaches greatly. In addition, it could be used to assess skin quality in terms of fine lines.

In Experiment II, an acne detection algorithm was proposed. The method was tested on 107 acne skin patches. The method is promising in detecting acne as it is able to detect blob-like structure. However, the small resolution and limited size of the available dataset were found to limit the potential of the algorithm.

While the algorithms presented in this chapter require manual adjustment of filter scales, [chapter 4](#) will explore machine intelligence methods, which will be able to skip the heuristic definition of filter scales.

4 Facial Skin Classification using Machine Learning and Deep Learning Techniques

This chapter is divided in two sections. The first will give an overview computer vision concepts and techniques based on machine learning, including classification approaches and their performance measures. The second will describe the first comparative study that aims to provide a pivot work on the basis of facial skin classification.

4.1 Introduction

The field of deep learning (i.e. algorithms that learn abstract representations from data without needing heuristic rules) has had, in recent years, an explosive growth, surpassing Conventional Machine Learning (CML) methods in many computer vision tasks like object classification, facial recognition, and face tracking (Wang and Sng, 2015). However, the most popular deep learning methods (e.g. (Krizhevsky, Sutskever, and Hinton, 2012)) use data that are unrelated to facial skin attributes, and many algorithms that analyse facial skin attributes are intended for clinical use. For example, Andre et al. (Esteva et al., 2017) developed a deep learning approach for the detection

of skin cancer from images, and obtained accuracy comparable to that of an ensemble of 21 dermatologists.

Indeed, there is little research on deep learning applied to facial skin analysis. Instead, most authors use CML methods for such tasks. The most common types of CML methods used for facial skin analysis are:

- Support Vector Machines (SVMs). They have been used extensively over the last decade (Schmidhuber, 2015), for example to categorise skin texture and colour for a melanoma detection task (Yuan et al., 2006; Khan et al., 2012).
- Feature representation. Feature representation, also known as feature learning, is a class of methods used for object classification and detection from raw data which determines features using automatic heuristic rules. The three main types of feature learning algorithms are:
 - Local Binary Pattern (LBP). LBP is an effective texture operator that divides input images into patches of 8 pixels and assigns a label to the pixel at the centre of each patch based on the pixels surrounding it (Tian et al., 2013). One of the main strengths of this method is its simplicity and speed of computation, which allows it to perform classification in real time (Ahonen, Hadid, and Pietikainen, 2006).
 - Histogram of Oriented Gradient (HOG). This feature detection method works by counting the instances of occurrence of gradient orientation in an image (Dalal and Triggs, 2005). The gradients G_x and G_y are found by using a 3×3 Sobel mask, and the orientation L is found as $L = \arctan\left(\frac{G_x}{G_y}\right)$, whereas the total magnitude of the gradient of an image is computed as $M = \sqrt{(G_x)^2 + (G_y)^2}$.

Though the above methods have been used extensively in the past, deep learning algorithms, and Convolutional Neural Networks (CNNs) in particular, greatly outperform CML methods in most classification tasks (Schmidhuber, 2015), and it would be of interest to apply such algorithms to facial skin analysis. The following section reviews some fundamental theoretical concepts of deep learning.

4.2 Brief Overview of Deep Learning

Deep learning is a subcategory of machine learning (Alpaydin, 2014), which is a class of algorithms that can learn mappings from inputs to outputs without relying on hand-crafted features or heuristic rules. Machine learning is divided into two main branches: supervised and unsupervised. Unsupervised learning algorithms learn to categorise data without it being labelled, simply by finding patterns in the data. The data points are then grouped based on the pattern class (or ‘cluster’) to which the algorithm thinks it belongs. On the other hand, supervised learning algorithms are shown examples of inputs and the corresponding desired outputs, and are asked to learn to output the correct value for a given input. This section will focus on supervised learning, and in particular on the task of classification, which is required for skin attributes detection and analysis.

The term ‘deep learning’ refers to artificial neural networks (ANNs) with many hidden layers, which makes their architecture ‘deep’; an example of such a model is shown in Figure 4.1. Each node of the network has connections to nodes of other layers, and an optimisation algorithm is used to update the weights of these connections by using a loss function that measures how well the network is currently predicting the targets. Convolutional neural networks (shown in Figure 4.2) are a particular type of deep learning algorithms that perform well on images. In a convolutional layer, a filter (usually

of size 3×3 or 5×5) is convoluted with an input image (i.e., it is applied to all 3×3 or 5×5 patches of the image). The values of the filter are randomised at first, and adjusted during the learning process. By using several different filters, the network learns to detect different types of features in the image. For example, a 3×3 filter may learn to recognise horizontal edges in the image, and a different filter may learn to recognise vertical edges. Successive convolutional layers then combine these low-level features into progressively higher level ones which have more semantic meaning; a deep hidden layer (i.e. a layer past the input one) may learn to detect an ear in an image, or eyes, or legs; deep layers still would then learn even more semantically complex features, such as a face, a house, a car. Based on these high-level, semantically rich features, the final layer is then able to predict the class of the object(s) in the image. For more information on convolutional neural networks, and on deep learning in general, refer to (Goodfellow, Bengio, and Courville, 2016).

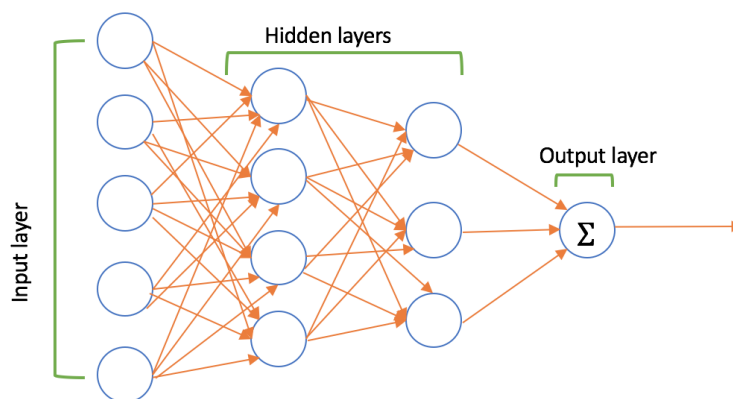


FIGURE 4.1: Feedforward Neural Network.

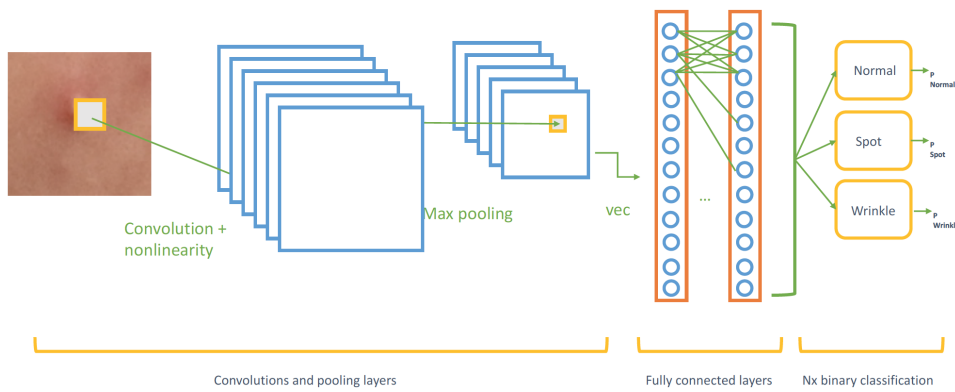


FIGURE 4.2: Convolutional Neural Network.

4.3 Classification evaluation

The basis for computing the performance of a classification algorithm is the confusion matrix:

TABLE 4.1: Confusion matrix

		Prediction outcome			total
		p	n		
actual value	p'	True Positive	False Negative	P'	
	n'	False Positive	True Negative	N'	
total		P	N		

where True Positives (TP) are inputs that the model correctly predicted as belonging to the 'positive' class¹; False Positives (FP), also known as Type I errors, are instances of the negative class that the model predicted as positive; False Negatives (FN), also known as Type II errors, are instances of the positive class that the model predicted as negative; and True Negatives (TN) are instances that the model correctly predicted as negative. From the elements of the confusion matrix different metrics can be derived in Table 4.2.

¹This assumes that the task at hand was binary classification, where samples either belong to a given class or not. For example, it could be an algorithm for determining if pictures of faces of people have acne (positive class, or 1) or not (negative class, or 0).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.2)$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} \quad (4.3)$$

$$\text{F1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (4.4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4.5)$$

$$\text{Jaccard Similarity Index (JSI)} = \frac{TP}{TP + FP + FN} \quad (4.6)$$

TABLE 4.2: Performance measures

4.4 Facial skin classification

Benign lesion experiences colour transformation, it may misdiagnosed as other cancerous conditions such as melanoma and basal cell carcinoma. Resulting in reducing the unnecessary biopsies and extraction of the lesions. Computer-aided tools may help in reducing the time and effort in analysing such lesions. One of the challenging tasks for dermatologists is to examine benign keratoses types. This section will focus on supervised learning, classification in particular that is required for skin attributes detection and analysis.

4.5 Proposed Method

This section compares two methods (SVMs and CNNs) on the classification of three classes of facial skin: normal, spot, wrinkle. These classes have major textural differences between them, making them an interesting target for the comparison. The goal of the experiment, ultimately, was to evaluate if deep learning models such as CNNs could outperform traditional machine learning methods such as SVMs on such a task.

For the SVM model, first features were extracted from images using the LBP (Guo, Zhang, and Zhang, 2010) and HOG (Yap et al., 2009) algorithms, described earlier in this chapter. Furthermore, different colour descriptors (i.e. noormalised RGB, HSV, and L^*u^*v) were used. To train the SVM, the Sequential Minimal Optimisation (SMO) algorithm was used (Hearst et al., 1998).

The CNN of choice was the state-of-the-art GoogLeNet, implemented via the Caffe framework (Jia et al., 2014). Different optimisation algorithms were tested: Stochastic Gradient Descent (SGD), Nesterovs Accelerated Gradient (NAG) and Adaptive Gradient (AdaGrad). SGD is one of the most commonly used optimisation algorithms for deep learning models (Singh et al., 2015),

but AdaGrad and NAG have also shown promising results in previous research (Goodfellow, Bengio, and Courville, 2016). Each optimizer was tested with its default settings for 60 epochs and 0.001 learning rate.

The datasets available for facial skin analysis have some limitations. For example, the DermNet dataset, which consists of 23,000 images of various skin diseases and conditions, contains images that were collected in poorly controlled environments, and their ground truths are not always reliable. Furthermore, most of the images focus on skin conditions that were not relevant for this experiment. Therefore, skin patches were cropped from the Social Habits dataset described in section 3.2. These were of size 100×100 pixels and were labelled with three categories: normal skin, skin with spots, and skin with wrinkles, as illustrated in Figure 4.3. The total number of skin patches were 325. 108, 108 and 109 for normal, wrinkle, and spot skin patches respectively. The data were split into 70% for training and 30% for testing. Given the limited amount of data available, a 10-fold cross-validation was used, where each fold had equal distribution of classes.

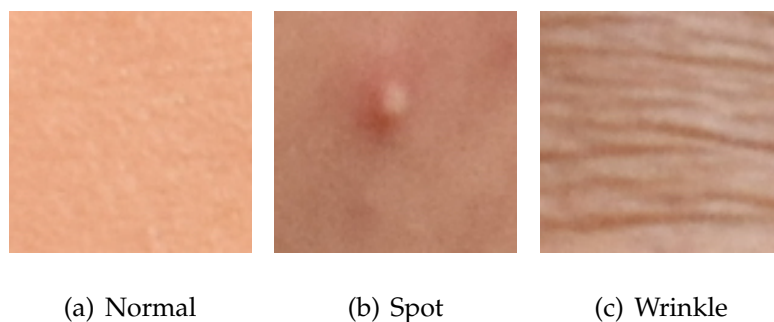


FIGURE 4.3: Sample skin patch from each three classes.

4.6 Results and Discussion

This section reports the results of the SVM model trained with different types of features as inputs (Table 4.3), as well as the results of the CNN trained with different optimisers (Table 4.4). The reported metrics are *Sensitivity*,

False Negative Rate (FNR), F-Measure. Recall, Precision, Matthews Correlation Coefficient (MCC), and Accuracy. The results reported here clearly show how, with careful tuning of the hyperparameters of a CNN, such a deep model easily surpasses the performance of traditional methods like SVMs even with scarce training data.

TABLE 4.3: SVM results.

<i>Method</i>	<i>Sensitivity</i>	<i>F-Measure</i>	<i>Recall</i>	<i>Precision</i>	<i>MCC</i>	<i>Accuracy</i>
LBP	0.742	0.741	0.741	0.740	0.597	0.815
LBP and HOG	0.736	0.738	0.742	0.742	0.591	0.811
LBP, HOG and Colour Descriptor	0.733	0.735	0.740	0.740	0.586	0.808

TABLE 4.4: GoogLeNet results.

<i>Solver</i>	<i>Epochs</i>	<i>Learning Rate</i>	<i>Sensitivity</i>	<i>F-Measure</i>	<i>Recall</i>	<i>Precision</i>	<i>MCC</i>	<i>Accuracy</i>
SGD	30	0.01	0.666	0.661	0.666	0.666	0.472	0.754
	60	0.01	0.833	0.835	0.835	0.835	0.745	0.884
	60	0.001	0.677	0.670	0.671	0.671	0.487	0.761
NAG	30	0.01	0.646	0.639	0.645	0.645	0.439	0.738
	60	0.01	0.854	0.852	0.856	0.856	0.779	0.899
	60	0.01	0.729	0.727	0.731	0.732	0.579	0.856
AdaGrad	30	0.01	0.521	0.425	0.375	0.375	0.192	0.624
	60	0.01	0.646	0.650	0.667	0.667	0.449	0.739
	60	0.001	0.708	0.703	0.707	0.707	0.545	0.790

4.7 Summary

This chapter introduced various concepts from machine learning and deep learning and compared two widely used algorithms from these fields (SVMs and CNNs) on a skin classification task. The results of the experiments indicated that the CNN model outperformed the SVM model. This is a strong indication that the CNN model can perform significantly better with additional training and data augmentation techniques. This is a step forward because it directs research toward exploring and improving the CNN method and toward developing tools for facial skin analysis in terms of skin classification.

5 Human perception: Face Age Estimation and Skin Quality

This chapter aims to understand the human perception of age and skin quality. It describes the experimental protocol and presents statistical analysis to compare the performance of dermatologists versus ordinary participants in face age estimation and skin quality rating.

5.1 Introduction

According to evolutionary psychology, facial features influence judgement when assessing facial skin. Machine learning approaches have also been proposed to handle the same task. These methods, however, focus on skin patches rather than the entire face. In this chapter, a novel approach is proposed to understand and compare the difference between human perception of facial skin assessment and the machine learning approach for facial skin assessment. The study is divided into two parts. The first experiment uses an eye-tracking system to learn about how humans perceive facial skin. This entailed asking participants to look at images of females and assess the facial skin quality using facial features (e.g., eyes, mouth), as well as assessing the same faces with cropped out facial features. After being informed of the age of the person in the image, the participants used a Likert scale to judge skin quality by assigning a score from 1 to 5 to each image. This

was chosen as the scoring method so that the results could be statistically analysed. This method is associated with several hypotheses, including that visibility of facial features improves skin quality judgement (and thus age estimation). The second experiment focuses on the machine learning approach, with additional analysis performed to comprehend and compare the results of the first experiment's human perception to those of the machine learning approach. This overall evaluates how the faces are judged, perceived, and judged by both approaches. This is accomplished by employing a state-of-the-machine learning approach, visualising the areas that contributed to the final estimation, and then comparing them to the human estimation from the first experiment.

5.2 Hypotheses

- H1: Facial features be included when it comes to analysis and evaluation of skin quality.
- H2: Skin quality assessment depend on facial features, age and gender.
- H3: There a correlation between the assessment of skin quality and age.

Experimental Protocol

Subjects¹ For all experiments, the Social Habits dataset [section 3.2](#) was used. However, the order of the pictures was randomised each time they were shown to participants, to avoid the participants remembering or associating the faces of people with previous experiments. The age range was (21-70), divided into 5 age groups 21-30, 31-40, 41-50, 51-60, and 61-70 consisting of 10 subjects per group, this is because there were not enough data per age group. The images were shown to participants on a screen display with resolution of

¹The word 'subject' will be used here to refer to people whose images form the dataset; the word 'participant' will be used to refer to people who perform the estimation tasks.

1783 × 2644 pixels. A random stimuli image example is shown in the mask used to generated the face image and featureless face that were used in the experiments Figure 5.1. The mask shown in in Figure 5.1 (a) was introduced by (Ng et al., 2018) to understand how humans estimate skin quality and age with and without facial features. Figure 5.1 (b) shows a random stimuli face image. Figure 5.1 (c) shows the mask over the image.

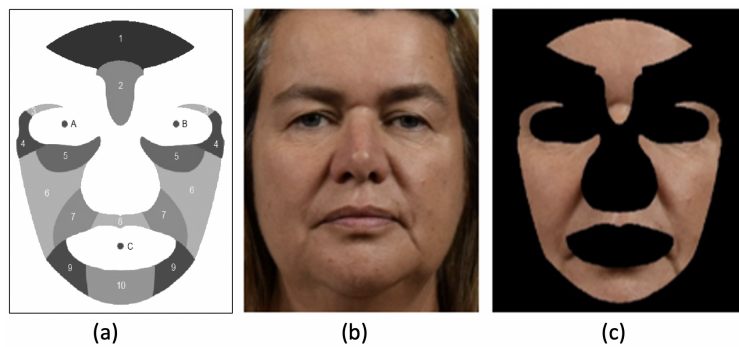


FIGURE 5.1: Random stimuli example. (a) ROI mask overview, (b) Frontal face image, (c): is featureless face after applying the mask.

5.3 Eye-tracking Experiments

Eighteen non-expert participants and ten dermatologists were recruited for this study. Upon arrival to the place of the experiments, each participant sat comfortably in front of the screen, where an eye-tracking software was operative (see Figure 5.2). The experiment was conducted in three phases: 1) participants were shown images of females where the whole face was visible, and were asked to estimate the age of the person in the image; 2) same as 1), but with images in which the facial features (nose, mouth, eyes) were masked (these images are also referred to here as ‘skin only’); and 3) participants were shown full images of female faces, told the age of the person in each image, and were asked to judge the quality of the facial skin (from 1 = very poor quality, to 5 = very good quality). Participants were informed that their gaze was being tracked and that an individual calibration was performed.

The SMI RED 250 (Mele and Federici, 2012) static eye-tracking remote device (with a sampling frequency of 120 Hz) was used for the first experiment; which is analysed in chapter 6, and the Tobii Pro Fusion (Pro, 2019) device (sampling frequency of 250 Hz) for the second experiment which is analysed in this chapter. . The screen display resolution was 1783×2644 pixels. The eye tracking raw data are available in Appendix C.

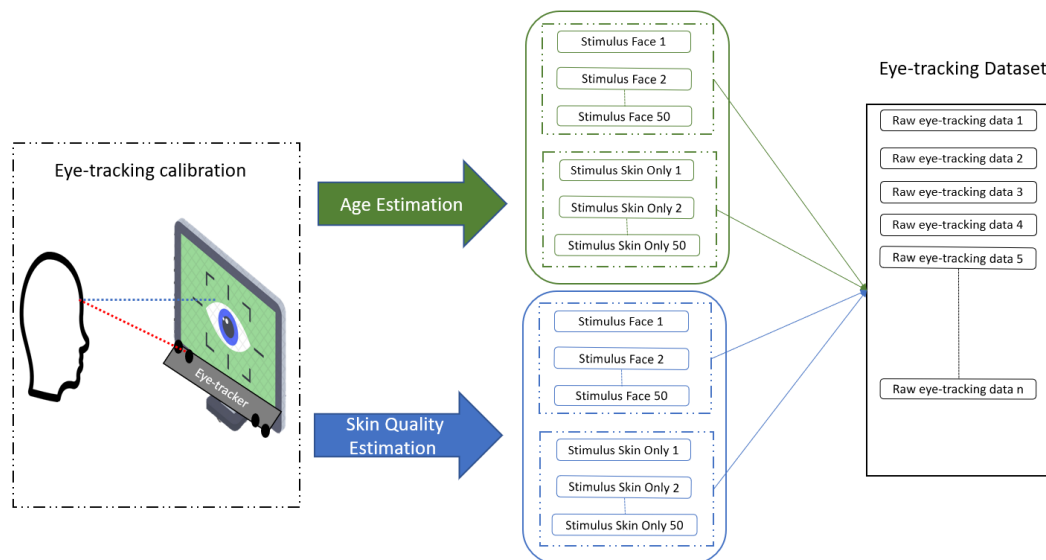


FIGURE 5.2: The experimental setting for the Experiment II.

5.4 Statistical Analysis

Numerous statistical tests which were used: T-test, Chi-square and mixed effect models. The Chi-square test was used wherever the response variable (Y) was discrete and the input variable (x) was also discrete, and it estimates how probable it is that any notable difference between the sets resulted from chance. Ordinal logistic regression was used wherever response variable was discrete and ordinal and input variable was numeric. An analysis of variance (ANOVA) test was also done by considering true age as the response

variable (Y) and skin quality as the input variable. This is a method to determine whether or not the means of two or more groups statistically differ from one another.

In order to understand the relationship between factors, such as gender and age, the data were examined quantitatively by the linear mixed-effect model in Equation 5.1.

$$Y_{npq} = \beta_1 X_{npq} + \beta_2 X_{npq} + \beta_3 X_{npq} + \beta_4 X_{npq} + IZ_{npq} + \varepsilon_{npq} \quad (5.1)$$

where β in the Equation 5.1 represents the fixed effects coefficients through X_{npq} that are the fixed effect predictors for observation p in group q , whereas I represents the random effect of the model through Z for the n^{th} observation variables p q , which are assumed to be multivariate normally distributed. The error for the case p in group q is indicated as ε .

For this experiment, two mixed effect models (from Equation 5.1) were applied, where the random effect for both models was the participant. The first model is illustrated in the fixed effect model results in the following sections.

In addition, the agreement among participants was measured using two tests: *Kendall's correlation coefficient* and *Fleiss' Kappa*. Table 5.1 illustrates the values for Kendall's correlation coefficient. For Fleiss' Kappa, values are colour-coded: green for fair agreement; yellow for slight agreement; and pink for poor agreement.

TABLE 5.1: This is the legend for Kendall’s correlation coefficient

Poor agreement =	Less than 0.20
Fair agreement =	0.21 to 0.40
Moderate agreement =	0.41 to 0.60
Good agreement =	0.61 to 0.80
Very good agreement =	0.81 to 1.00

5.4.1 Age estimation

Overall, the agreement of both experts and non-experts was found to be good and very good agreement based on the scale on Table 5.1. The Fleiss’ Kappa agreement (Table 5.2 has good agreement overall but not among all age groups, the agreement among all age groups for experts and non-expert. However, on Table 5.3, Kendall’s Correlation showed very good agreement in experts and good agreement in non-experts.

TABLE 5.2: Fleiss’ Kappa Statistics - Absolute Agreement Summary for Age Estimation

Comparison Parameters	Overall	Expert Overall	Expert without mask	Expert With Mask	Default Overall	Default without mask	Default With Mask
Between respondents	0.2728	0.306125	0.394783	0.237103	0.254968	0.298098	0.233939
Between respondents and true age	0.3241	0.362234	0.443077	0.281391	0.305058	0.356985	0.253132
21-30		0.13962	0.204735	0.139579		0.0351256	0.046442
31-40		0.076541	0.269337	-0.0396734		0.131914	0.042044
41-50		0.184485	0.161669	0.181877		0.09519	0.164097
51-60		0.0011167	-0.0457936	-0.008874		0.0437836	0.078161
61-70		0.0221663	0.0411732	-0.0148656			

TABLE 5.3: Kendall’s Correlation Coefficient for age estimation.

Comparison Parameters	Overall	Expert Overall	Expert without mask	Expert With Mask	Default Overall	Default without mask	Default With Mask
Between respondents	0.79304	0.833652	0.891409	0.836839	0.780427	0.82924	0.779883
Between respondents and true age	0.72843	0.748174	0.801833	0.694516	0.718555	0.771117	0.665994
21-30		0.331818	0.318574	0.549334		0.143676	0.378841
31-40		0.359652	0.588767	0.403875		0.34651	0.29118
41-50		0.721875	0.701132	0.808704		0.46261	0.639698
51-60		0.205129	0.278755	0.326583		0.383998	0.460933
61-70		0.153652	0.299142	0.214474		0.0728618	0.165333

From the statistical analysis, the following conclusions can be drawn:

- The accuracy of age prediction dependent on the facial features. As the p-value of the Chi-Square test ($P = 0.00$) was less than 0.05, the null hypothesis is rejected and it can be concluded that there is an association between age prediction accuracy and Mask/ No Mask (features). The proportion of match is higher if there is no mask.
- Hypothesis 3: There is a difference in age prediction by gender. As the p-value of the Chi-Square test ($P = 1.00$) was greater than 0.05, the null hypothesis is accepted and it can be concluded that there is no association between the gender of participants with the accuracy of age prediction.
- Hypothesis 4: The accuracy of age prediction depend upon expertise level. As the p-value of the Chi-Square test ($P = 0.03$) was less than 0.05, the null hypothesis is rejected and it can be concluded that there is an association between accuracy of age predicted with expertise level. The percentage of accurate match is significantly higher for experts in comparison to non-experts.
- Hypothesis 5: There is a correlation between participants' age and their accuracy in predicting age. As the p-value of the Chi-Square test ($P = 0.738$) is greater than 0.05, the null hypothesis is accepted and it can be concluded that there is no association between the age of the participants and the accuracy of age predicted.
- There is a relationship between the experience level of the experts and their accuracy in predicting age and is there an interaction between the experience level of experts and their accuracy in predicting age with/without faical feaures. As the p-value of the binary logistic regression test ($P = 0.218$) was greater than 0.05, the null hypothesis is accepted and it can be concluded that there is no relationship between experience of experts with accuracy of predicted age.

- Hypothesis 7: There is an association between accuracy in predicting the age of the people with a smoking habit. This is broken down into two:
 - Accuracy versus smoking habits without mask: As the p-value of the Chi-Square test ($P = 0.005$) was less than 0.05, the null hypothesis is rejected and it can be concluded that there is an association between accuracy of age prediction and whether the person whose age is being estimated smokes or not.
 - Accuracy versus smoking habits with mask: As the p-value of the Chi-Square test ($P = 0.34$) was greater than 0.05, the null hypothesis can be accepted and concludes that, if a mask is used, there is no association between accuracy in predicting the age of a person with a smoking habit.
- Hypothesis 8: There is an association between accuracy in predicting the age of the participant with a drinking habit. This is broken down into two:
 - Accuracy of predicting age of participant with a drinking habit (without mask): As the p-value of the Chi-Square test ($P = 0.231$) was greater than 0.05, the null hypothesis is accepted and it can be concluded that, without a facial mask, there is no association between accuracy in predicting the age of people with drinking habits.
 - Accuracy of predicting age with drinking habits (with mask): As the p-value of the Chi-Square test ($P = 0.227$) was greater than 0.05, the null hypothesis can be accepted and it can be concluded that, with a facial mask, there is no association between accuracy in predicting the age of people with drinking habits.

- Hypothesis 9: There is a difference in the error in prediction within age group intervals with/without facial features. As the p-value of the paired T-test ($P = 0.00$) was less than 0.05, the null hypothesis can be rejected and it can be concluded that there is a difference in the error in prediction within age group intervals with mask or without mask. In particular, when a face mask is applied faces are generally predicted to be older than when no face mask is used.

Mixed Model Effect summary:

Model I: Model I is built by encoding the predicted age interval to its mid point. For example, if the age predicted is 21-30, then the recorded estimated age is 25; this was done to convert the categorical data to continuous data to run the mixed effect model. The Y or response variable considered in this model was the predicted age (encoded), whereas the random variables taken were "Participants" and "Image" (as they don't have any fixed levels), while other factors like gender, mask/ no mask, group, and participant age were taken as fixed effects, as they have fixed levels. Conclusion of Random Effect: Both random effects were found to be significant for predicting the age, with the Image variable contributing to 72.56% of the variance in the predicted age.

Tests of Fixed Effects

Conclusion of Fixed Effect: In terms of fixed effect, the terms which were found to be significant are the age group of the participants, and whether a mask was applied to the images.

Model II:

Model II is built by taking the accuracy (match or no match) of the prediction as the response variable. The accuracy is converted in 1 (Match) and 0

(No Match), to convert the categorical data to numeric data to run mixed effect model. The random variables and fixed effects were the same used in model I. Conclusion of Random Effect: Both random effects are significant for predicting age.

Tests of Fixed Effects

Conclusion of Fixed Effect: Only the presence/absence of a face mask was found to be significant for predicting age.

5.4.2 Skin Quality

During the study, both experts and non-experts were unable to agree on similar skin quality score. Thus, Fleiss’ Kappa agreement has poor agreement among all age groups for experts and non-expert. However, Kendall’s Correlation showed moderate agreement among experts when the mask was applied.

TABLE 5.4: Kendall’s Correlation Coefficient for Skin Quality

Comparison Parameters	Overall	Expert Overall	Expert without mask	Expert With Mask	Default Overall	Default without mask	Default With Mask
Between respondents	0.185	0.230	0.365	0.443	0.175	0.296	0.298
21-30		0.398	0.396	0.578	0.278	0.364	0.298
31-40		0.141	0.267	0.404	0.130	0.321	0.173
41-50		0.093	0.322	0.503	0.110	0.267	0.316
51-60		0.305	0.423	0.397	0.200	0.362	0.349
61-70		0.185	0.324	0.373	0.105	0.201	0.221

TABLE 5.5: Fleiss’ Kappa Statistics - Absolute Agreement Summary for Skin Quality

Comparison Parameters	Overall	Expert Overall	Expert without mask	Expert With Mask	Default Overall	Default without mask	Default With Mask
Between respondents	0.04	0.032	0.051	0.067	0.033	0.061	0.062
21-30		0.060	0.031	0.082	0.051	0.066	0.052
31-40		0.004	0.061	0.024	0.020	0.040	0.051
41-50		0.005	0.014	0.075	0.023	0.061	0.074
51-60		0.045	0.062	0.081	0.037	0.093	0.068
61-70		0.012	0.021	0.041	0.015	0.022	0.025

1. Hypothesis 1: Is there is a difference in the skin quality rating with full features (no mask) versus with mask for each respondent? As the

p-value of the paired T-test ($P = 0.00$) was less than 0.05, the null hypothesis is rejected and it can be concluded that there is difference in the skin quality rating with mask or without mask.

2. Is there a difference in skin quality rating by gender? As the p-value of the Chi-Square test ($P = 0.642$) was greater than 0.05, the null hypothesis can be accepted and it can be concluded that there is no association between the gender of the participants and the skin quality rating given to subjects.
3. Does the skin quality depend upon expertise level? As the p-value of the Chi-Square test ($P = 0.067$) was greater than 0.05, the null hypothesis can be accepted and it can be concluded that there is no association between expertise level and skin quality rating given to subjects.
4. Is there correlation between respondents age group and skin quality rating? As the p-value of the Chi-Square test ($P = 0.00$) was less than 0.05, the null hypothesis is rejected and it can be concluded that there is an association between the age group of the participants and the skin quality rating.
5. Is there an association between skin quality of the participant with his smoking habit? This is broken down into:
 - Skin Quality Rating versus smoking habit without mask: As the p-value of the Chi-Square test ($P = 0.569$) was greater than 0.05, the null hypothesis can be accepted and it can be concluded that, when no mask is used, there is no association between skin quality rating and the smoking habits of the subjects.
 - Skin Quality Rating versus smoking habits with mask: As the p-value of the Chi-Square test ($P=0.01$) is less than 0.05, the null hypothesis is rejected and it can be concluded that, when a mask is

used, there is an association between skin quality rating and the smoking habits of the subjects.

6. There an association between skin quality rating and drinking habit.

This is broken down into:

- Skin Quality Rating versus drinking habit (without mask): As the p-value of the Chi-Square test ($P=0.00$) is less than 0.05, the null hypothesis is rejected and it can be concluded that, without a mask, there is an association between skin quality rating and drinking habits of subjects.
- Skin Quality Rating versus drinking habit (with mask): As the p-value of the Chi-Square test ($P=0.00$) is less than 0.05, the null hypothesis is rejected and it can be concluded that, with a mask, there is an association between skin quality rating and drinking habits of subjects.

7. There is a relationship between the true age of subjects and their skin quality given by participants. (With and without mask) As the p-value of the ANOVA ($P = 0.112$) and Ordinal logistic regression ($P= 0.12$) tests was greater than 0.05, it can be concluded that there is no relationship between the true age and skin quality rating of the subjects.

8. There a relationship the between age group of the participants and their skin quality rating of subjects. As the p-value of the Chi-Square test ($P = 0.00$) was less than 0.05, the null hypothesis is rejected and it can be concluded that there is relationship between age group of respondents and skin quality rating. The respondents with the age group of 51-60 tend to rate skin quality as 4 or 5 mostly.

9. There a relationship between accuracy of age prediction and the skin quality rating, without mask. As the p-value of the Chi-Square test (P

= 0.668) was greater than 0.05, the null hypothesis can be accepted and it can be concluded that there is no relationship between skin quality rating and accuracy of age prediction without mask.

10. There a relationship between accuracy of age prediction and the true age of the subject without mask. As the p-value of the Chi-Square test ($P = 0.00$) was less than 0.05, the null hypothesis is rejected and it can be concluded that there is a relationship between the true age of subjects and the accuracy of age prediction without mask.

Mixed Model Effect Summary

The mixed effect model for skin quality was built by assuming skin quality as numerical data (1-5) and considering "Image" and "Participants" as random effect as they don't have any fixed level. Other factors like gender, mask/ no mask, group, participant age, accuracy of age prediction are taken as fixed effect as they have fixed levels. From the model, the following can be concluded: With regards to random effect, both Images as well as participants are significant for the skin quality rating. With regards to fixed effects, subjects with or without mask, age of participants, drinking habit of the subject, and accuracy of age prediction were found to be significant.

5.5 Eye-tracking results

To better compare eye movements for images with and without a mask, a mask image has been generated for the images of full faces to get the overlap. The goal of the mask is compare the eye-tracking results on the same regions with and without facial features. This will allow evaluation of what areas of the face experts and non-experts look at when the image is or is not masked. The images have been divided into six parts, as shown in Figure 5.3:

- (1) Eye mask: includes the eyes, eyebrows and nose regions.

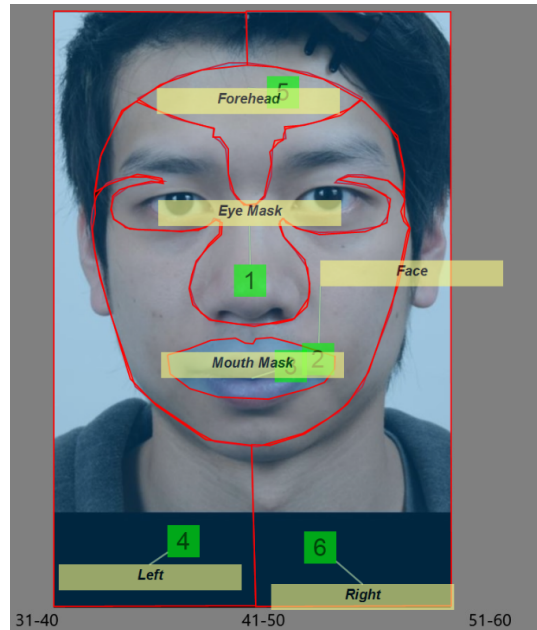


FIGURE 5.3: Eye-tracking sections on the face.

- (2) Face: includes the skin area of the cheeks and chin.
- (3) Mouth mask: includes only the mouth area
- (4) Left: includes all the areas on the left side of the face that is not included in the mask.
- (5) Forehead: includes only the forehead area.
- (6) Right: includes all the areas on the right side of the face that is not included in the mask.

The eye-tracking metrics considered include the average fixation count (i.e. for how long the eyes focused on a given region), first fixation (i.e. where did the eyes first focus) and the number of revisits (i.e. how many times the eyes went back to a particular feature).

5.5.1 Age estimation

Humans tend to estimate age by many different factors such as facial features, ethnicity, skin colour or social habits (Dantcheva and Dugelay, 2015).

This section aims to use the objectivity of eye-tracking measures to create a deep insight into how humans perceive age. Due to the complexity of the age estimation problem, this study investigates the most critical factors. First, a template of Areas of Interest (AOI) is created and applied to all images. The AOI template applied divides the images presented to the participant into different areas (eyes, face, forehead, left, mouth, right) as displayed in Figure 5.3. According to these AOI, the eye-tracking data are collected across the two sets of 50 images, once with full images and once with the masks that cover the nose, mouth and eyes.

5.5.2 Fixation count

Figure 5.4 shows the fixation counts of participants. The time spent fixating is also categorized by where the fixation time was spent (in ms) on each component of the face according to the AOI. The eyes were the feature on which participants, both experts and non-experts, fixated the most. Although both experts and non-experts tend to fixate more on the AOI of facial features, experts spent more time on skin areas compared to the non-experts (even when facial features are not masked), which indicates that experts rely more on skin quality to judge age, whereas non-experts rely more on facial features and on the face as a whole. Experts also pay more attention than non-experts to areas such as the forehead and the cheeks.

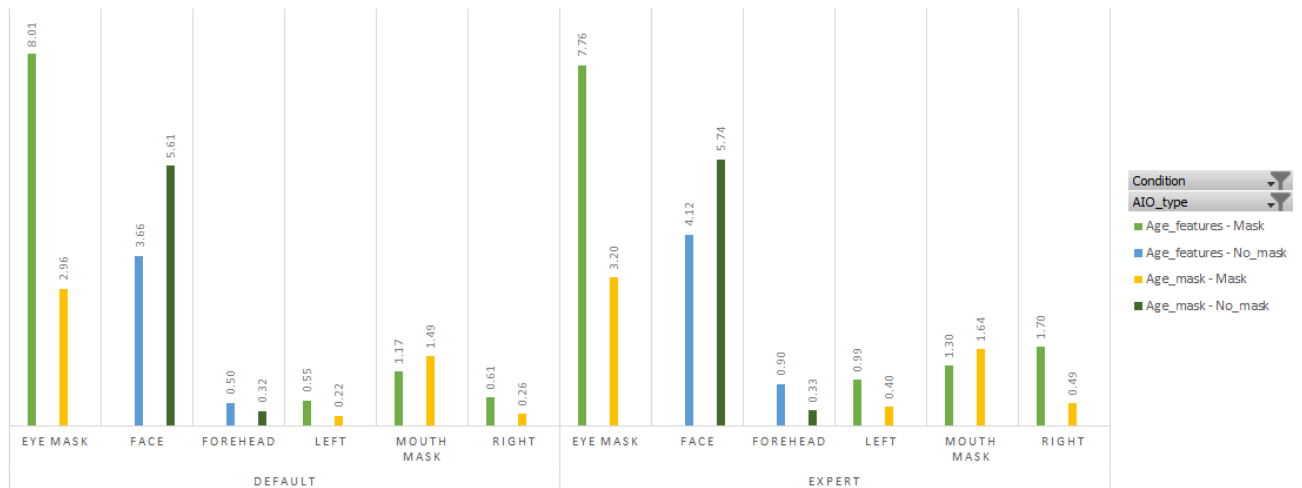


FIGURE 5.4: Average fixation count for experts and non-experts on the face parts for mask and full face.

5.5.3 First fixation

Time of the first fixation for Figure 5.5: The graph shows the time participants spent during their first fixation in milliseconds. The images are organized into six different categories that represent the areas of interest. The time spent in the first fixation is compared between "non-expert" participants and dermatologists "expert" category. It is relatively straightforward that experts spent more time in their first fixation on each facial component compared to the non-expert group. For the eye mask, the non-expert groups attempted to look at the eye mask longer time, even when the eyes are covered with a mask. Both right and left sides were fixated at more by experts to assess the skin and estimate the age compared to the non-expert group, almost ignoring these areas. Facial features get more attention from the non-expert group while experts are distributing their time during their first-time fixation and spending time of the first fixation wisely to make a judgment of estimating age. Experts are viewing the images paying attention to certain areas more than the default group, it is represented in the forehead area, where experts spent more time. The data illustrates that experts spend more time in all of

the AOI categories during the first fixation relative to the default group. Experts also are spending more time on the masked images corresponding to the default group, indicating that experts are paying more attention to facial details when the facial features, wherein the default group spends nearly equal time on each image in both masked images and facial features. However, this was not the case in the left and right side of the images, the default group seems to not spend time looking at those sides especially in masked images, and that their attempts to rely on facial features are shown that can reflect on the hypothesis that human perception of age estimation of default group can be biased when facial features are present.

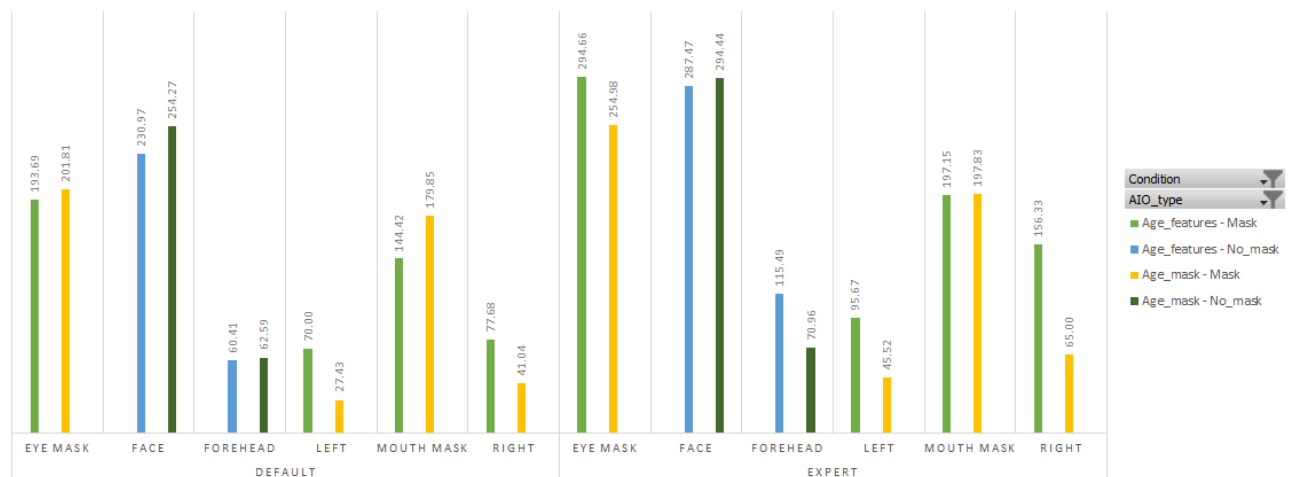


FIGURE 5.5: Average time fixation in ms for experts and non-experts on the face parts for mask and full face.

5.5.4 Revisits fixation count

The number of revisits Figure 5.6: The graph below shows the number of revisits to each component of the images collected from the eye-tracking data. This variable is heavily used to determine how affirmative the decision is, meaning that the higher number of revisiting an image component, the more uncertain the participant is about this area as supported by (vcepeda2020eye). It can also indicate the value of the category of the AOI compared to the other areas that the participant is revising the category to

confirmation of a particular decision that the participant is making to estimate the age of the image presented. The number of revisits is shown in the Figure 5.6. The eye area ranked the most revisited AOI in both groups regardless of the image was masked or not, which means participants are revisiting the eyes region to confirm their age estimation judgment the most. Experts, though, are generally revisiting areas more than the default group, which displays a comprehensive approach that they are using to make their age estimation. The default group are not revisiting important skin areas such as right, left and forehead whereas experts are paying attention to those. The general scanning to the face area is almost equal for both groups experts and default, while the number of revisits to the mouth and forehead region is higher when experts are estimating age. The differences between experts and default group are that experts are revising the areas more than the default in masked images to confirm their estimation of age where the default group tend not to verify their judgements by revisits, especially when facial features are masked, and the default group focuses more on facial features even with facial features masked images.

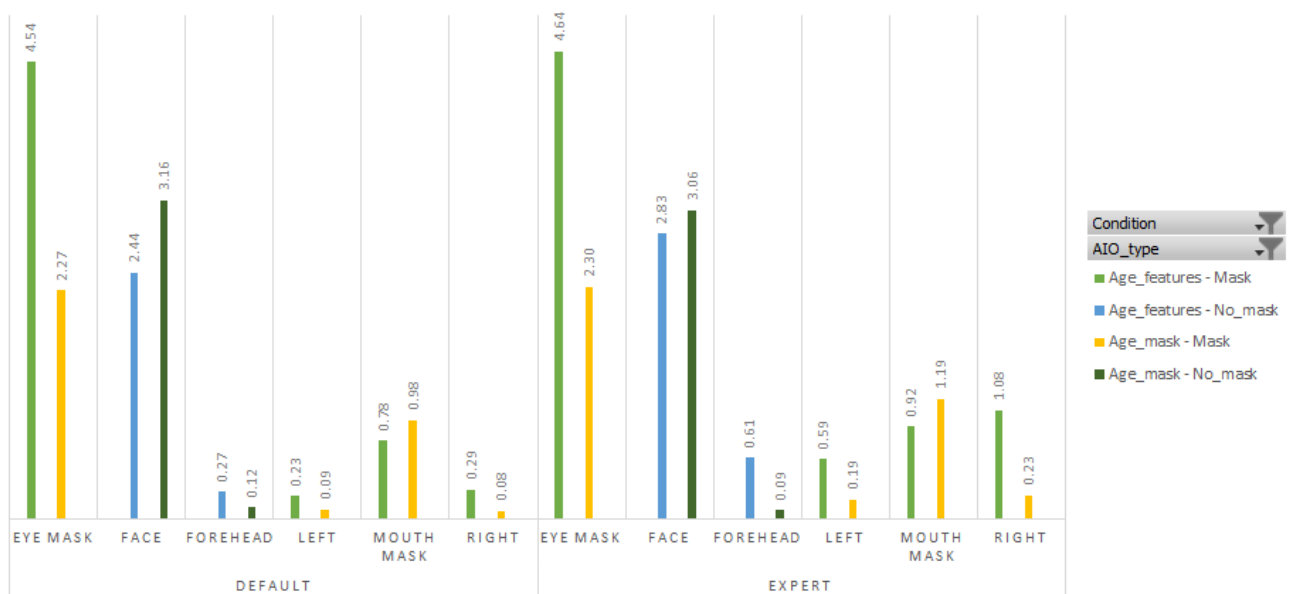


FIGURE 5.6: Average count of fixation revisits for experts and non-experts on the face parts for mask and full face.

5.5.5 Skin Quality

It is different how humans perceive the face images and reflect on their judgment of beauty assessment and attractiveness. It is a hugely complex, and a multi-element problem to be solved regarding human perception of the evaluation of skin quality (Geng, Smith-Miles, and Zhou, 2008). In this study, we aim to investigate the human perception of skin quality. Therefore, we sought to find the practical elements that experts and non-experts are focusing on to rank skin quality by depending on the time to first fixation, fixation count and the number of revisits to six different AOI of the face images that are presented to participants. Participants are asked to rank skin from (0 = very poor skin Quality up to 5 = excellent skin quality). The face images were presented to participants to rate skin quality once with the full face image presented. Another time with facial features (eyes, nose and mouth) are masked, and only regions of the skin are shown. All images included the age of the person on the image presented, which participants can make judgments according to their vision and age fact given to participants to determine their ranking. Multiple comparisons are displayed in the graphs below between experts and the default group that participated in this study.

5.5.6 Fixation count

The graph in Figure 5.7 displays the eye-tracking data that presents how much time in total was spent fixating on each AOI. The results show that the eye region is the highest AOI. Next is the overall face region that is fixated at the most, then comes the left side of the image as a third factor in terms of fixation count. As expected, experts are fixating more on the regions with larger skin areas included, and are paying attention to the actual skin regions when the face images are masked. Moreover, experts are spending more time fixating on the masked images to rank skin quality and less time fixating on

the facial features that default groups are spending their most time fixating at. The specific distribution of the fixation count is distributed in the graph below.

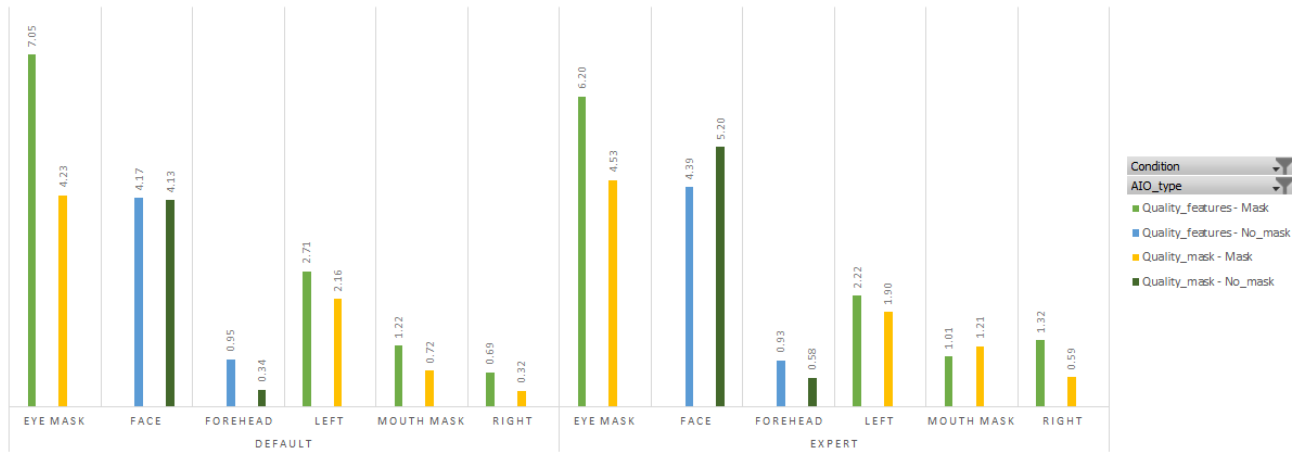


FIGURE 5.7: Average fixation count for experts and non-experts on the face parts for mask and full face.

5.5.7 First fixation duration

Figure 5.8 shows the number of milliseconds spent on each of the six AOI categories. It is clear that, in general, experts spent more time looking at skin regions compared to the default group. In addition, experts also spent more time on some areas that the default group tend to scan fast, particularly forehead AOIs. This is relevant because experts reported that the first fixation was a more critical variable they used to perform their estimates. The non-experts focused their first fixation more on the eye region, mouth, and face as a whole. The fact that non-experts spend considerable time fixating the eye region first, even when it is masked, confirms that facial features are most significant for non-experts.

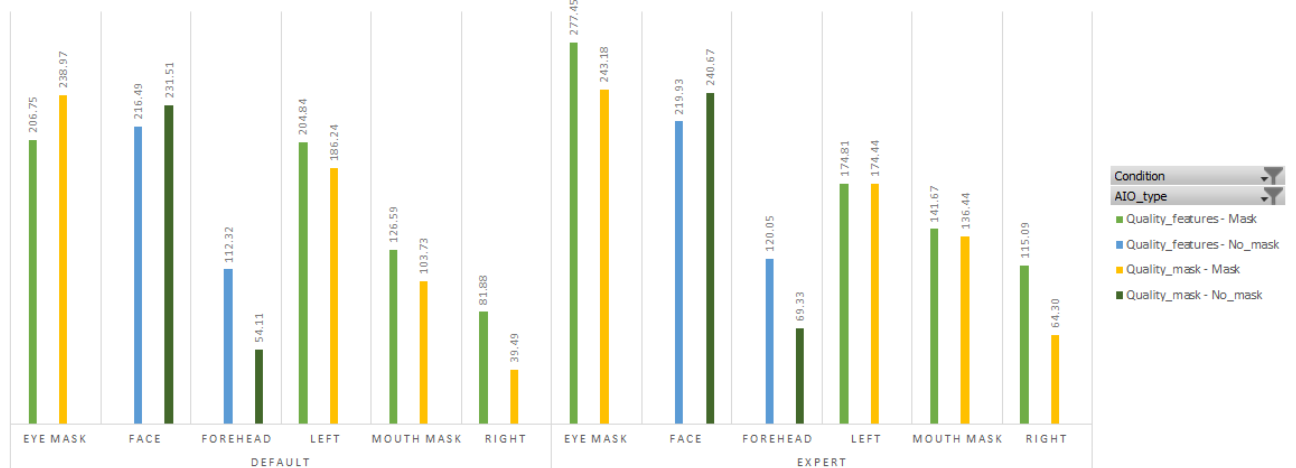


FIGURE 5.8: Average time fixation in ms for experts and non-experts on the face parts for mask and full face.

5.5.8 Revisits fixation count

Figure 5.9 shows the details of the number of revisits that the experts and non-experts made to each feature. The data shows that experts made more revisits in general, especially when looking at the masked images, which indicates that they are looking for an evidence-based ranking of the skin instead of a subjective judgment. The non-experts were more engaged and made more revisits when shown full images than when shown masked images, even when the age of the subjects is given.

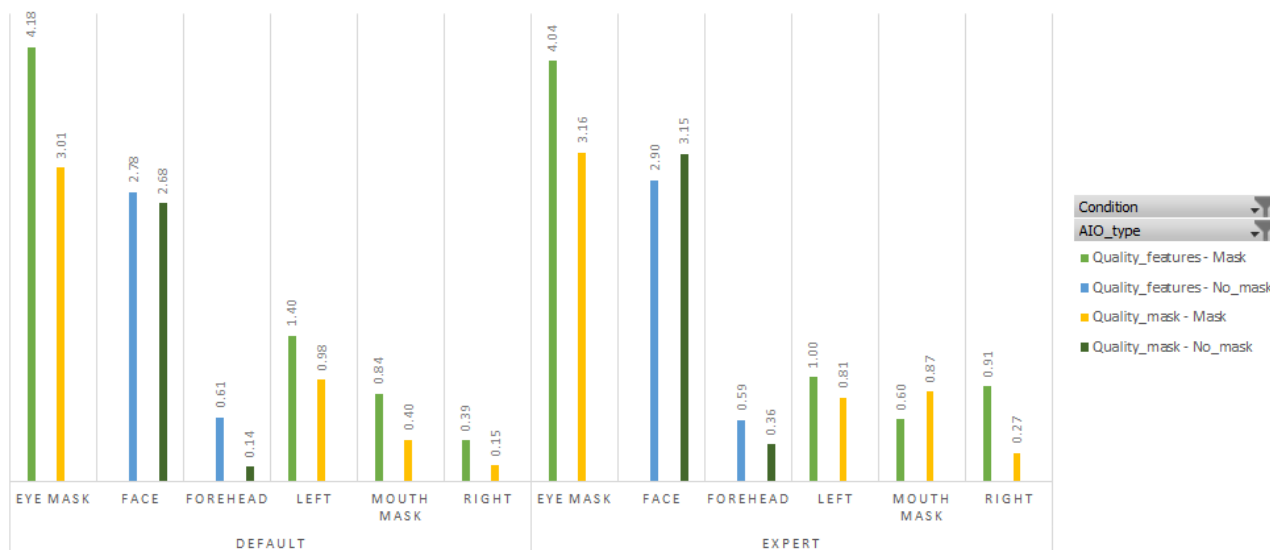


FIGURE 5.9: Average count of fixation revisits for experts and non-experts on the face parts for mask and full face.

5.6 Summary

This chapter studied human perception on age and skin quality. Age prediction accuracy was found to be impacted by appearance of facial features: prediction accuracy was higher when the full face was visible and not covered by a mask that hid features such as the nose, mouth and eyes. Furthermore, experts (i.e. dermatologists) provided significantly more accurate age predictions in comparison to default group of non-experts, but were also significantly dependent on facial features. The level of accuracy of experts increased with the level of expertise, regardless of whether the images were masked or not.

Human perception on skin quality assessment ratings were affected by facial features, and no association was found between expert and default group skin quality ratings. Experts were revisiting areas of the face to confirm their initial intuitions; non-experts relied heavily on facial features even when they were masked. Also, experts spent more time looking at skin regions rather than on features such as the eyes and mouth.

Given this information, it can be concluded that humans rely on facial features for age prediction, but not necessarily for skin quality assessment. However, how much humans rely on facial features depends on the experience level.

6 Machine Perception of Skin Quality

This chapter will focus on dermatologists perspective from the experiment in [chapter 5](#) on skin quality assessment. Then, we will examine whether machine learning is able to predict skin quality from the experts labelling on the age group with and without mask. Lastly, the dermatologists performance on machine generated skin patches.

6.1 Introduction

This chapter aims to evaluate factors that can help understand human perception of facial skin quality from the point of view of human experts, considering factors such as age, facial features, and gender. After this, the chapter will describe how machine learning was used to assess skin quality. Because there are limited amounts of data available on skin quality, a Generative Adversarial Network (GAN) was used to generate skin lesions for dermatologists to assess. The reason for generating images of skin lesions was that if dermatologists cannot tell the difference between real and machine generated images, there could be potential in using the approach to generate data for various skin attributes.

GANs were first proposed by (Goodfellow et al., [2014](#)), by training generative models where two models are trained concurrently: a generative model G

that captures the distribution of training samples and learns to generate a new sample of data; and a discriminative model D that labels the generated data as fake. The GAN model can be illustrated by the Equation 6.1, where \mathbb{E} is the entropy, $x \sim p_{data}(x)$ is the real data distribution, $z \sim p(z)$ is the generated data distribution.

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

6.2 Skin Quality Evaluation

The validation is done by correlating the experts' skin quality rating results. This section will evaluate the dermatologists mean skin quality rating for face and their corresponding mask. This will allow us to understand the rating for each image in the experiment for the full face and the mask. The top number of the below images indicates the image ID followed by mean skin quality score.

As shown in Figure 6.1, in the 21-30 age group one out of 10 subjects received the same skin quality score with and without a mask, whereas 5 of the subjects received higher skin quality score with the mask, and the 4 remaining received lower skin quality score with a mask. Subjects with prominent facial hair received higher ranking of skin quality relative to subjects without, which may indicate possible gender bias. The subjects that were scored the highest without masks, 0131 and 0133, were scored lower in images with masks. However, subject 0144, who has noticeably more prominent dark facial hair, scored higher in images with masks as opposed to images without, which may further indicate gender bias.





















0001	0119	0084	0126	0108	0005	0114	0144	0131	0133
2.5	2.7	3	3	3.3	3.4	3.5	4.2	4.4	4.4
									
0114	0119	0001	0108	0084	0126	0131	0133	0005	0144
2.2	2.2	2.5	2.8	3.7	3.7	3.9	4	4.1	4.5
									

FIGURE 6.1: Comparison of skin quality score from experts for faces and masks for age group (21-30).

Figure 6.2 shows the results for the 21-40 age group. Here, 7 out of 10 subjects received higher skin quality score with a mask, and the remaining 3 subjects received lower skin quality score without the mask. Subject 0148, with prominent facial hair, was scored significantly lower without a mask, indicating a bias towards facial hair and/or gender; this was also seen with subject 0087, who received a higher skin quality score with mask.



















0088	0128	0148	0098	0087	0092	0115	0146	0147	0121
2.1	2.7	2.7	2.9	3.1	3.1	3.3	3.7	4	4.3
									
0147	0088	0087	0098	0128	0121	0146	0115	0092	0148
2.5	3.1	3.3	3.3	3.5	3.6	3.6	3.8	4.1	4.7
									

FIGURE 6.2: Comparison of skin quality score from experts for faces and masks for age group (31-40).

Figure 6.3 shows the results for the 41-50 age group. Here, 4 (two males and two females) out of the 10 subjects scored better in skin quality when the mask was on. However, 5 females scored lower in images with masks. Four of them, subjects 0109, 0090, 127, and 0112 can be classified as pale

complexions, indicating a bias towards skin colour within this age group, which also display signs of ageing (i.e. wrinkles and sagging skin). Subject 0018, a female with a darker complexion, received a higher score with the mask, which further suggests skin colour bias within this age group.





















0018	0009	0013	0149	0096	0046	0112	0127	0090	0109
2.4	3	3.2	3.2	3.6	3.7	3.8	3.9	4	4.5
									
0127	0009	0109	0112	0090	0096	0013	0046	0018	0149
1.9	2.6	2.6	2.6	3.3	3.5	3.8	3.9	4.2	4.5
									

FIGURE 6.3: Comparison of skin quality score from experts for faces and masks for age group (41-50).

Figure 6.4 shows the results for the 51-60 age group, in which 5 out of 10 subjects received higher skin quality scores in images with masks, while 1 subject obtained the same score with or without mask, and the remaining 4 received lower scores with masks. As seen in the age group 21-30, male subjects with facial hair (0061, 0100, and 0101) scored better in skin quality images with mask, suggesting the presence of a gender bias. Male subjects with no facial hair (094 and 0118) obtained lower and same scores in images with masks, respectively. Furthermore, female subjects with fair skin tones 0130, 0083, and 0034 scored lower in images with masks, which was also observed in age group 41-50; this suggests a skin colour and/or gender bias and highlights the effect of facial features in skin quality estimation.

0061	0100	0038	0094	0118	0130	0101	0044	0083	0034
2	2.5	2.6	3	3	3.1	3.2	3.4	4	4.5
0130	0083	0094	0061	0038	0100	0118	0101	0034	0044
2.3	2.6	2.6	2.9	3	3	3	3.7	4.4	4.4

FIGURE 6.4: Comparison of skin quality score from experts for faces and masks for age group (51-60).

Figure 6.5 shows the results of the age group 61-70, of which 5 out of 10 subjects scored higher for images with mask. Within this age group, male subjects with and without facial hair (0040 and 0047, respectively) scored lower in skin quality images with masks relative to images with no masks. 5 out of 10 subjects obtained lower scores in masked skin quality images, and the remaining subjects scored higher. Subject 0122 with significant appearance of wrinkles scored highest in face images, but scored much lower in skin quality image with a mask; this variation in results highlights the influence of facial features when measuring skin quality.

0057	0034	0029	0110	0026	0040	0020	0022	0047	0122
3	3.1	3.4	3.4	3.6	3.7	4	4	4.6	4.6
0057	0029	0040	0047	0122	0034	0026	0110	0020	0022
2.7	2.8	3.2	3.2	3.2	3.7	3.8	4.1	4.3	4.3

FIGURE 6.5: Comparison of skin quality score from experts for faces and masks for age group (61-70).

6.3 Skin quality prediction

In this section, we used the data from the eye-tracking experiment; the skin quality score, age group and face and mask data. Machine learning classifiers (SVM, ANN and XGBoost [chapter 4](#)) were used to investigate if the classifiers are able to predict skin quality. All experiments are available in [??](#). The accuracy of each the prediction were approximately 30%. As indicated in [chapter 5](#)), there was only a fair agreement between dermatologists. As skin quality definition is a difficult task for human experts and with such inconsistency between the dermatologists, the research is yet ongoing for machine learning algorithms. Another key aspect is the data availability.

One way to overcome the data limitation issue associated with machine learning [chapter 4](#) is to use machine generated images. As we could not define skin quality due to human subjectivity, we want to understand the human performance in identifying the real images from machine generated images (fake). The next section will examine the ability of machine to generate photo-realistic images, i.e. the use of GANs in generating skin lesions. The objective is to validate the ability of humans (dermatologist) in identifying machine generated skin lesions.

6.4 Machine generated skin lesions

This experiment is using skin lesions as GAN requires a large-scale images to train. Dermoscopic images are publicly available and also contains detailed information of the skin, which makes it easier for the machine to learn how to generate these low level features.

GANs have been used for data generation to help the classification task. (Odena, Olah, and Shlens, [2017](#)) indicates the need of high resolution images to improve the classifier performance. Only a few research have showed

promising results when it comes to high-resolution picture production. (Karras et al., 2017) proposed using progressive-growing GANs to produce celebrity faces with a maximum resolution of 1024×1024 pixels. Another noteworthy piece of work (Ledig et al., 2017) super-resolution GAN. During training, the suggested network amplifies the training samples within the appropriate convolutional layers in order to have an effect on the output, which is the generated image in both of these networks (Wang et al., 2018), for example, employed semantic segmentation and instance mapping to generate high-resolution images. The latter argued for the use of several discriminators and generators operating at various scales to assess fine-grained characteristics and therefore enhance the global consistency of produced (synthetic) images.

We chose to conduct our experiment using a super-resolution GAN (Ledig et al., 2017) based on the findings of these networks. We chose super-resolution GAN (SRGAN) because it is open source and easy to train, which is critical given our limited computer resources. We reasoned that for synthetic skin lesions to be clinically meaningful, the lesions' images needed to be visually appealing. As a consequence, we strive to create high-quality images. We want to crop it down to the smallest possible size so that we can capture every pixel. In contrast to previous face generating GANs, where facial features are readily visible and easily generated, skin lesions are critical and need painstaking attention to detail down to the last pixel. To be aesthetically pleasing to the dermatologist, the lesions' border features must be thoroughly captured. Our objective is to create a network that can learn malignancy markers while also accounting for the unique characteristics of lesion borders. As a consequence, we chose a high-resolution GAN and fed the network a low-resolution picture. As a result, the difficulty of synthesising

skin lesions in this situation is limited to image super-resolution. Additionally, the super-resolution GAN takes use of pre-trained VGG19 features (Simonyan and Zisserman, 2014), which have proven outstanding performance in previous face generation problems and are well-known for their usage in computing the loss function for a large number of GAN networks. In addition, the VGG19 network is a good feature extractor, which is critical to the task of generating skin lesion in this thesis.

We make adjustments to the layers and the values of the momentum parameters to accommodate our computational resources. This is because we planned to slow down the activation functions' firing while raising the learning rate in order to make the network more successful at learning. The images used for this network are from the (Codella et al., 2018) dataset which are resized to 64×64 low-resolution images and amplified to 128×128 sized images. We trained the model for 30000 iterations because the dataset is too small for a GAN. When given a larger dataset, GANs learn more successfully. Note that due to lack of computational resources, we did not augment the image to 1024×1024 or 512×512 which is the closest size to the original image of the dataset.

6.5 Dermatologists evaluation

Thirty-five dermatologists were asked to diagnose fifteen images, of which ten fake images and five real images. The sample selection is based on the approach by (Song, Mukerji, and Hou, 2021), where a majority of generated images were selected and combined with a smaller sample size of real images. The outcome was that the generated images were incorrectly classified to be real instead of fake. The objective is to test the dermatologist's ability to correctly classify the generated images from the GAN. The Figure 6.7 shows examples of five real and ten fake images used in the questionnaire.

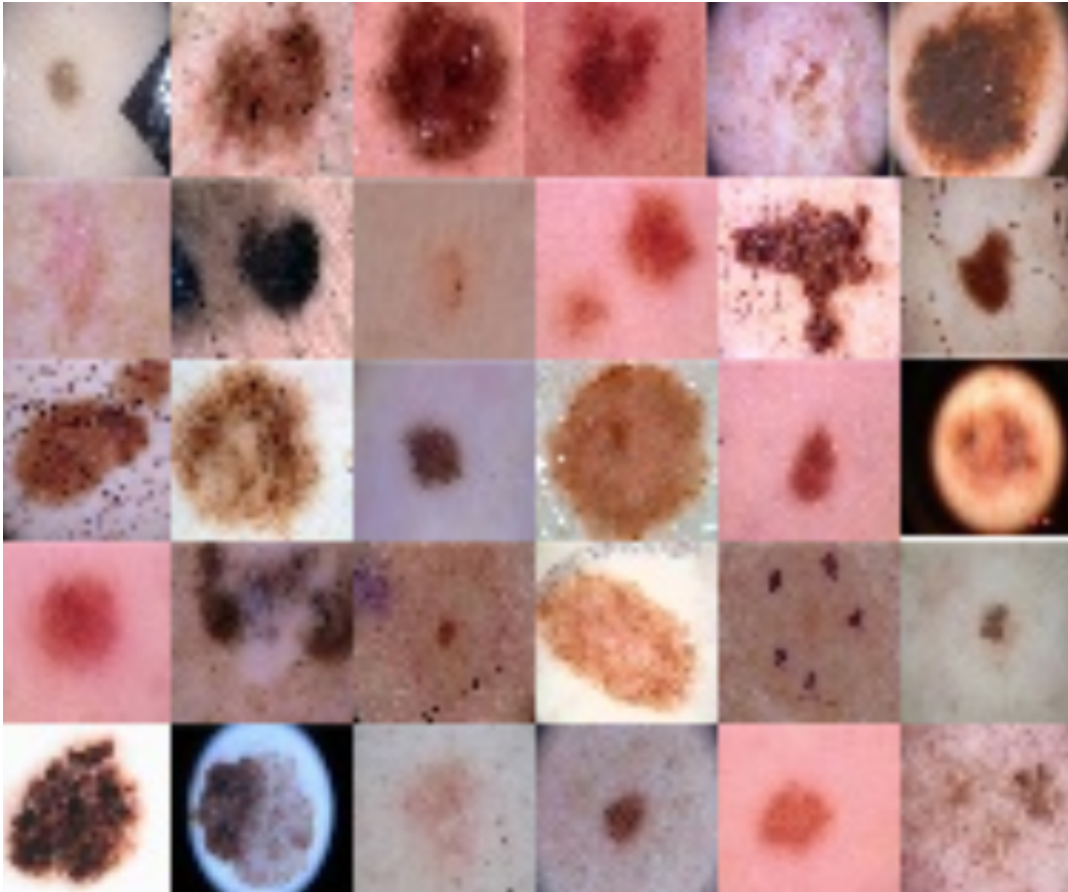


FIGURE 6.6: Examples of the fake and real images.

Future directions include the use of a local and global discriminator in conjunction with a generator (encoder-decoder) trained end-to-end as a super-resolution GAN. The generated skin lesion is segmented into foreground and background regions. The foreground region is passed through the local discriminator; while the generated skin lesion as a whole combined with the ground-truth is passed through the global discriminator

The performance of the dermatologists in predicting real and fake skin lesion images was measured. The results shows that two-third of the images were mis-classified. This seems to demonstrate even that human experts are not able to differentiate between real and fake images. This illustrates the potential of GANs to overcome the issue of data limitation. Thus, different skin quality patches can be generated with promising results.

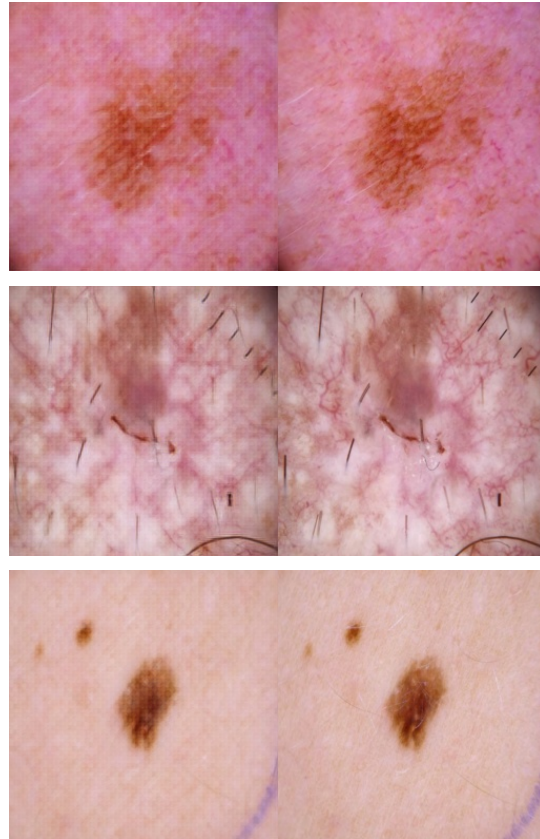


FIGURE 6.7: Comparison of generated skin lesions using SR-GAN Ledig et al., 2017 (Left) compared to ground-truth images (Right)

6.6 Summary

The first section of the chapter illustrated the mean skin quality score for the same subject with and without facial features, focusing on understanding the rating of skin quality. The results shows that facial feature influences the way human make judgment on perceived age and skin quality. Therefore, skin quality definition is very subjective and associated with facial features. The second section investigated the ability of machine learning algorithms to predict skin quality. This research recommends more future work in the area of quantitative measure to correctly quantify GAN generated images. The final section of the chapter estimated the accuracy of dermatologists assessment on machine generated skin lesions. [chapter 7](#) will focus on comparing the visual perception of humans and machine for age estimation.

7 Human Perception and Machine Prediction

In this chapter, we will look at the accuracy of age estimation for people in further detail. We will compare the human perception component to the machine perception component in order to better grasp the estimated similarities. The purpose of this chapter is to explore which face characteristics are significant in the assessment of the age of women in five age groups, both for humans and for CNNs. After that, we compare the heat maps produced by the human eye gaze with those produced by the CNN. We take into consideration two major research questions: (i) What parts of the face do people glance at when they are estimating their age?. (ii) When it comes to assessing age, do humans and machines focus on the same face areas or are they different?

7.1 Introduction

In the field of computer vision, many applications have been developed to estimate the age of people based on certain skin regions of their face. In particular, there are a number of studies that analysed how the appearance of the facial skin of females, in terms of colour and texture, can influence how others judge their age and attractiveness (Fink et al., 2012; Matts and Fink, 2010; Samson, Fink, and Matts, 2010). (Nkengne et al., 2008) found that facial

features have an effect on the overall human judgment of the face, especially in females.

To estimate the age of people from images of their faces, two types of approaches (traditional and deep learning) are being researched. Traditional computer vision methods rely on hand-crafted features that are designed based on the best understanding available of how humans perceive and define age. Deep learning methods learn to perform age estimation by directly mapping unknown data distributions to real data distributions without the use of hand-crafted features. The model learns its own features by being fed several input/output examples in which the error between the initial output of the generated data distribution and the desired output of the real data distribution is computed iteratively to tune the model's features to achieve better learned representations.

This chapter adopts such a model with the goal of predicting the age of people based on a single image of their face. Furthermore, we wish to understand what features (skin quality, eyes, mouth, etc.) humans look at when assessing the age of a person's face, and then see if these are the same features that a machine learns on its own when using a Convolutional Neural Networks (CNNs). Researchers in the field of human age perception will benefit from this chapter's comparison of human perception with that of CNNs.

7.2 Human perception of face age

(Nkengne et al., 2008) studied the influence of facial skin attributes on the perceived age of Caucasian women. The study had two objectives: to understand the influence of facial features when estimating age, and to understand the influence of gender in terms of how they estimate face age. The study consisted of 173 images of Caucasian women between 20 and 74 years old.

The participants of the study, 20 men and 28 women, were asked to estimate the age group of the woman in each image and to classify it as young (< 35 years), middle-aged (35-50 years), or senior (> 50 years). Female participants estimated age more accurately than males. However, this could be due to gender bias: As (Wright and Sladden, 2003) discovered, people are more accurate when estimating the age of people of their own gender. Nkengne et al.'s study concluded that features such as the eyes and the mouth influenced the process of age estimation. But can skin quality also be a predictor of the age of someone's face? The answer to this question, which (Mattis, 2008) suggest to be yes, would be of considerable importance in the cosmetology industry.

7.2.1 Computer methods for face age estimation

There are a number of studies which took inspiration from the human perception of age to develop computational methods for age estimation (Han, Otto, and Jain, 2013; Lanitis, Taylor, and Cootes, 2002; Geng, Zhou, and Smith-Miles, 2007)¹. Unlike these studies, which considered the whole face as global feature, (Ng et al., 2015a; Ng et al., 2018) used only some parts of the face by applying a mask to the face, obtaining what are known as local features. However, their study only focused on understanding the predictive power of wrinkles when estimating face age, and did not include other facial features, such as spots and pigmentation, that are also related to skin ageing.

CNNs accelerated the performance on a number of computer vision tasks, such as virtual recognition, using GoogLeNet developed by (Szegedy et al., 2015) and VGG16 by (Simonyan and Zisserman, 2014). VGG16 in particular has been advocated as a strong candidate for several computer visions tasks. (Zhou et al., 2016), for example, found that VGG16 outperformed all other

¹For a detailed review on computerised face age estimation using traditional, non-deep-learning-based methods, please refer to (Osman and Yap, 2018).

state-of-the-art networks in object localisation, mainly due to its low error rate. (Yang et al., 2015) developed a state-of-the-art CNN that is able to estimate the age of a person from a single image. Their model was based on the VGG16 architecture and used 0.5 million images of celebrities from the IMDB and Wikipedia datasets for training, obtaining a Mean Absolute Error (MAE) of 3.22 years of age. Thus, we chose to use VGG16 as the core architecture of our model. Because VGG16 is a network typically used for object classification, we used transfer learning to adapt it to our task. Transfer learning, which is the process of taking an off-the-shelf model and re-training its last few layers on a new, specific dataset, has been widely used in the literature. For example, (Oquab et al., 2014) showed that transferring the virtual recognition task on large-scale annotated dataset would help other recognition tasks.

One of the key advantages of CNNs over other learning-based methods is that the features that they learn when working with 2D data can be visualised. It is our goal, then, to determine if the features that our CNN model learns are the same features that humans use to predict face age.

7.3 Method

In this section, we designed two experiments in order to answer the research questions associated with the study.

7.3.1 Experiment I: Human perception

The features that humans use to predict face age can be discovered by using an eye-tracking software that tracks where on the image the eyes of the participants focus, and for how long (John et al., 2017). (Dreiseitl, Pivec,



FIGURE 7.1: The data collected from eye-tracking. Left panel: visualisation about fixation. The numbers indicate the number of ms that the gaze was focused on that area. Also, a longer gaze on an area corresponded to a larger red circle around that area. Right panel: image data converted into heat-maps to visualise the areas where the participants focused on the longest (the longer a participant's eyes focused on an area, the more red the heat-map)

and Binder, 2012) recruited sixteen participants of different expertise to diagnose twenty-eight digital dermoscopy images of pigmented skin regions, and used eye-tracking software to determine what skin regions the participants focused the most on. (Krupinski et al., 2014) used eye-tracking to understand if online training increased the accuracy and performance of dermatologists when assessing skin lesions.

Following the literature in chapter 2, we also employed eye-tracking software. The dataset we used for our study was the Social Habits dataset chapter 3, which has high-resolution images from five different age groups. From this dataset, we selected ten random images of females for visualization.

The experiment comprised thirty-six participants (19 males, 17 females; age between 21 and 70 years) who voluntarily took part in the study. The participants were asked to observe each image for fifteen seconds and then estimate the age of the person in the image. The images were displayed on a screen and an eye-tracking recorder was used to record eye movements.

7.3.2 Experiment II: CNN model

Data preprocessing

The MORPH dataset (Albert and Ricanek Jr, 2008) was used to train the CNN model. The total number of images was 44806 of which 37503 male and 7303 female, the age groups of dataset is illustrated in Table 7.1. We adopted the following pre-processing steps: face detection using Haar-based cascade classifiers (**viola2001rapid**), followed by face alignment based on the location of the eyes; lastly, resizing of the coloured image to $224 \times 224 \times 3$ to train the networks for age estimation. For testing, we used the same images that were used during the eye-tracking experiment (i.e. the same 10 random images).

TABLE 7.1: Age and gender information of 44806 samples from Album2 of MORPH dataset

	(21-30)	(31-40)	(41-50)	(51-60)	(61-70)	Total
Male	12985	12429	9150	2535	404	37503
Female	2238	2898	1806	346	15	7303
Total	15223	15327	10956	2881	419	44806

Network architecture

Visual Geometry Group (VGG16) is a Convolution Neural Network architecture that has been used to win the ILSVR (Imagenet) competition in 2014. It is commonly considered as one of the most excellent vision model architectures created to date (Rangarajan and Purushothaman, 2020). They concentrated on having convolution layers of 3×3 filter with stride 1 instead of having a huge number of hyper-parameters, and they always utilized the same padding and maxpool layer of 2×2 filter with stride 2. This is the most distinctive part about VGG16. The convolution and max pool layers are arranged in this manner across the whole design, and this is continuous throughout the architecture. After that, it has two FC (completely connected layers), which are followed by a softmax for output. The 16 in VGG16 alludes to the fact

that it contains 16 layers with different weights. This network is rather huge, with around 138 million (approximately) parameters, and it is quite complex.

The network was used for the classification and pre-trained on ImageNet (Russakovsky et al., 2014). The network was then fine-tuned to learn how to classify the five age groups on the face data.

Machine Specification

We run our experiment on a GPU machine with the following specifications: (1) Hardware: CPU - Intel i7-6700 @ 4.00Ghz, GPU - NVIDIA TITAN ×12 Gb, RAM - 32Gb DDR5 (2) Deep Learning Framework: Tensor flow.

7.4 Results

This section describes the results of both experiments and provides a visual and quantitative comparison of human and machine performance.

7.4.1 Experiment 1: eye-tracking

The eye-tracking results of Experiment I in [chapter 5](#) were analysed using BeGaze (Sensomotoric Instruments, Teltow, Germany) (Mele and Federici, 2012), which uses the raw data from the eye-tracking software to show where on the image the participants were looking at all times, and for how long. For comparison purposes, we average the thirty-six heat-maps images as shown in [Figure 7.2](#) (leftmost two columns).

The overall accuracy of the participants in estimating the age group of the women in the images was 61%. Based on the results, it is observed that it was easier for the participants to predict age group 1 (21-30) and age group 5 (61-70), with an accuracy of 86% and 65%, respectively. It was harder for

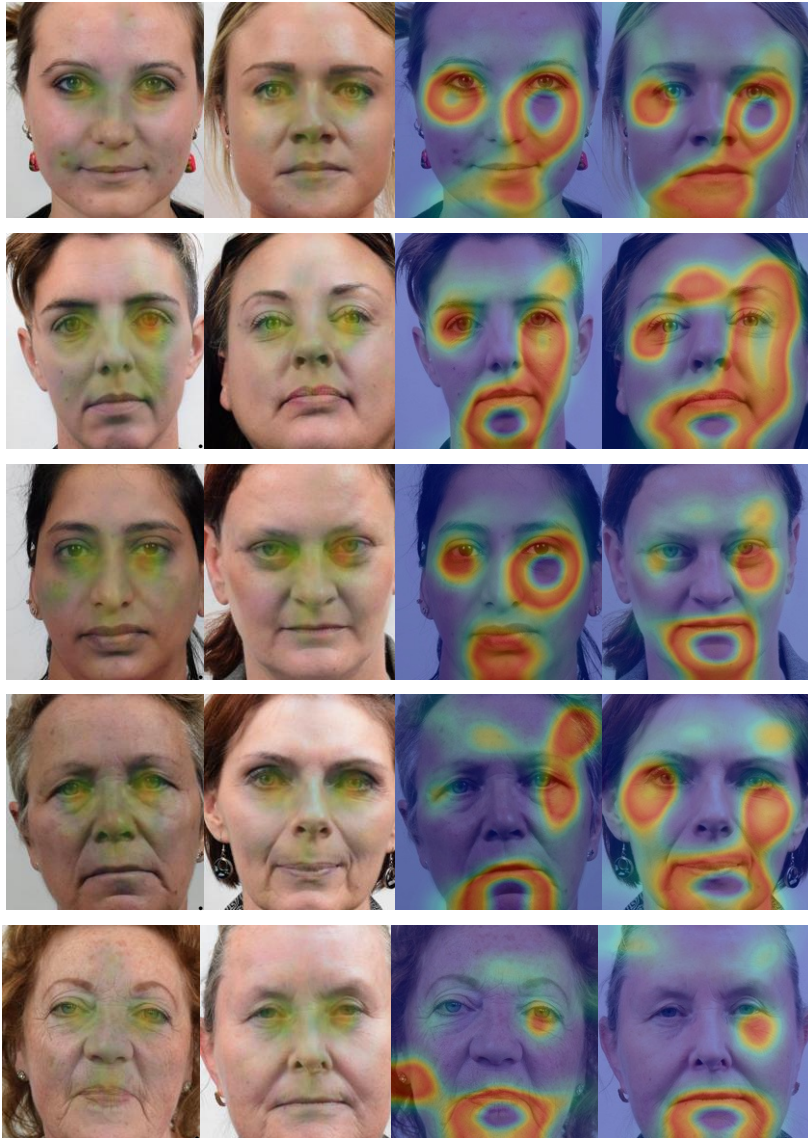


FIGURE 7.2: Comparison of human eye gaze pattern with computer model activation map. The first row is for age group (21-30), second row (31-40), third (41-50), fourth (51-60), and fifth (61-70). The number indicates the age group (leftmost two columns). Average heat-map from eye-tracking results and Grad-CAM (rightmost two columns).

them to predict age group 3 (41-50): 50% of the images of group 3 were labeled as group 2 (31-40). The grouping was used to compute the overall accuracy for age group prediction is 57%. In addition, we are focusing on the heat map visualisation as these can be compared to the activation in deep learning model. To find the main features observed by the participants in decision making, the average of each stimuli image for all the participants were generated as in Figure 7.2 (Left).

7.4.2 Experiment 2: CNN visualisation

Grad-CAM was first introduced in (Selvaraju et al., 2017) to visualise the activations of their network to gain a deeper understanding of what those features represented. In short, Grad-CAM provides a visualisation of the network's features by displaying a gradient-based class activation map.

We generated the Grad-CAM as illustrated in Figure 7.2 (rightmost two columns). The overall accuracy of the classification of the network was 60%.

7.5 Discussion

The goal of this study was to investigate a CNN-based method for the automatic estimation of face age of women, and furthermore to compare the type of features learned by this model with the features that humans use when performing the same task. The visualization results reported in Figure 7.2 show that humans considered mainly the eyes and the mouth (T-section of the face) when estimating age. A number of observations can be made by analysing:

1. The age prediction of women from a single image by the VGG16 model was comparable to the prediction by humans.

2. The results of the human eye-tracking experiment showed that when the participants concentrated their gaze on facial features such as the eyes or the mouth, their accuracy in predicting the age of the person increased; conversely, if their gaze focused more on facial skin, their accuracy dropped. This provides some evidence for the fact that indeed humans predict age based on facial features, as first suggested by the work by (Wright and Sladden, 2003). In addition, our study showed that it is easiest for humans to estimate the age of people from groups 1 and 5, and hardest to estimate the age of people from groups 2 and 4.
3. We evaluated the accuracy of the CNN model and compared its classification accuracy with that of the participants. Visually, the heat-maps showed similarities between the two experiments, except for group 5, where the features learned by the CNN were markedly different from those observed by the participants. This was mainly due to fact that group 5 is the one with the lowest number of examples, and therefore the CNN model did not have enough data to learn how to generalise well to this type of input.
4. When comparing both methods, we observed that the eye region is an important feature for both humans and CNNs. We also observed that the CNN gives a lot of importance to the mouth region, while the human gaze takes into consideration the nose region. In addition, we observed that facial skin is an important feature in the prediction done by the CNN, while it is barely taken into consideration by humans.

We answered these questions by conducting two experiments. In the first experiment, we used an eye-tracking software to detect where on the face the gaze of human participants focused the most when they were asked to assign the person in the image to an age class. In the second experiment, we used transfer learning on the network pre-trained on ImageNet, then fine-tuned

the network on a benchmark face age dataset to classify the same images shown to the participants. The heat-maps of the VGG16 network were then visualised using Gradient-based Class Activation Map (Grad-CAM). The results showed how our model was almost as accurate as humans in predicting the age of a person from a single image of their face (CNN: 60%; humans: 61%). The results also showed that people mainly look at the eyes and nose when predicting a person's age, while the features learned by the CNN included the eyes, the mouth, and the skin surface. This chapter compared human and CNN-based face age estimation. We observed similarities between the eye gaze patterns and the model's activation features when assessed on ten high-resolution images. Whilst both of the methods focused on the eye regions, the CNN approach focused more on the mouth and facial skin regions than human gaze did.

There are a number of limitations associated with this work, including the limited number of images used for both experiments (10, selected at random from the MORPH dataset). In addition, other bias factors such as the age and gender of the participants should be taken into consideration in designing future experiments.

In the future, analysing the tracking results statistically will provide a more thorough understanding of the type of features that can be extracted using eye-tracking data when humans predict the age of a person's face. In addition, novel CNN models may further improve the results obtained in this study. According to the state of the art research (Ng et al., 2018), local features on facial skin play an important role in face age estimation, particularly as a complementary feature to global face features. Hence, designing CNNs that better incorporated these regions into their predictions may further increase the accuracy of computerised face age estimation.

7.6 Summary

This chapter presents a novel perspective on human perception and machine prediction in the context of an age estimation problem. The observation is that there are similarities and differences in how humans and machines use eye areas to make decisions. Furthermore, it was discovered that the process of face ageing does not remain consistent over time, making the distinction between machine and human perception more difficult. Our algorithm was nearly as good as humans at estimating a person's age from a single image of their face, according to the results. We can reasonably conclude that the variety of factors influencing individual ageing patterns influenced human perspective. The fact that the data has been labelled by humans will have an impact on the machine learning outcome. Furthermore, confounding variables such as the participants' age and gender influenced the experiment's outcome. The scarcity of data had a significant impact on the outcomes. It would be interesting to investigate human performance on various age prediction tasks if the demographics of the participants were more diverse. In order to improve the efficacy of the proposed methodology, future research should focus on developing a large dataset. Furthermore, when dealing with smaller datasets, data re-scaling and data augmentation are important considerations. Because feature importance is critical in this type of task, future research should concentrate on developing a neural network with efficient feature extraction capabilities. Other optimization techniques, such as weight initialization or pre-training with unsupervised networks, changing the learning rate between training times, training for more epochs with varying batch sizes, and experimenting with different activation functions with and without batch normalisation, can improve network performance. All of the techniques mentioned above have the potential to improve machine performance in future research.

8 Conclusion and Future Work

This chapter describes the unique contributions by the thesis to the fields of facial skin evaluation and age estimate. In addition, it addresses possible future improvements and research areas in regards to the work that has been presented. This thesis explains its contributions, analyses their respective merits, and offers suggestions for future research. It is divided into three sections.

8.1 Introduction

The primary goal of this study was to identify the key factors that aid in quantifying skin quality efficiency. The study looked at the relationship between age and skin quality by analysing facial features to better understand human perception. Thus, another experiment identified similarities between the eye gaze pattern and the model activation when assessed on high-resolution images. Whilst both of the methods focused on eye regions, CNN's approach focused more on the mouth region and the skin region rather than human perception.

8.2 Research Findings

According to the state-of-the-art research (Ng et al., 2018), local features on facial skin plays an important role in face age estimation, particularly as a complementary feature for global face features. Hence, the design for deep

learning network architecture on skin regions has the potential to significantly improve computerized face age estimation. The future of the establishment of the machine rating of skin quality of different ages look promising. The research agrees on the fact that the human perception, which includes dermatologists, are biased towards certain facial features. The eyes and nose provide more information for age estimation than any of the other face components (forehead, eyebrows, mouth, and shape) (Nkengne et al., 2008; Han, Otto, and Jain, 2013). As a result, while examining the skin on the face, it is more objective to consider facial patches or masks. The prediction abilities of humans and machines are remarkably comparable, according to yet another research finding. We demonstrated this by employing a machine learning approach that produced results that were comparable to human performance.

Objectives versus outcomes

Objective 1. To establish high-resolution datasets for facial skin assessment	<p>1.1 The Make-up dataset section 3.2: was collected in a controlled environment and various angles of the face were captured. In addition, three make-up settings were collected (natural, with foundation and with full make-up).</p> <p>1.2 The Social Habits dataset section 3.3 was collected including social habits information about participants.</p>
Objective 2. To propose new computer methods for objective quantification of facial skin assessment	<p>2.1 A computer method for objective quantification of acne is being explored on section 3.2 dataset.</p> <p>2.2 The first to propose facial skin classification using Convolutional Neural Networks chapter 4.</p>
Objective 3. To investigate human perception on skin quality and face age estimation with and without facial features.	<p>3.1 A new insight into human perception into age estimate and skin quality assessment chapter 5.</p> <p>3.1 An eye-tracking dataset; for skin quality and age estimation. chapter 5.</p>

Objectives versus outcomes

- | | |
|---|--|
| Objective 4. To establish an objective quantification technique on skin quality assessment for various age groups | 4.1 Evaluation of skin quality assessment from dermatologists perspective.
4.2 Using machine learning for quantification of skin quality based on experts' labelling.
4.3 Thirty-five dermatologists evaluated machine generated skin lesions. |
| Objective 5. To conduct a comparative study on the performance of human and machine in skin assessment | 5.1 New findings on similarities and differences between human judgment and machine prediction on face age estimation, where eye regions are important cues for both chapter 7 .
5.2 Human and machine perception (based on the heat maps) are very similar in visual comparison. This is because the heat maps by humans and machines in chapter 7 show that when performing the age estimation task, both machines and humans made predictions based on the facial features. This is consistent with previous findings. |
-

8.3 Future Work

There were a number of limitations associated with the study of which, the number of participants took part in the experiment was limited. Other factors such as ethnicity and skin colour were not considered due to the limited data available of high resolution images. In the future, the limitation could be taken into consideration when compared the performance of human estimation to the machine prediction. Thus, using computerized methods to objectively define skin quality could be promising. To be more specific, we will explain the upcoming work in the subsections that follow.

8.3.1 Data collection

Although it was challenging to recruit participants to take part in the data collection, using computer vision methods such as GANs is promising in augmenting high-resolution images. The most difficult aspect of data collecting was convincing individuals to come to the data collection site and participate in the three steps of the data collection procedure (natural, semi make-up and full makeup). When it came to the few persons who consented to join, convincing them to wear cosmetics proved to be a tough task. Additionally, this was time-consuming, and many people were hesitant to convince others to join the data collection process. We do urge that additional data be gathered over time in the future as a potential direction. Such datasets are time-consuming to acquire, and it is anticipated that it will take 5 years before full-scale research in this subject can be undertaken fully. Other methods of data collecting have been employed, such as web crawling for CelebA-HQ datasets, however we are not certain of the accuracy of the age estimation when using this dataset, because celebrities can have cosmetic surgery or wear heavy makeup, which can cause the conclusions to be inaccurate or misleading.

8.3.2 Human perception

Eye-tracking data could be explored further to better understand the visual perception of humans, specially for age estimation. The study of human perception, in particular with regard to age estimate, is becoming increasingly popular, as a large number of researchers are looking into it (Jongerius et al., 2020, Klaib et al., 2021, Jarodzka, Skuballa, and Gruber, 2020, Horsley et al., 2013; Yadav et al., 2018; Murphy, 2011). However, the information gathered from the experiment will be used for further investigation. For example, it might be beneficial in the fields of psychology and Human-Computer Interaction (HCI). Furthermore, the assessment of human age and the development of computer vision-based human perception on age estimation are also areas that require further investigation. This will entail reviewing the meta-data from the eye tracking studies, which can be found in [Appendix C](#) of this thesis. However, because human perception of skin quality is subjective, research into it is still ongoing. This is due to the fact that the scale developed for subjective evaluation is difficult to use and can be quite costly. Therefore, we recommend that a more precise scale be developed in order to guide future research in this direction. However, there is no assurance that this will achieve the desired result because we are still confident that age estimation by dermatologists will vary from person to person and that ethnicity plays a significant part in this process. In addition, the VGG network in [chapter 7](#) is developed based on VGG16 and VGG19 to estimate age and the results show the performance of machine versus humans. Its excellent feature extractor demonstrates that machine can outperform humans in age estimation.

8.3.3 Objective skin assessment

There is potential in future work to use GANs techniques to objectively measure skin quality. As a result, it will be advantageous if researchers could investigate various objective measurement techniques that will better evaluate the quality of skin generated images by GANs. As a result, computer vision and deep learning methods will be able to better predict individual age in relation to skin quality. However, the observation is that it is subjective, and that there is potential to utilise GANs to generate synthetic data of different sorts of skin features for the benefit of machine learning in order to scientifically measure skin quality. This can be accomplished through the use of semi-supervised or supervised learning techniques to quantify skin quality. Future studies should develop skin assessments for different patches in order for research to proceed in this direction, we recommend. Unfortunately, we are aware that a shortage of data sets might be a problem. As a result, we propose that using a GAN to create additional synthetic data of skin patches, this problem might be resolved.

8.4 Summary

The research presented in this thesis contributes to major advancements in the evaluation of facial skin and age estimation. We reviewed our findings and made recommendations for further research in each contribution chapter. Our findings were reached using a combination of techniques, including a GAN and classification algorithms. Furthermore, we proved that GAN-generated images are indistinguishable from real-world data, which makes dermatologists' decision-making more difficult. This field has produced a significant amount of work, and it still has a great deal of promise for expansion into skin analysis applications that make use of GAN technology.

We are witnessing an increase in the number of consumer-based skin assessment devices being used in areas such as health care and law enforcement, as technology develops. As a result, applications such as age estimation-assisted machine prediction and skin quality analysis are likely to become more widely used in the not too distant future.

A Publications

This thesis is based on material from the following publications:

[P01] Alarifi, J.S., Goyal, M., Davison, A.K., Dancey, D., Khan, R. and Yap, M.H., 2017, July. Facial skin classification using convolutional neural networks. In International Conference Image Analysis and Recognition (pp. 479-485). Springer, Cham.

[P03] Alarifi, J., Fry, J., Dancey, D. and Yap, M.H., 2019, August. Understanding face age estimation: humans and machine. In 2019 International Conference on Computer, Information and Telecommunication Systems (CITS) (pp. 1-5). IEEE.

Other Publication

[P02] Yap, M.H., Alarifi, J.S., Ng, C.C., Batool, N. and Walker, K., 2018, September. Automated Facial Wrinkles Annotator. In ECCV Workshops (4) (pp. 676-680).

B Data Collection

This appendix includes the participant information sheet used for the Makeup dataset collection.

Thank you for agreeing to take part in this research, if you have any questions arising from the information sheet then please do not hesitate to ask the researcher before you conduct any part of the experiment. You will be given a copy of this form, along with a copy of the Release Agreement to keep and refer to during this study, and at a later date.

Study Background

The aim of this research is to analyse skin quality for female capture visual data using a high-resolution camera. Also looking at makeup application effect on skin quality and appearance.

Who can take part?

We are looking for female over 18 years old who are able to take part with and without makeup application. We will ask you to come to the designated experimental setting where we will take photographs of your face. This should last no longer than 30 minutes. You are free to withdraw from the study at any time up until the end of the study on 26th September 2020 and do not have to give us a reason. We will be happy to answer any questions you may

have and we will only begin the experiment and each condition when you are ready.

What is involved?

Upon arrival, you will fill in two sets of questions; one mainly includes daily skin and makeup routine and Social habits questions. These questionnaires are required to enable us to investigate more about facial skin. Once this has been completed, you will be sat in the photographing area, which will be lit with box lights and LEDs. Before taking any photographs, we will ensure you are comfortable in how you are sitting and take picture before and after makeup. Assuming that the participants will wear makeup at some point before or after the experiment. You will be required to bring your makeup (mainly foundation) with you, as you need to apply the makeup at some point of the experiment. The study will be explained to you again when you arrive and we will answer any questions you may have.

Data Collection and Usage

The 'data' collected in this study will mainly comprise of frontal face images taken with a digital camera. The main purpose of using these images is to use computer vision algorithms to analyse skin quality and the effect of cosmetics on facial skin features. Your images will never be coupled with your name, however we may use images in research publications to show how the pipeline of algorithm development. As your face may be included, you could be recognised by anyone who can link your visual appearance with your name. Image Matrix is non-commercial company we will work with to annotate the face. However, Image Matrix will not share the information with other partners. The data will be password protected and access will require a legal agreement to be signed. An example of a frontal face image can

be seen below.

Are there any risks in taking part in the study? There are no risks out of the ordinary for this experiment, as it only involves filling out a questionnaire and having your picture taken.

Contact

If you have any questions or require further information please contact Jhan Alarifi, a PhD candidate at Manchester Metropolitan University. You can reach him on e-mail: Jhan-saad.a.alarifi@stu.mmu.ac.uk. Alternatively, the study is supervised by Dr Moi Hoon Yap who can be contacted via: m.yap@mmu.ac.uk.

Consent Form

Before the experiments begin, each participant will fill out a questionnaire so that we are able to identify their skin quality i.e. with and without makeup. Once this has been collected they will be directed to the photographing area where they will be asked if they are comfortable before any pictures are taken.

A high-resolution digital camera will be setup to take photographs of the participant's faces. The distance of the camera will be so that a high-quality and well lit photograph will be taken of the participant's face. Several photographs may be taken for quality assurance. Then the participants will be asked to apply makeup on the face and take few photos. All data will be recorded at MMU within the usability lab. The data will be password protected, and can be accessed for use on the MMU network.

The length of time the data will be held is the duration of the project, plus additional time for use in publications. Therefore, the overall time the data will be held for is 10 years. The digital data will be anonymised and will have

a unique ID for each participant. When each participant fills in the questionnaire before the experiment, they will be assigned a participant ID written on the paper copy with their basic information (including their name). This paper copy can then be used to cross-reference the participant with the digital database created from the experiment if the participant wishes to be removed from the study.

The experiment should last no longer than 30 minutes. This includes greeting the participant, filling in the questionnaire, seating the participant before taking photographs (with and without makeup), ensuring the lighting is correct, taking photographs, and finally thanking the participant.

The study requires participants to agree to the following terms before taking part. This study is governed by the rules and regulations of Manchester Metropolitan University. Please take the time to read each point and initial to confirm you have read and agree.

1. I have read and understood the document entitled “Automated Facial Skin Analysis– Participant Information”.
2. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.
3. I acknowledge the risks associated with the study and they have been explained to me.
4. I understand that my participation is voluntary, and that I am free to withdraw at any time without giving any reason up until the end of the study on 26th September 2020.
5. I allow my photographic images (the ‘data’) to be retained as part of a dataset for research use.

6. I realise that although no images will be linked with my name, anyone who knows what I look like may be able to identify me in research publications.
7. I understand that the data collected during the experiment will be held for 10 years during the project and that the dataset will be available for academic purposes, publication to journals and conferences.
8. I agree that Manchester Metropolitan University may provide summary results in journal or conference publications of my data.
9. I understand that my data will be anonymised throughout the study.
10. I understand that the data stored digitally will be password protected, and will only have a unique participant ID associated with it.
11. I understand that the unique participant ID allocated to me on the participant questionnaire will be the only means of cross-referencing myself to the digitally stored data.
12. I understand that point (11) is done so that I may have my data removed from the study if I so wish.

I have read each point in the agreement and initialled after each one. I hereby give my consent to participate in this study.

Questionnaire

1/12/2017

The Effect of Makeup on Skin Quality

The Effect of Makeup on Skin Quality

1. What is your age?

.....

2. What is your occupation?

.....

3. What is your ethnic group?

Mark only one oval.

- White
- Asian / Asian Brithish
- Black / African / Black British
- Other:

4. Do you consume any health supplements for non-medical purpose? (eg. vitamins)

Mark only one oval.

- Yes
- No

5. If yes, what do you take?

.....

Section A

This section contains questions about cosmetics usage.

6. Do you wear makeup regularly ?

Mark only one oval.

- Yes
- No, please go to section B

7. If yes, how often ?

Mark only one oval.

- At least once a day
- More than once a day
- Once a week
- Option 4
- Only at special occasions

C The Eye-tracking raw data

This appendix includes the eye-tracking meta data collected from the experiment.

Gender, Trial, Stimulus Age, Stimulus Age group, Age Estimation, Age Group Estimation, Stimulus, Export Start Trial Time [ms], Export End Trial Time [ms], Participant Color, Fixation Count, Fixation Frequency [count/s], Fixation Duration Total [ms], Fixation Duration Average [ms], Fixation Duration Maximum [ms], Fixation Duration Minimum [ms], Fixation Dispersion Total [px], Fixation Dispersion Average [px], Fixation Dispersion Maximum [px], Fixation Dispersion Minimum [px], Saccade Count, Saccade Frequency [count/s], Saccade Duration Total [ms], Saccade Duration Average [ms], Saccade Duration Maximum [ms], Saccade Duration Minimum [ms], Saccade Amplitude Total [°], Saccade Amplitude Average [°], Saccade Amplitude Maximum [°], Saccade Amplitude Minimum [°], Saccade Velocity Total [°/s], Saccade Velocity Average [°/s], Saccade Velocity Maximum [°/s], Saccade Velocity Minimum [°/s], Saccade Latency Average [ms], Blink Count Blink Frequency [count/s], Blink Duration Total [ms], Blink Duration Average [ms], Blink Duration Maximum [ms], Blink Duration Minimum [ms], Left Mouse Click Count Left Mouse Click Frequency [count/s], Right Mouse Click Count Right Mouse Click Frequency [count/s] and Scanpath Length [px].

References

- Acne, Pimples, Skin Hairfall Treatment: CureSkin - Apps on Google Play*. URL: <https://play.google.com/store/apps/details?id=com.heallo.skinexpert>.
- Ahonen, Timo, Abdenour Hadid, and Matti Pietikainen (2006). "Face description with local binary patterns: Application to face recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 28.12, pp. 2037–2041.
- Albert, A Midori and Karl Ricanek Jr (2008). "The MORPH database: investigating the effects of adult craniofacial aging on automated face-recognition technology". In: *Forensic Science Communications* 10.2.
- Albert, A Midori, Karl Ricanek Jr, and Eric Patterson (2007). "A review of the literature on the aging adult skull and face: Implications for forensic science research and applications". In: *Forensic science international* 172.1, pp. 1–9.
- Alpaydin, Ethem (2014). *Introduction to machine learning*. MIT press.
- Amini, Mohammad et al. (2018). "Automated facial acne assessment from smartphone images". In: *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVI*. Vol. 10497. International Society for Optics and Photonics, 104970N.
- An Ongoing Commitment to Equity in Medicine*. URL: https://www.visualdx.com/diversity/?gclid=EAIaIQobChMIqJ_1-7H08QIVlobVCh2cjAsREAAYASAAEgLZiPD_BwE.

- Angulu, Raphael, Jules R Tapamo, and Aderemi O Adewumi (2018). "Age estimation via face images: a survey". In: *EURASIP Journal on Image and Video Processing* 2018.1, pp. 1–35.
- Antipov, Grigory, Moez Baccouche, and Jean-Luc Dugelay (2017). "Face aging with conditional generative adversarial networks". In: *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, pp. 2089–2093.
- Appelle, Stuart and Martin Countryman (1986). "Eliminating the haptic oblique effect: Influence of scanning incongruity and prior knowledge of the standards". In: *Perception* 15.3, pp. 325–329.
- Arakawa, Kaoru (2004). "Nonlinear digital filters for beautifying facial images in multimedia systems". In: *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on*. Vol. 5. IEEE, pp. V–V.
- Avezov, K, AZ Reznick, and D Aizenbud (2014). "Time and dose effects of cigarette smoke and acrolein on protein carbonyl formation in HaCaT keratinocytes". In: *Environmental Biomedicine*. Springer, pp. 57–64.
- Aznar-Casanova, Jose, Nelson Torro-Alves, and Sergio Fukusima (2010). "How much older do you get when a wrinkle appears on your face? Modifying age estimates by number of wrinkles". In: *Aging, Neuropsychology, and Cognition* 17.4, pp. 406–421.
- Barata, Catarina, Jorge S Marques, and M Emre Celebi (2019). "Deep Attention Model for the Hierarchical Diagnosis of Skin Lesions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0.
- Bastanfard, Azam, Melika Abbasian Nik, and Mohammad Mahdi Dehshibi (2007). "Iranian face database with age, pose and expression". In: *2007 International Conference on Machine Vision*. IEEE, pp. 50–55.
- Batool, Nazre and Rama Chellappa (2012). "A Markov point process model for wrinkles in human faces". In: *2012 19th IEEE International Conference on Image Processing*. IEEE, pp. 1809–1812.

- Batool, Nazre and Rama Chellappa (2014). "Detection and inpainting of facial wrinkles using texture orientation fields and Markov random field modeling". In: *IEEE transactions on image processing* 23.9, pp. 3773–3788.
- (2015). "Fast detection of facial wrinkles based on Gabor features using image morphology and geometric constraints". In: *Pattern Recognition* 48.3, pp. 642–658.
- (2016). "Modeling of Facial Wrinkles for Applications in Computer Vision". In: *Advances in Face Detection and Facial Image Analysis*. Springer, pp. 299–332.
- Behrents, Rolf Gordon (1985). *Growth in the aging craniofacial skeleton*.
- Berry, Diane S and Leslie Z McArthur (1986). "Perceiving character in faces: the impact of age-related craniofacial changes on social perception." In: *Psychological bulletin* 100.1, p. 3.
- Bertacchi, Marcello G. and Ismar F. Silveira (2019). "Facial Makeup Detection using the CMYK Color Model and Convolutional Neural Networks". In: *2019 XV Workshop de Visão Computacional (WVC)*, pp. 54–60. DOI: [10.1109/WVC.2019.8876943](https://doi.org/10.1109/WVC.2019.8876943).
- Bianco, Simone (2017). "Large Age-Gap Face Verification by Feature Injection in Deep Networks". In: *Pattern Recognition Letters* 90, pp. 36–42. DOI: [10.1016/j.patrec.2017.03.006](https://doi.org/10.1016/j.patrec.2017.03.006).
- Bingham, G et al. "Preliminary Studies on a Large Face Database MORPH-II". In: ().
- Braun, C. et al. (2001). *Beauty check - causes and consequences of human facial attractiveness (summary)*. The German Students Award. URL: http://www.beautycheck.de/index_eng.php.
- Brunyé, Tad T et al. (2019). "A review of eye tracking for understanding and improving diagnostic interpretation". In: *Cognitive research: principles and implications* 4.1, pp. 1–16.

- Burt, D Michael and David I Perrett (1995). "Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information". In: *Proc. R. Soc. Lond. B* 259.1355, pp. 137–143.
- Carcagnì, Pierluigi et al. (2019). "Classification of Skin Lesions by Combining Multilevel Learnings in a DenseNet Architecture". In: *International Conference on Image Analysis and Processing*. Springer, pp. 335–344.
- Caumes, Eric (2020). "Skin Lesions in Returning Travelers". In: *Hunter's Tropical Medicine and Emerging Infectious Diseases*. Elsevier, pp. 1102–1107.
- Chen, Bor-Chun, Chu-Song Chen, and Winston H Hsu (2014). "Cross-age reference coding for age-invariant face recognition and retrieval". In: *European conference on computer vision*. Springer, pp. 768–783.
- Chen, Cunjian, Antitza Dantcheva, and Arun Ross (2013). "Automatic facial makeup detection with application in face recognition". In: *Biometrics (ICB), 2013 International Conference on*. IEEE, pp. 1–8.
- Chen, Cunjian et al. (2017). "Spoofing faces using makeup: An investigative study". In: *Identity, Security and Behavior Analysis (ISBA), 2017 IEEE International Conference on*. IEEE, pp. 1–8.
- Chennamma, HR and Xiaohui Yuan (2013). "A survey on eye-gaze tracking techniques". In: *arXiv preprint arXiv:1312.6410*.
- Cho, SI et al. (2019). "Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network". In: *British Journal of Dermatology*.
- Codella, Noel CF et al. (2018). "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, pp. 168–172.

- Coleman, Sydney R and Rajiv Grover (2006). "The anatomy of the aging face: volume loss and changes in 3-dimensional topography". In: *Aesthetic surgery journal* 26.1_Supplement, S4–S9.
- Cortez, Daniel Nogueira et al. (2020). "Costs of treating skin lesions in Primary Health Care". In: *Estima–Brazilian Journal of Enterostomal Therapy* 17.
- Cox, Sue Ellen and J Charles Finn (2005). "Social implications of hyperdynamic facial lines and patient satisfaction outcomes". In: *International ophthalmology clinics* 45.3, pp. 13–24.
- Cula, Gabriela O et al. (2013). "Assessing facial wrinkles: automatic detection and quantification". In: *Skin Research and Technology* 19.1, e243–e251.
- Dalal, Navneet and Bill Triggs (2005). "Histograms of oriented gradients for human detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, pp. 886–893.
- Dantcheva, Antitza, Cunjian Chen, and Arun Ross (2012). "Can facial cosmetics affect the matching accuracy of face recognition systems?" In: *2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS)*. IEEE, pp. 391–398.
- Dantcheva, Antitza and Jean-Luc Dugelay (2015). "Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics". In: *Multimedia Tools and Applications* 74.24, pp. 11331–11355.
- Deffenbacher, Kenneth A et al. (1998). "Facial aging, attractiveness, and distinctiveness". In: *Perception* 27.10, pp. 1233–1243.
- Doherty, Stephen, Sharon O'Brien, and Michael Carl (2010). "Eye tracking as an MT evaluation technique". In: *Machine translation* 24.1, pp. 1–13.
- Dreiseitl, Stephan, Maja Pivec, and Michael Binder (2012). "Differences in examination characteristics of pigmented skin lesions: results of an eye tracking study". In: *Artificial intelligence in medicine* 54.3, pp. 201–205.
- Ebner, Natalie C, Michaela Riediger, and Ulman Lindenberger (2010). "FACES—A database of facial expressions in young, middle-aged, and older women

- and men: Development and validation". In: *Behavior research methods* 42.1, pp. 351–362.
- Ekiz, Ö et al. (2012). "Factors influencing skin ageing in a Mediterranean population from Turkey". In: *Clinical and Experimental Dermatology: Clinical dermatology* 37.5, pp. 492–496.
- Ekman, Paul (1999). "Facial expressions". In: *Handbook of cognition and emotion* 16, pp. 301–320.
- Escalera, Sergio et al. (2016). "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–8.
- Esteva, Andre et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature*.
- Face Detection. URL: <https://www.faceplusplus.com/face-detection/>.
- Fink, B et al. (2012). "Visible skin colouration predicts perception of male facial age, health and attractiveness". In: *International journal of cosmetic science* 34.4, pp. 307–310.
- Fink, Bernhard, Karl Grammer, and Randy Thornhill (2001). "Human (Homo sapiens) facial attractiveness in relation to skin texture and color." In: *Journal of Comparative Psychology* 115.1, p. 92.
- Fink, Bernhard and Paul J Matts (2008). "The effects of skin colour distribution and topography cues on the perception of female facial age and health". In: *Journal of the European Academy of Dermatology and Venereology* 22.4, pp. 493–498.
- Flament, Frederic et al. (2013). "Effect of the sun on visible clinical signs of aging in Caucasian skin". In: *Clinical, cosmetic and investigational dermatology* 6, p. 221.

- Fokuo, J Konadu (2009). "The lighter side of marriage: Skin bleaching in post-colonial Ghana". In: *Institute of African Studies Research Review* 25.1, pp. 47–66.
- Frangi, Alejandro F (2001). "Three-dimensional model-based analysis of vascular and cardiac images". PhD thesis.
- Frolov, Nikita S et al. (2019). "Dynamics of functional connectivity in multilayer cortical brain network during sensory information processing". In: *The European Physical Journal Special Topics* 228.11, pp. 2381–2389.
- Fu, Yanwei et al. (2014). "Interestingness Prediction by Robust Learning to Rank". In: *ECCV*.
- Fu, Yun, Guodong Guo, and Thomas S Huang (2010). "Age synthesis and estimation via faces: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 32.11, pp. 1955–1976.
- Fu, Yun and Nanning Zheng (2006). "M-face: An appearance-based photorealistic model for multiple facial attributes rendering". In: *IEEE Transactions on Circuits and Systems for Video technology* 16.7, pp. 830–842.
- Gao, Wen et al. (2007). "The CAS-PEAL large-scale Chinese face database and baseline evaluations". In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38.1, pp. 149–161.
- Gartstein, Vladimir and Steven A Shaya (1996). "Assessment of visual signs of skin aging". In: *Bioengineering of the Skin: Cutaneous Blood Flow and Erythema*, pp. 331–344.
- Geng, Xin, Kate Smith-Miles, and Zhi-Hua Zhou (2008). "Facial age estimation by nonlinear aging pattern subspace". In: *Proceedings of the 16th ACM international conference on Multimedia*, pp. 721–724.
- Geng, Xin, Zhi-Hua Zhou, and Kate Smith-Miles (2007). "Automatic age estimation based on facial aging patterns". In: *IEEE Transactions on pattern analysis and machine intelligence* 29.12, pp. 2234–2240.

- Gold, Jason M, Patrick J Mundy, and Bosco S Tjan (2012). "The perception of a face is no more than the sum of its parts". In: *Psychological science* 23.4, pp. 427–434.
- Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep learning*. MIT press.
- Goodfellow, Ian et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*, pp. 2672–2680.
- Guo, Zhenhua, Lei Zhang, and David Zhang (2010). "A completed modeling of local binary pattern operator for texture classification". In: *IEEE Transactions on Image Processing* 19.6, pp. 1657–1663.
- Han, Hu, Charles Otto, and Anil K Jain (2013). "Age estimation from face images: Human vs. machine performance". In: *2013 international conference on biometrics (ICB)*. IEEE, pp. 1–8.
- Hansen, Dan Witzner and Qiang Ji (2009). "In the eye of the beholder: A survey of models for eyes and gaze". In: *IEEE transactions on pattern analysis and machine intelligence* 32.3, pp. 478–500.
- Hearst, Marti A. et al. (1998). "Support vector machines". In: *IEEE Intelligent Systems and their applications* 13.4, pp. 18–28.
- Home: MoleMap New Zealand. URL: <https://www.molemap.co.nz/>.
- Hoon Yap, Moi et al. (2018). "Automated Facial Wrinkles Annotator". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0.
- Horsley, Mike et al. (2013). *Current trends in eye tracking research*. Springer.
- Huang, Gary B et al. (2008). "Labeled faces in the wild: A database for studying face recognition in unconstrained environments". In: *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*.
- Huang, Gary B. et al. (2012). "Learning to Align from Scratch". In: *NIPS*.

- Igarashi, Takanori, Ko Nishino, Shree K Nayar, et al. (2007). "The appearance of human skin: A survey". In: *Foundations and Trends® in Computer Graphics and Vision* 3.1, pp. 1–95.
- Jarodzka, Halszka, Irene Skuballa, and Hans Gruber (2020). "Eye-Tracking in Educational Practice: Investigating Visual Perception Underlying Teaching and Learning in the Classroom". In: *Educational Psychology Review*, pp. 1–10.
- Jia, Yangqing et al. (2014). "Caffe: Convolutional architecture for fast feature embedding". In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, pp. 675–678.
- John, Kevin K et al. (2017). "Do Pattern-Focused Visuals Improve Skin Self-Examination Performance? Explicating the Visual Skill Acquisition Model". In: *Journal of health communication* 22.9, pp. 732–742.
- Jongerius, Chiara et al. (2020). "The measurement of eye contact in human interactions: a scoping review". In: *Journal of Nonverbal Behavior*, pp. 1–27.
- Kanitakis, Jean (2001). "Anatomy, histology and immunohistochemistry of normal human skin." In: *European journal of dermatology: EJD* 12.4, pp. 390–9.
- Karras, Tero et al. (2017). "Progressive growing of gans for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196*.
- Kasinski, Andrzej, Andrzej Florek, and Adam Schmidt (2008). "The PUT face database". In: *Image Processing and Communications* 13.3-4, pp. 59–64.
- Kawahara, Jeremy, Aicha BenTaieb, and Ghassan Hamarneh (2016). "Deep features to classify skin lesions". In: *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*. IEEE, pp. 1397–1400.
- Khan, Rehanullah et al. (2012). "Color based skin classification". In: *Pattern Recognition Letters* 33.2, pp. 157–163.

- Klaib, Ahmad F et al. (2021). "Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies". In: *Expert Systems with Applications* 166, p. 114037.
- Korotkov, Konstantin and Rafael Garcia (2012). "Computerized analysis of pigmented skin lesions: a review". In: *Artificial intelligence in medicine* 56.2, pp. 69–90.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.
- Krupinski, Elizabeth A et al. (2014). "Understanding visual search patterns of dermatologists assessing pigmented skin lesions before and after online training". In: *Journal of digital imaging* 27.6, pp. 779–785.
- Kumar, Neeraj et al. (2009). "Attribute and simile classifiers for face verification". In: *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, pp. 365–372.
- LaBerge, David and S Jay Samuels (1974). "Toward a theory of automatic information processing in reading". In: *Cognitive psychology* 6.2, pp. 293–323.
- Landau, Marina (2007). "Exogenous factors in skin aging". In: *Environmental Factors in Skin Diseases* 35, pp. 1–13.
- Langner, Oliver et al. (2010). "Presentation and validation of the Radboud Faces Database". In: *Cognition and emotion* 24.8, pp. 1377–1388.
- Lanitis, A., C.J. Taylor, and T.F. Cootes (2002). "Toward automatic simulation of aging effects on face images". In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24.4, pp. 442–455.
- Lanitis, Andreas (2008). "Comparative evaluation of automatic age-progression methodologies". In: *EURASIP Journal on Advances in Signal Processing* 2008, p. 101.

- Learned-Miller, Erik et al. (2016). "Labeled faces in the wild: A survey". In: *Advances in face detection and facial image analysis*. Springer, pp. 189–248.
- Ledig, Christian et al. (2017). "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690.
- Leung, W-C and Ian Harvey (2002). "Is skin ageing in the elderly caused by sun exposure or smoking?" In: *British Journal of Dermatology* 147.6, pp. 1187–1191.
- Li, Wenjin et al. (2020). "Eye Tracking Methodology for Diagnosing Neurological Diseases: A Survey". In: *2020 Chinese Automation Congress (CAC)*. IEEE, pp. 2158–2162.
- Liao, D et al. (2020). "How Old Do I Look? Exploring the Facial Cues of Age in a Tasked Eye-Tracking Study." In: *Facial Plastic Surgery & Aesthetic Medicine* 22.1, pp. 36–41.
- Liu, Ziwei et al. (2015). "Deep Learning Face Attributes in the Wild". In: *Proceedings of International Conference on Computer Vision (ICCV)*.
- Lueberding, Stefanie, Nils Krueger, and Martina Kerscher (2014). "Comparison of Validated Assessment Scales and 3D digital fringe projection method to assess lifetime development of wrinkles in men". In: *Skin Research and Technology* 20.1, pp. 30–36.
- Makino, Taro et al. (2020). "Differences between human and machine perception in medical diagnosis". In: *arXiv preprint arXiv:2011.14036*.
- Maksimenko, Vladimir A et al. (2020). "Dissociating cognitive processes during ambiguous information processing in perceptual decision-making". In: *Frontiers in Behavioral Neuroscience* 14, p. 95.
- Malik, Aamir Saeed et al. (2014). "Digital assessment of facial acne vulgaris". In: *2014 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings*. IEEE, pp. 546–550.

- Mark, Leonard S et al. (1980). "Wrinkling and head shape as coordinated sources of age-level information". In: *Perception & Psychophysics* 27.2, pp. 117–124.
- Maron, Roman C et al. (2019). "Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks". In: *European Journal of Cancer* 119, pp. 57–65.
- Mathew, Juby Susan (2016). "Detection and Inpainting of Facial Wrinkles". In:
- Matts, Paul J (2008). "New insights into skin appearance and measurement". In: *Journal of Investigative Dermatology Symposium Proceedings*. Vol. 13. 1. Elsevier, pp. 6–9.
- Matts, Paul J and Bernhard Fink (2010). "Chronic sun damage and the perception of age, health and attractiveness". In: *Photochemical & Photobiological Sciences* 9.4, pp. 421–431.
- McKenzie, Naja E et al. (2011). "Development of a photographic scale for consistency and guidance in dermatologic assessment of forearm sun damage". In: *Archives of dermatology* 147.1, pp. 31–36.
- Mele, Maria Laura and Stefano Federici (2012). "Gaze and eye-tracking solutions for psychological research". In: *Cognitive processing* 13.1, pp. 261–265.
- Mena-Chalco, JP, R Cesar-Jr, and Luiz Velho (2008). "Banco de dados de faces 3D: IMPA-FACE3D". In: *IMPA-RJ, Tech. Rep.*
- Merinville, Eve et al. (2018). "What makes Indian women look older—an exploratory study on facial skin features". In: *Cosmetics* 5.1, p. 3.
- Meyers, Ethan and Lior Wolf (2008). "Using biologically inspired features for face processing". In: *International Journal of Computer Vision* 76.1, pp. 93–104.

- Mizukoshi, Koji and Kazuhiro Takahashi (2014). "Analysis of the skin surface and inner structure around pores on the face". In: *Skin Research and Technology* 20.1, pp. 23–29.
- Murphy, Cynthia A (2011). "The role of perception in age estimation". In: *International Conference on Digital Forensics and Cyber Crime*. Springer, pp. 1–16.
- Naidoo, Khimara and Mark A Birch-Machin (2017). "Oxidative stress and ageing: the influence of environmental pollution, sunlight and diet on skin". In: *Cosmetics* 4.1, p. 4.
- Nevatia, Ramakant (1982). "Machine perception." In: *PRENTICE-HALL, INC., ENGLEWOOD CLIFFS, NJ 07632, 1982, 209*.
- Ng, Choon-Ching et al. (2014). "Automatic wrinkle detection using hybrid Hessian filter". In: *Asian Conference on Computer Vision*. Springer, pp. 609–622.
- (2015a). "Will wrinkle estimate the face age?" In: *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC2015)*, pp. 2418–2423.
- (2015b). "Wrinkle detection using hessian line tracking". In: *IEEE Access* 3, pp. 1079–1088.
- Ng, Choon-Ching et al. (2018). "Hybrid ageing patterns for face age estimation". In: *Image and Vision Computing* 69, pp. 92–102.
- Nguyen, Anh, Huong Thai, and Thanh Le (2021). "Severity Assessment of Facial Acne". In: *International Conference on Computational Collective Intelligence*. Springer, pp. 599–612.
- Nkengne, A et al. (2008). "Influence of facial skin attributes on the perceived age of Caucasian women". In: *Journal of the European Academy of Dermatology and Venereology* 22.8, pp. 982–991.

- Odena, Augustus, Christopher Olah, and Jonathon Shlens (2017). "Conditional image synthesis with auxiliary classifier gans". In: *International conference on machine learning*. PMLR, pp. 2642–2651.
- Ohchi, Shuji, Shinichiro Sumi, and Kaoru Arakawa (2010). "A nonlinear filter system for beautifying facial images with contrast enhancement". In: *Communications and Information Technologies (ISCIT), 2010 International Symposium on*. IEEE, pp. 13–17.
- Oquab, Maxime et al. (2014). "Learning and transferring mid-level image representations using convolutional neural networks". In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, pp. 1717–1724.
- Osman, O. F. et al. (2017). "Automated assessment of facial wrinkling: A case study on the effect of smoking". In: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1081–1086. DOI: [10.1109/SMC.2017.8122755](https://doi.org/10.1109/SMC.2017.8122755).
- Osman, Omaira FathElrahman and Moi Hoon Yap (2018). "Computational Intelligence in Automatic Face Age Estimation: A Survey". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 99, pp. 1–15.
- Panis, Gabriel and Andreas Lanitis (2014). "An overview of research activities in facial age estimation using the fg-net aging database". In: *European Conference on Computer Vision*. Springer, pp. 737–750.
- Perrett, David I et al. (1998). "Effects of sexual dimorphism on facial attractiveness". In: *Nature* 394.6696, p. 884.
- Phillips, P Jonathon et al. (2000a). "An introduction evaluating biometric systems". In: *Computer* 33.2, pp. 56–63.
- Phillips, P Jonathon et al. (2000b). "The FERET evaluation methodology for face-recognition algorithms". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.10, pp. 1090–1104.

- Prats-Montalbán, JM et al. (2009). "Prediction of skin quality properties by different Multivariate Image Analysis methodologies". In: *Chemometrics and Intelligent Laboratory Systems* 96.1, pp. 6–13.
- Premaladha, J and KS Ravichandran (2016). "Novel approaches for diagnosing melanoma skin lesions through supervised and deep learning algorithms". In: *Journal of medical systems* 40.4, p. 96.
- Pro, Tobii (2019). *Tobii Pro Fusion*. URL: <https://www.tobiipro.com/product-listing/fusion/>.
- Puizina-Ivic, N (2008). "Skin aging". In: *Acta Dermatovenerologica Alpina Panonica et Adriatica* 17.2, p. 47.
- Quer, Giorgio et al. (2017). "Augmenting diagnostic vision with AI". In: *The Lancet* 390.10091, p. 221.
- Ramanathan, Narayanan and Rama Chellappa (2008). "Modeling shape and textural variations in aging faces". In: *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, pp. 1–8.
- Ramli, Roshaslinie et al. (2012). "Acne analysis, grading and computational assessment methods: an overview". In: *Skin research and technology* 18.1, pp. 1–14.
- Rangarajan, Aravind Krishnaswamy and Raja Purushothaman (2020). "Disease classification in eggplant using pre-trained VGG16 and MSVM". In: *Scientific reports* 10.1, pp. 1–11.
- Rexbye, Helle and Jørgen Povlsen (2007). "Visual signs of ageing: what are we looking at". In: *Int J Ageing Later Life* 2.1, pp. 61–83.
- Rhodes, Gillian, Alex Sumich, and Graham Byatt (1999). "Are average facial configurations attractive only because of their symmetry?" In: *Psychological Science* 10.1, pp. 52–58.
- Rhodes, Gillian and Tanya Tremewan (1996). "Averageness, exaggeration, and facial attractiveness". In: *Psychological science* 7.2, pp. 105–110.

- Rinnerthaler, Mark et al. (2015). "Oxidative stress in aging human skin". In: *Biomolecules* 5.2, pp. 545–589.
- Roh, Myung-Cheol and Seong-Whan Lee (2007). "Performance analysis of face recognition algorithms on Korean face database". In: *International Journal of Pattern Recognition and Artificial Intelligence* 21.06, pp. 1017–1033.
- Rothe, Rasmus, Radu Timofte, and Luc Van Gool (2015). "Dex: Deep expectation of apparent age from a single image". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 10–15.
- Russakovsky, Olga et al. (2014). "Imagenet large scale visual recognition challenge". In: *arXiv preprint arXiv:1409.0575*.
- Russell, Richard et al. (2019). "Differential effects of makeup on perceived age". In: *British Journal of Psychology* 110.1, pp. 87–100.
- Samson, Nadine, Bernhard Fink, and Paul J Matts (2010). "Visible skin condition and perception of human facial appearance". In: *International Journal of Cosmetic Science* 32.3, pp. 167–184.
- Savran, Arman, BüLent Sankur, and M Taha Bilge (2012). "Comparative evaluation of 3D vs. 2D modality for automatic detection of facial action units". In: *Pattern recognition* 45.2, pp. 767–782.
- Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". In: *Neural networks* 61, pp. 85–117.
- Selvaraju, Ramprasaath R et al. (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In: *ICCV*, pp. 618–626.
- Shaw Jr, Robert B et al. (2010). "Aging of the mandible and its aesthetic implications". In: *Plastic and reconstructive surgery* 125.1, pp. 332–342.
- Shen, Xiaolei et al. (2018a). "An automatic diagnosis method of facial acne vulgaris based on convolutional neural network". In: *Scientific reports* 8.1, pp. 1–10.

- Shen, Yujun et al. (2018b). "FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 821–830.
- Sim, Terence, Simon Baker, and Maan Bsat (2002). "The CMU pose, illumination, and expression (PIE) database". In: *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*. IEEE, pp. 53–58.
- Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.
- Singh, Bharat et al. (2015). "Layer-Specific Adaptive Learning Rates for Deep Networks". In: *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*. IEEE, pp. 364–368.
- Šitum, Mirna et al. (2010). "Skin changes in the elderly people—how strong is the influence of the UV radiation on skin aging?" In: *Collegium antropologicum* 34.2, pp. 9–13.
- Skin Cancer Melanoma Detection App* (2021). URL: <https://www.skinvision.com/>.
- Sofka, Michal and Charles V Stewart (2006). "Retinal vessel centerline extraction using multiscale matched filters, confidence and edge measures". In: *IEEE transactions on medical imaging* 25.12, pp. 1531–1546.
- Song, Suihong, Tapan Mukerji, and Jiagen Hou (2021). "GANSim: Conditional facies simulation using an improved progressive growing of generative adversarial networks (GANs)". In: *Mathematical Geosciences*, pp. 1–32.
- Suo, Jinli et al. (2007). "A multi-resolution dynamic model for face aging simulation". In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, pp. 1–8.
- Suo, Jinli et al. (2008). "Design sparse features for age estimation using hierarchical face model". In: *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, pp. 1–6.

- Suppa, M et al. (2011). "The determinants of periorbital skin ageing in participants of a melanoma case-control study in the UK". In: *British Journal of Dermatology* 165.5, pp. 1011–1021.
- Sveikata, Kestutis, Irena Balciuniene, Janina Tutkuvienė, et al. (2011). "Factors influencing face aging. Literature review". In: *Stomatologija* 13.4, pp. 113–116.
- Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.
- Taister, Michael A, Sandra D Holliday, and HIM Borrman (2000). "Comments on facial aging in law enforcement investigation". In: *Forensic science communications* 2.2, pp. 1–11.
- Taylor, Karen T (2000). *Forensic art and illustration*. CRC press.
- Tian, Dong ping et al. (2013). "A review on image feature extraction and representation techniques". In: *International Journal of Multimedia and Ubiquitous Engineering* 8.4, pp. 385–396.
- Tobin, Desmond J (2017). "Introduction to skin aging". In: *Journal of tissue viability* 26.1, pp. 37–46.
- Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler (2018). "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific data* 5, p. 180161.
- Tsukada, Akihiro et al. (2011). "Illumination-free gaze estimation method for first-person vision wearable device". In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, pp. 2084–2091.
- Wang, Li and Dennis Sng (2015). "Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey". In: *arXiv preprint arXiv:1512.03131*.
- Wang, Lipo (2005). *Support vector machines: theory and applications*. Vol. 177. Springer Science & Business Media.

- Wang, Qiuzhen et al. (2014). "An eye-tracking study of website complexity from cognitive load perspective". In: *Decision support systems* 62, pp. 1–10.
- Wang, Ting-Chun et al. (2018). "High-resolution image synthesis and semantic manipulation with conditional gans". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807.
- Watson, Stevie, Corliss G Thornton, and Brian T Engelland (2010). "Skin color shades in advertising to ethnic audiences: The case of African Americans". In: *Journal of Marketing Communications* 16.4, pp. 185–201.
- Wright, Daniel B and Benjamin Sladden (2003). "An own gender bias and the importance of hair in face recognition". In: *Acta psychologica* 114.1, pp. 101–114.
- Yadav, Daksha et al. (2018). "Unraveling Human Perception of Facial Aging Using Eye Gaze". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2221–22217. DOI: [10.1109/CVPRW.2018.00288](https://doi.org/10.1109/CVPRW.2018.00288).
- Yan, Yiqi, Jeremy Kawahara, and Ghassan Hamarneh (2019). "Melanoma Recognition via Visual Attention". In: *International Conference on Information Processing in Medical Imaging*. Springer, pp. 793–804.
- Yang, Xu et al. (2015). "Deep label distribution learning for apparent age estimation". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 102–108.
- Yap, Moi Hoon et al. "Automated Facial Wrinkles Annotator." In:
- Yap, Moi Hoon et al. (2009). "A short review of methods for face detection and multifractal analysis". In: *CyberWorlds, 2009. CW'09. International Conference on*. IEEE, pp. 231–236.
- Yi, Dong et al. (2014). "Learning face representation from scratch". In: *arXiv preprint arXiv:1411.7923*.

-
- Yu, Lequan et al. (2016). “Automated melanoma recognition in dermoscopy images via very deep residual networks”. In: *IEEE transactions on medical imaging* 36.4, pp. 994–1004.
- Yuan, Xiaojing et al. (2006). “SVM-based texture classification and application to early melanoma detection”. In: *Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE. IEEE*, pp. 4775–4778.
- Zhou, Bolei et al. (2016). “Learning deep features for discriminative localization”. In: *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. IEEE*, pp. 2921–2929.