big data and cognitive computing

MDPI

Review

Facial Age Estimation Using Machine Learning Techniques: An Overview

Khaled ELKarazle [1,*], Valliappan Raman [2] and Patrick Then [1]

[1] School of Information and Communication Technologies, Swinburne University of Technology (Sarawak Campus), Sarawak 93350, Malaysia
[2] Department of Artificial Intelligence and Data Science, Coimbatore Institute of Technology, Coimbatore 641014, India
* Correspondence: kelkaeazle@swinburne.edu.my

Abstract: Automatic age estimation from facial images is an exciting machine learning topic that has attracted researchers' attention over the past several years. Numerous human–computer interaction applications, such as targeted marketing, content access control, or soft-biometrics systems, employ age estimation models to carry out secondary tasks such as user filtering or identification. Despite the vast array of applications that could benefit from automatic age estimation, building an automatic age estimation system comes with issues such as data disparity, the unique ageing pattern of each individual, and facial photo quality. This paper provides a survey on the standard methods of building automatic age estimation models, the benchmark datasets for building these models, and some of the latest proposed pieces of literature that introduce new age estimation methods. Finally, we present and discuss the standard evaluation metrics used to assess age estimation models. In addition to the survey, we discuss the identified gaps in the reviewed literature and present recommendations for future research.

Keywords: automatic age estimation; age estimation review; deep learning; facial recognition; features extraction; image processing

Citation: ELKarazle, K.; Raman, V.; Then, P. Facial Age Estimation Using Machine Learning Techniques: An Overview. *Big Data Cogn. Comput.* **2022**, *6*, 128. https://doi.org/10.3390/bdcc6040128

Academic Editor: Fabrizio Marozzo

Received: 22 September 2022
Accepted: 21 October 2022
Published: 26 October 2022

1. Introduction

Ageing is an inevitable process that can be defined as the formation of wrinkles on the face surface and changes in core facial structures due to gravitational force, exposure to sunlight, and bone reformation [1].

In the context of machine learning, facial age estimation can be described as the process of training a model to produce a value representing one's age. This value can either be an age range (classification problem) or an exact age value (regression problem). In order to build an automatic age estimation model, the first step is acquiring a suitable training and testing dataset. Currently, there are plenty of publicly available datasets with labelled samples of various illuminations, head poses, and conditions.

The next step is the pre-processing stage, which includes cropping the images to avoid background noises after detecting the face in an image. This step is essential to remove any unwanted background noise. The detected face in each image is then rotated and aligned to normalize all the samples and ease the training process. Following this step is the feature extraction stage, in which we extract discriminative ageing features such as wrinkles or the head structure. Finally, the features are fed to a classifier or a regression-based model, which learns the different patterns of each age group. After training, testing commences, and various evaluation metrics are available to assess the performance of the final model. We summarise this process in Figure 1.

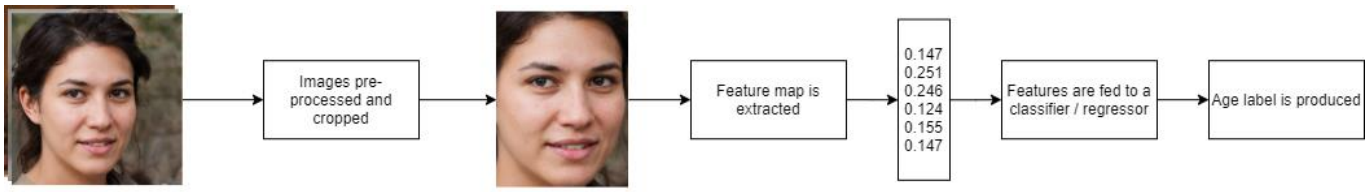

Figure 1. Overview of training a typical age estimation model.

Although the process of building an age estimation model is straightforward, there are several issues researchers encounter when attempting to build these models. Ageing is unique to each individual since it depends on various internal factors such as ethnicity, gender, health condition, lifestyle, and external factors such as degree of sunlight exposure, makeup, overall physical environment, or facial scars. These factors contribute to the increase in inaccurate predictions during testing. In addition to the abovementioned factors, data inconsistency and the lack of enough diverse samples covering different genders and ethnicities are also reasons for poor performance.

This paper surveys several recently proposed age estimation methods as well as some of the common challenges. In addition, we present a list of benchmark datasets available to train and test age estimation models. Moreover, we discuss the gaps identified in the reviewed literature to better understand the current state of research.

The methods reviewed in this manuscript are listed in ascending order based on their publication date. The oldest papers are reviewed first, and the most recent ones are reviewed last. The reviewed papers are selected based on the following criteria:

(1) Relevance of the proposed method to the problem of age estimation.
(2) Novelty of the proposed method compared to similar methods.
(3) The impact the reviewed method has made based on the number of citations and mentions.
(4) The challenges a method is attempting to overcome.

The paper is divided into eight sections. Section 1 introduces the content of the manuscript and the objective of this paper. Section 2 lists our contributions. Section 3 presents the common challenges researchers face when building age estimation models. Section 4 describes the existing benchmark datasets that are available to build age estimation systems. Section 5 presents the different techniques used to build age estimation models. Section 6 explains the evaluation metrics. Section 7 provides a literature review of the existing methods. Section 8 presents a discussion of the findings. Finally, Section 9 concludes the paper by highlighting the main items discussed during this study.

2. Contributions

The contributions to the body of the literature are summarised below:

1. Several architectures and methods used to build facial age classifiers are reviewed.
2. Several common challenges faced by researchers are described.
3. An analysis of the benchmark datasets used to build age estimation systems is presented.
4. A survey on several recently proposed age estimation methods is provided.
5. A discussion of the existing gaps and the trends based on the surveyed papers is demonstrated.

3. Common Challenges

There are many challenges researchers encounter when building age estimation models. Some of these challenges are considered "controllable", in which further tweaking of the model or introducing new measures is required to overcome them. Examples of these controllable challenges include head pose, image quality, or image resolution. In contrast, "uncontrollable" challenges are those that scientists at the current stage of machine learning research cannot control; thus, creating significant roadblocks to achieving better age estimation accuracies.

3.1. Head-Pose and Alignment

One of the common controllable challenges is the head pose and alignment, which refers to the position of the face in a given image. As images are captured in real-life conditions, the position of faces tends to vary in terms of alignment. This issue is usually resolved during the pre-processing stage, in which the face is detected and realigned. In [2], the authors reported that head poses and alignment contributed to the decrease in performance.

3.2. Image Resolution

In any computer vision problem, the resolution of training samples plays a significant role in shaping the accuracy of an age estimation model. Images of lower quality tend to lose critical ageing features such as wrinkles or the shape of a face, preventing the model from learning all the necessary discriminative features to tell ages apart. This issue is caused by the various resolution image capturing devices poses; however, it can be solved by normalising the resolution of both training and testing samples through different image upsampling methods. Researchers in [3] have shown that enhancing the resolution of the images prior to training improves classification accuracies.

3.3. Lifestyle and Health Condition

A person's lifestyle is one of the considerable uncontrollable challenges that prevent most age estimation models from performing well. Depending on one's health condition and lifestyle, they can appear older or younger than their actual age; thus, confusing the model that attempts to predict their age.

3.4. Lack of Data

Overfitting is a severe issue in age estimation, and it occurs as a result of the lack of diverse facial images for training. Although large-scale facial datasets exist for various facial analysis problems, most samples are either unlabelled or do not cover enough examples, such as subjects from different genders and ethnicities. Several studies such as [4] and [5] have stated that the lack of data is the main cause of most of the misclassified samples.

3.5. Genetics

The ageing pattern of an individual depends primarily on one's genes. This issue is one of the roadblocks that existing machine learning models cannot solve since it depends on the person's background, gender, and the climate in which they grew up. In [6], the researchers showed a relationship between the subject's gender and their facial age.

3.6. Facial Modifications

The presence of facial accessories or facial hair often causes certain features to become unclear or disappear entirely from an image. This issue is considered uncontrollable; however, there is room for more research to develop a model capable of normalising all images by removing particular facial accessories such as beards, piercings, or makeup. The findings in [7] demonstrated a relationship between the degraded accuracy and facial modifications.

4. Datasets

Acquiring a suitable training dataset is the most critical step in building a machine learning model. However, it is challenging to obtain a perfect dataset as almost every dataset suffers from data disparity or uneven distribution of samples. This section analyses 17 different datasets by looking into the number of images, the distribution of samples, and the condition of the images. We provide a summary of each dataset in Table 1.

Table 1. Datasets available for building and training age estimation models.

Dataset	Number of Samples	Age Group	Condition
IMDB-WIKI	523,051	1–90	Unconstrained
Human and Object Interaction Processing (HOIP)	306,600	15–64	Constrained
The Asian Face Age Dataset (AFAD)	164,432	15–40	Unconstrained
Cross-age Celebrity Dataset (CACD)	163,446	16–62	Unconstrained
WebFace	494,414	1–80	Unconstrained
MORPH	55,134	16–17	Constrained
Specs on Face (SoF)	42,592	1–52	Unconstrained
MegaAge	41,941	0–70	Unconstrained
Adience	26,580	0–60	Unconstrained
UTKFace	23,000	0–116	Unconstrained
AgeDB	16,488	1–101	Unconstrained
MSU LFW+	15,699	0–20	Unconstrained
Facial Recognition Technology (FERET)	14,126	1–66	Unconstrained
YGA	8000	0–93	Unconstrained
Images of Group (IoG)	5080	0–66	Unconstrained
Iranian Face Database (IFDB)	3600	2–58	Constrained
FGNET	1002	0–69	Constrained

4.1. IMDB-WIKI

By far, IMDB-WIKI [8] is the largest publicly available dataset with up to 523,051 labelled samples of 20,284 individuals aged between 1 and 90 years old. This dataset combines 460,723 and 62,328 samples taken in unconstrained conditions of celebrities from IMDB and Wikipedia, respectively. Most of the samples are of individuals between 20 and 50 years old, with fewer images of individuals aged 20 and below. The dataset is available online for academic research use.

4.2. Human and Object Interaction Processing (HOIP)

HOIP [9] contains 306,600 images of 300 individuals between the age of 15 and 64 years. The images were taken in controlled conditions such as illumination and head pose. In this dataset, there are ten age groups, and each age group consists of 30 images, with 15 samples belonging to females while the rest belonging to males.

4.3. The Asian Face Age Dataset (AFAD)

AFAD [10] is another relatively large dataset with 164,432 images of individuals between 15 and 40 years old. About 38% (approximately 63,000) of the images represent female subjects, while the remaining 62% (approximately 100,752) represent male subjects. The images were taken in uncontrolled environments with various illuminations and head poses.

4.4. Cross-Age Celebrity Dataset (CACD)

CACD [11] was initially introduced for facial recognition tasks; however, it was then used to train age estimation models. It consists of 163,446 facial images of 2000 celebrities aged 16 and 62 years old. Samples in this dataset are taken in both controlled and uncontrolled conditions. There is no clear breakdown of how the samples are distributed among the age groups and genders.

4.5. WebFace

WebFace [12] consists of 494,414 facial images of 10,575 individuals taken in uncontrolled conditions. The dataset covers ages between 1 and 80 years old, and it is a result of scraping images from Google and Flicker.

4.6. MORPH

MORPH [13] is by far the most used dataset to build and train age estimation models. The dataset consists of 55,134 images taken in controlled conditions. The samples in this dataset represent 13,618 individuals between the age of 16 and 77 years old. The images in this dataset are distributed over two albums, MORPH and MORPH-II.

4.7. Specs on Face (SoF)

SoF [14] consists of 42,592 facial images, 112 of which 66 are males, and 46 are females, taken in uncontrolled environments with extreme variations in illumination and face occlusions. The dataset is free for academic and research use.

4.8. MegaAge

MegaAge [15] contains 41,941 unconstrained images of subjects between 0 and 70 years of age. The images are all taken in unconstrained conditions, and each image is annotated with posterior labels. The publishers of MegaAge also released a single ethnicity dataset titled MegaAge-Asian, which exclusively contains samples of Asian subjects.

4.9. Adience

Adience [16] comprises 26,580 facial images of 2284 subjects taken in uncontrolled conditions. The images are all labelled with binary gender labels and age groups. The age groups are 0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, and 60+.

4.10. UTKFace

UTKFace [17] contains over 20,000 images of individuals between 0 and 116 years old. The images are taken in unconstrained conditions with various illuminations, occlusions, and resolutions. The images are denoted by age, gender, ethnicity, and timestamp.

4.11. AgeDB

AgeDB [18] consists of 16,488 manually collected unconstrained images of 568 subjects between the age of 1 and 101 years old. The images are labelled with both age and gender. The training images in this dataset were collected through a web search, and according to the authors, several labels may be inaccurate.

4.12. MSU LFW+

MSU LFW+ [19] is an extension of the LFW [20] database, and it is used widely for training facial recognition models, including age, gender, and ethnicity estimation. The dataset consists of 15,699 unconstrained images of 8000 individuals denoted by age, ethnicity, and gender.

4.13. Facial Recognition Technology (FERET)

FERET [21] contains a total of 14,126 images of 1199 individuals. Several images were taken in uncontrolled conditions, while some were captured in controlled environments. This database consists of images of several ethnicities; however, the ethnicity's label is not present. Nevertheless, this database is common in tasks related to facial recognition and age estimation.

4.14. YGA

YGA [22] consists mainly of Asian individuals with 8000 images of 1600 individuals between 0 and 93 years of age, each contributing to five approximately labelled images. The photos were taken outdoors, and the database is evenly divided into 800 women and 800 men. The images in this database contain different illuminations and facial expressions.

4.15. Images of Group (IoG)

The images IoG [23] contain more than a single labelled face. The dataset consists of 5080 images with a total number of 28,231 faces. This dataset has seven age groups: 0–2, 3–7, 8–12, 13–19, 20–36, 37–65, and 66+. The images in this dataset were all taken in uncontrolled conditions.

4.16. Iranian Face Database (IFDB)

IFDB [24] contains 3600 coloured images of 616 individuals between 2 and 85 years. It is mainly used in age classification and ethnicity estimation tasks because of the diversity of the samples. The database comprises samples for 787 men and 129 women. The data are diverse, and there are variations in poses, expressions, and facial accessories; however, all images were captured under controlled conditions.

4.17. FG-NET

FG-NET [25] is commonly used to build age estimation models. It contains 1002 coloured and grayscale images of 82 subjects aged between 0 and 69 years old. The images were all taken in controlled conditions.

5. Age Estimation Models

In this section, we present some of the techniques used to build age estimation models. Furthermore, we discuss the advantages and disadvantages of each of these methods. On a high level, we can classify age estimation models as handcrafted and deep learning-based models.

Handcrafted-based methods usually combine filters, such as Histogram of Oriented Gradients (HOG) or Local Binary Pattern (LBP), to extract edges and shapes from a facial image. A learning algorithm of choice such as K-nearest neighbour or support vector machine is then added to learn the extracted features.

In contrast, deep learning facial extraction methods rely more on using algorithms such as convolutional neural networks (CNNs) or Multilayer preceptors (MLPs) to extract useful information from a given image. In deep learning-based methods, a fully-connected neural network is employed to learn the extracted features.

Usually, handcrafted methods tend to be less computationally expensive and more efficient but less accurate and time consuming to build. On the other hand, deep learning methods are more accurate; however, they require more computational power to process images. The performance of every model, regardless of whether it is handcrafted or based on transfer learning depends on the dataset used and the design of the model. Therefore, it is not possible to report an average accuracy or performance of an age estimation model without knowing details such as datasets and training mechanism.

In addition, handcrafted methods are usually built for low-end devices with limited computational resources. While deep learning models built-from-scratch and pre-trained models using transfer learning are built for devices with decent computational power.

In Table 2, a comparison of deep learning models built-from-scratch, pre-trained, and handcrafted models is presented.

Table 2. Comparison of the different methods of build age estimation models.

Method	Advantages	Disadvantages
Built-from-scratch	More control over the model's design and all features can be extracted without manually deciding the features.	Requires more time to train and build.
Pre-trained models	Works without the need for lots of samples and the learning time is shorter.	Performs badly if it was pre-trained on a different domain and no control over the fine details of each layer.
Handcrafted	Can be deployed to low-end devices.	Does not perform well since features are manually extracted.

5.1. Handcrafted Models

In handcrafted models, we extract various ageing features manually. Extracting facial features begins by using either a single filer or a combination of different filters such as Gabor [26], Histogram of Oriented Gradients (HOG) [27], or Sobel filters [28]. The filter's parameters are manually tweaked to acquire as many features as possible. The filters are slid over an image to extract features such as wrinkles, head shape, textures, or edges that indicate the subject's age. Handcrafted models generally require less computational power as they are not as complex as their deep-learning counterparts. However, one drawback is that these models are usually less accurate. This section analyses several handcrafted methods used to manually extract facial features from a photo. Generally, there are four methods:

1. **Anthropometric models**: Anthropometric measurements are quantifiable dimensions of the bones, muscles, and the human body's overall structure [24]. These measurements help us understand the geometrical structure of the human body and can help us distinguish between different age groups and genders. Several studies [25,26] used anthropometric models We can represent the anthropometric models in Equation (1) in which R is the circle's radius (face) and theta is the initial angle formed with the vertical axis. K is a variable that increases over time and R' represents the growth of the human face over time.

$$R' = R(1 + k(1 - \cos\theta)) \quad (1)$$

2. **Texture-based models**: Unlike Anthropometric models, which use the distance between facial points, texture-based models depend on the texture of facial images, which the pixel intensity can represent. In texture-based models, various image texture operators such as local binary patterns [29] (LBP) or biologically inspired features [30] (BIF) are employed to extract skin areas such as spots, lines, or edges that might represent wrinkles. Several studies, such as [31,32], used texture-based models to estimate age. Unlike anthropometric models, texture-based models can perform well on images taken in uncontrolled conditions; however, they are incapable of discriminating different shapes and distances between facial points.
3. **Active appearance models (AAM)**: Active appearance models are statistical models commonly used in facial image representations, combining anthropometric and texture-based model descriptors. A dimensionality reduction algorithm such as the principal component analysis (PCA) learns the extracted texture and shape features. AAMs are ubiquitous in various facial recognition, facial verification, and age estimation tasks due to their flexibility in working with textures and shapes. However, reducing the feature's dimensionality results in having several ageing features, such as wrinkles going unnoticed. Studies such as [33,34] used AAMs with various other methods to predict age.
4. **Ageing pattern subspace (AGES)**: In a study conducted by [35], ageing pattern subspace (AGES) was proposed to identify a person's ageing pattern based on a set of facial images taken at the age of 2, 4, and 8 sorted in ascending order. The reason for placing images in this order is to learn the ageing pattern of an individual, defined as a "sequence of personal face images sorted in time order" according to [36]. This method generates missing age samples by learning the subspace representation of a single image when constructing a sequence of the subject's ageing facial images. AGES can help estimate missing samples; however, it does not work well with wrinkles; therefore, integrating texture-based models with AGES is common. Studies such as [37,38] have used AGES by itself to extract facial features from photos. While studies such as [33,34] used AAMs with various other methods to predict age.

Age manifold: Age manifold focuses on treating ageing patterns as a trend for several subjects at various ages instead of finding a specific ageing pattern for each person. The flexibility of the ageing manifold method allows for subject representation to be in the form of

one image or several images at different ages. Comparing the age manifold to a close equivalent (AGES), we find that models based on this method can learn low-dimensional ageing patterns that AGES could ignore. The low subspace is defined using conformal embedding analysis (CEA) [39], which obtains features and reduces dimensionality through discriminating analysis and conformal mapping. Both these techniques project high-dimensional data onto a unit hypersphere. Unsupervised dimensionality reduction techniques are not effective for handling discriminators. Instead, [40] suggested an alternative to reduce dimensionality using a supervised algorithm, denoted as orthogonal locality preserving projections (OLPP), whereby age is first predicted using a regression function and is then locally adjusted to match the correct values within a boundary.

5.2. Deep Learning Models

Deep learning models work differently than their manual counterparts. After acquiring and pre-processing the images, we feed them to a deep neural network. The network may consist of several convolutional neural network layers (CNN) [41], pooling layers, dropout layers, batch normalisation layers, or residual connections. We then define the number of filters and the size of each kernel. The layers in the network will automatically extract ageing features as the network continue processing the input images. Deep learning models are usually more accurate than their handcrafted counterparts as we allow the model to decide on the essential features to learn. One of the significant drawbacks to this is the high computational cost, as these models can become enormous and consume more computational power and time. Deep learning models can be either trained from scratch or based on pre-trained models.

1. **Built-from-scratch models**: One way to extract facial features with deep learning is to develop a deep learning algorithm from scratch. In this method, we can define a convolutional neural network that may consist of several convolutional layers, dropout layers, activation functions, pooling layers, and fully connected layers. The convolutional and pooling layers construct the feature maps, and we use dropout layers to avoid overfitting by turning off randomly selected neurons. The fully connected network receives the extracted features and learns a mapping function. Using a network from scratch could be computationally expensive and time-consuming as we will have to keep fine-tuning the parameters, and the network could grow exponentially.
2. **Pre-trained models**: Pre-trained models are time and space-efficient alternatives to the deep learning models built from scratch. We use and fine-tune a network that has been trained on a more complicated task to extract features from an image. The model's hyperparameters are usually fine-tuned and adjusted. VGG-16 [42], VGG-19 [43], ResNet50 [44], and AlexNet [45] are examples of pre-trained models that have shown state-of-the-art accuracies in facial recognition tasks, including age estimation. Table 3 presents some of the well-known neural networks used to build age estimation models.

Table 3. Examples of commonly used pre-trained networks.

Network Name	Number of Layers	Number of Parameters (Millions)	Size (MB)	Default Input Size (Height × Width)
VGG16	16	144	515	224 × 224
VGG19	19	144	535	224 × 224
ResNet50	50	25.6	50	227 × 227
Xception	71	22.9	85	299 × 299
AlexNet	8	61	227	227 × 227
GoogLeNet	22	7	27	224 × 224
MobileNetV2	53	3.5	13	224 × 224

6. Performance Evaluation Metrics

Evaluating the performance of any machine learning model is essential to determine the efficiency and generalisation of a proposed model. There are several ways to evaluate

the performance of age estimation models, and each method depends on the model's architecture. In this section, we look into the frequently used metrics and their definitions.

Since age estimation is a task that can be treated as a classification or a regression problem, researchers have a vast array of loss functions they can use to evaluate the performance of their proposed models. Mean absolute error and mean squared error are both available for regression tasks while accuracy and cumulative score are available for classification problems.

The definition of MAE, which primarily evaluates regression models, is as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{2}$$

Y represents the predicted age, x represents the actual age, and n is the number of images. MSE, on the other hand, is defined as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i)^2 \tag{3}$$

Similar to Equation (2). Y represents the actual age, while X represents the predicted age of the ith element. In Equations (4) and (5), a lower value indicates that the model performs well, while a higher MAE or MSE indicates that the error margin is large; thus, the model is not performing well. If the model is training on several age classes, then Equations (2) and (3) might be unsuitable. Therefore, a metric such as cumulative score (CS) may be more suited for the model. The equation of CS is as follows:

$$\text{CS} = \frac{n}{N} * 100\% \tag{4}$$

"n" represents the total number of correctly classified images, while "N" represents the total number of testing images. The CS equation can be rewritten to calculate the number of correctly classified images where the error is not greater than a specific value representing years. The new CS equation is as follows:

$$\text{CS} = \frac{N_{e \leq j}}{N} * 100\% \tag{5}$$

"N" still represents the total number of testing images and $N_{e \leq j}$ is the number of correctly predicted images where the error is not more than j years old.

In addition, [45] introduced an evaluation metric denoted as one-off accuracy. This metric evaluates age group classifiers, and the output is the error of one age category. The one-off accuracy equation is as follows:

$$1 - \text{off} = \frac{n_o}{N_x} * 100\% \tag{6}$$

where n_o is the number of classified images with one class error N_x is the total number of testing images. In Equations (4)–(6), we aim to obtain a higher value that indicates how many images were classified correctly.

7. Age Estimation Literature Review

In this section, we review several recently published papers that attempt to solve the problem of age estimation using various techniques. The review consists first of a brief introduction to the paper that has been surveyed, followed by the proposed method. The second part of the review details the datasets the authors used and the result. Finally, an analysis of the strengths and weaknesses concludes the review of each method. This section is divided into two subsections: (1) Handcrafted and (2) Transfer learning-based models. The handcrafted subsection surveys the papers utilising algorithms built from scratch to

estimate facial age. On the other hand, the transfer learning-based subsection lists the papers that use pre-trained architectures to predict facial age.

7.1. Handcrafted Methods

In 2015, a paper by [5] presented an age estimation model based on a modified version of the support vector machine algorithm [46]. The proposed method uses Viola and Jones [47] to detect and extract the subject's face. The detected face is aligned using a 68-landmarks facial detector [48].

The facial features are extracted using a local binary pattern (LBP) operator. The authors use the LBP of studies [49–51] with the related four-patch LBP codes of [52]. The authors claim that these LBP codes were used due to their robustness with various face recognition problems and are computationally inexpensive.

The pre-processed images are fed to a modified linear support vector machine (SVM) for classification. The classifier is equipped with a dropout layer to reduce the model's complexity and overfitting. The authors experiment with two different dropout rates. The authors set the dropout rate to 80%. The training and testing were conducted on the Adience dataset, consisting of more than 20,000 samples taken in unconstrained conditions.

When the model was tested on the Adience dataset, the authors reported a classification accuracy of 45.1% and a one-off accuracy of 79.5%. Additionally, the authors reported a classification accuracy of 66.6% when the model was validated on the Gallagher dataset [23].

This method retains several advantages and disadvantages. The first advantage is that the model is trained and tested on images taken in unrestrained conditions. In real-life applications, the quality of images, head poses, and facial accessories are not controlled and are usually very noisy. The second advantage is that the model is not computationally expensive, and the custom dropout layer helped to generalise the model to new samples.

In contrast, the main disadvantage lies in the feature extraction phase. As the facial features are extracted manually, several important discriminative features are left behind, which causes the model not to learn all the essential ageing features, thus resulting in low accuracy. The second weakness is the distribution of the Adience dataset samples. The dataset is missing samples from specific age groups, such as those between 20 and 25 years old or individuals between 43 and 48 years old. This distribution of samples limits the model from being tested on samples of subjects within the abovementioned age groups.

The authors of the previous study experimented with convolutional neural networks and reported the findings in [5]. The authors opted for a smaller and much simpler network design than much larger architectures such as [53–57]. The classifier consists of three convolutional layers followed by two fully connected layers. The final output layer maps to either eight age classes or two gender classes. Before feeding the network with the images for training, the samples are first rescaled to 256×256, and a crop of 227×227 is fed to the network's input layer. The first hidden convolutional layer consists of 96 filters with a kernel size of 7×7. This layer is activated using the rectified linear operator (ReLU) function followed by a max-pooling layer with a pool size of 3×3 and two-pixel strides. The authors then add a local response normalisation layer. The second hidden layer consists of 256 filters with a kernel size of $96 \times 5 \times 5$, and it is activated using the ReLU activation function. A max-pooling layer and a local response normalisation layer are added. The final hidden layer contains 384 filters with a kernel size of $256 \times 3 \times 3$.

The next portion of the classifier is the fully connected module consisting of two fully connected layers, each with 512 neurons. Each layer is followed by a dropout layer and activated using ReLU. The layer that maps to either the age groups or the gender class is a soft-max layer that assigns a probability to each gender or age class. The authors use the Adience dataset to train and test the model. The testing is performed using K-Fold validation with five splits.

The authors recorded a higher accuracy of 50.7% compared to their previous attempt in [4] with SVM.

Although the authors of this study attempted to extract the facial features using convolutional neural networks, which are proven to be much better than handcrafted models, the presented results remain far from perfect for several reasons. By observing the confusion matrix provided by the authors, we notice that the model is capable of identifying samples of subjects in the 0–2, 4–6, 60+, and 8–13 age groups accurately; however, we see lots of misclassifications of samples taken from the 15–20, 38–43, and 48–53 age groups. This observation might indicate that the model did not learn certain discriminative features that would allow it to learn the key features that make these age groups different.

Despite the observed weakness, this model's significant advantage is that it can classify images taken in real-life conditions of various qualities and illuminations. This study confirms that convolutional neural networks outperform traditional machine learning techniques in which the features are manually extracted.

In a study by [58], the authors introduced a built-from-scratch network trained on facial images of celebrities captured in unconstrained conditions. This method consists of four steps. First, a face detection method, introduced by the authors and known as the deep pyramid deformable parts model, is employed to locate a subject's face in an image. The second step is face alignment, which is carried out using the dlib C++ library. The third step is feature extraction, which uses a ten layers CNN network. The final step is estimating facial age, and for this step, a custom-built three-layer neural network trained on Gaussian loss is employed. The input layer of the network takes an input vector of size 320. The first hidden layer consists of 160 units, while the second layer consists of 32 units. Every layer is activated using the parametric rectified linear unit (PReLU) function. The proposed network is generalised by adding several dropout layers after each neural layer with rates of 40%, 30%, and 20% for the input, first, and second hidden layers, respectively. The network has been trained on the CASIA-WebFace dataset and cross-validated on ICCV 2015 ChaLearn challenge dataset. The authors reported an error rate of 0.373 on the test and validation portions of the ChaLearn dataset.

This study has shown that using Gaussian loss improves age prediction performance. Based on the reported results, the proposed method has been robust to poses and resolutions compared to similar methods. However, this method cannot appropriately handle images of extreme illuminations and poses. In addition, the authors insisted that the lack of training samples of individuals older than 70 caused the model to misclassify individuals in that age group.

The authors of [3] attempted to tackle the issue of poor-quality images by introducing a pre-trained super-resolution GAN (SRGAN) [59] layer into the pre-processing stage. The authors introduced a custom-built CNN classifier that can distinguish between six age classes. The proposed method begins by pre-processing the images through face detection, alignment, and resizing. The next pre-processing stage is passing the image to a pre-trained SRGAN generator which reconstructs a higher resolution equivalent of the input image. The image is then fed to a custom CNN classifier with two hidden layers. The first layer consists of 96 filters, while the second consists of 128 filters. The authors used the UTKFace dataset for training and testing and reported an accuracy of 72%. Based on the experiments, the introduction of SRGAN has improved the model's performance. However, it is evident from the confusion matrices that the lack of enough samples and data disparity contributed to the decrease in accuracy.

In another study by [60], the authors introduced a concept known as Deep Expectation (DEX) to estimate apparent age. The network was inspired by the VGG-16 network design, and it was trained to treat age estimation as a regression problem. The authors utilised the IMDB-WIKI dataset for training and testing. Out of the 500,000+ samples, 260,282 images were used for training, while the remaining samples were dedicated to testing. According to the authors, the experiment showed that the DEX network improved the rate of age prediction compared to traditional regression. The authors of this study reported an MAE of 3.2 years. One of the limitations mentioned in this paper is the demand for higher computational power to carry out the face identification process. In addition, it

was mentioned in this paper that extreme illumination and various poses were the main contributors to failed predictions.

7.2. Transfer Learning-Based Methods

Pre-trained models such as VGG19 or ResNet50 have solved various machine learning problems. The interest in utilising pre-trained models lies in their ability to produce better accuracy without needing a lot of labelled training samples. In age estimation, one of the common issues is the availability of adequate training samples. Several studies suggested using pre-trained models to tackle the problem of insufficient data.

The authors of [61] introduced a multi-stage age and gender estimation model that uses a pre-trained VGG19 model. The first component of the proposed method is a saliency detection network capable of extracting regions of interest, which, in the case of this problem, is a subject's face. The second component is an age estimation model, a pre-trained VGG19 network.

The saliency detection network is a deep encoder–decoder segmentation network that has proven successful in semantic segmentation, hole filling, and computer vision tasks that require segmentation.

The age and gender estimators are combined into a modified VGG-19 network, where the last fully connected layers are replaced with average pooling layers, and two extra separated layers are added to predict age and gender. The output of the last convolutional layer is encoded from $14 \times 14 \times 512$ to $14 \times 14 \times 1$ using a 1×1 convolutional layer. A max-pooling layer is then used to encode the feature map to $7 \times 7 \times 1$.

These two separate layers are introduced to reduce the number of parameters and complete the linear combination of the 512 feature maps. The output of the last convolutional layer is flattened to a 49×1 vector, and each element in the vector is a 32×32 features region. The 49×1 vector is then expanded to 2058×1 to represent the features of the reach region using the local region interaction (LRI) operation. The introduced LRI ignores the interactions among local regions in the same row, which better represents the features by eliminating redundant information.

The study's authors treat age as a continuous variable; therefore, they consider this problem a regression task in which the final layer outputs a single value instead of a class probability.

The authors used the FG-NET, Adience, and CACD datasets to train and test the proposed method. The age label in the CACD dataset was estimated using the year information acquired when the dataset was collected through a web search. All three datasets were equally divided into 80% training and 20% testing data with a mini-batch stochastic gradient descent of patch size 224×224 and a batch size of 10. The learning rate of the network is 2.5×10^{-4}, whereas the momentum is 0.9, and the number of epochs is 200.

The authors reported an MAE of 1.84 years, mainly due to the implementation of the saliency detection network, which ensures that only faces are extracted.

Compared to the other models, this method's main advantage is adding a subject-background segregation mechanism to better pre-process the input images. Ensuring that only the pixels representing one's face are fed to the learning algorithm helps ease the training process.

Despite reporting a relatively low MAE, this method suffers from several issues. Firstly, the authors insisted that this method performs only on images that contain a single face due to the lack of a face detection component. This issue limits the model from performing in real-life scenarios where more than one face might exist in a single image. In addition, this method does not consider non-frontal or misaligned faces since it only extracts regions of interest (faces) without aligning or rotating them. The lack of a pre-processing stage to correct alignment and rotation is a limitation since images acquired from different datasets or taken in real-life scenarios are of various poses and angles.

The authors in [62] proposed a regression-based method. The first step of this method is detecting and aligning the faces in a given image, followed by feature extraction from the input images. For this step, the authors run different experiments using several pre-trained architectures. The first network the authors experiment with is the VGG16, which consists of 13 convolutional layers and 3 fully connected layers. The second architecture is VGG19, which contains more convolutional layers. The third architecture is ResNet50, which is an architecture that combines convolutional layers and residual connections. Next is the InceptionV3, which consists of different convolutional sequences that perform separately on their given input. The last network architecture is the Xception, which uses depthwise separable convolutional layers convolving separately on each input channel.

Since all the above-mentioned architectures are pre-trained on ImageNet and have a softmax output of 1000 classes, the authors had to fine-tune the model to fit the nature of the age estimation problem. The fine-tuning of the model was carried out by replacing the last layer with a one-layer regression neural network to learn an age regression function from the extracted features. A dropout layer is added before the last output layer is added as an extra measure to avoid overfitting, which might occur due to the small number of training samples. In addition, early stopping was implemented to end the training once the accuracy stopped improving after ten epochs. Each network uses the Adam optimiser and the mean squared error loss function with a learning rate of 0.001. In order to save time and computational power, the transfer learning process is segregated into two steps.

The first step is pre-training, in which the models are randomly initialised by a related task that owns enough labelled data. In this case, the networks were trained on the ImageNet dataset, which contains 14 million images.

The second step is fine-tuning parameters to fit the nature of the new problem, which is age estimation.

Three datasets were used in this study to train and test the pre-trained networks. The first dataset is MORPH, divided into 80% training and 20% testing. In training, the dataset is further divided into 90% training and 10% validation. FACES is the second dataset used in this method; however, it was mainly used to evaluate the performance of each network. The third and final dataset used in this study is the FGNET dataset. However, the train-test split was adjusted to 90% training and 10% testing because of the low number of samples.

The VGG-16 and VGG-19 models achieved the best MAE of 4.43 and 4.84, respectively, when 0% of the hidden layers were unfrozen.

InceptionV3 and Xception were the worst-performing models, scoring an MAE above ten years when 0% of their hidden layers were unfrozen. However, InceptionV3 and Xception scored the best MAE of 2.47 and 2.53, respectively, the lowest MAE compared with the other three models when 100% of the layers were frozen at training. Although Xception and InceptionV3 are among the best-performing networks in this study, the authors insisted that it does not work well with Gaussian noise; therefore, the mean absolute error increases when the variance increases. Similar to previous studies, the findings in this paper have shown that pre-trained models possess many advantages. Pre-trained models similar to the one used in this study are usually faster. These models require less computational power to train since only fine-tuning the hyperparameters or modifying the layers is needed.

The main limitation the authors concluded their study with is the drop in the accuracy of all models when tested on images of subjects of mixed ethnicities or cross-genders. In addition, all the networks could not perform well on images in which the face is not frontal or the faces are of different poses and facial expressions. The authors also noticed that external circumstances, such as lighting and illumination, increased the mean absolute error of all five networks regardless of whether the layers were frozen.

Another method that utilises pre-trained models was proposed by [63] to estimate age and gender from facial images. Their study introduced three novel methods that use an architecture similar to the VGG16 design pre-trained on the facial recognition task. The first method is denoted as pure per-year classification; however, it is referred to as 0/1-CAE. In this method, the authors treat every age value as a single class and every label as a one-hot

1D-vector. The size of such a vector depends on the number of classes. The authors chose the number of classes to be 100 (between 0 and 99 years old).

The second method is called pure regression, which predicts the value y of an image based on input x. In this case, the regression model maps facial images to one of the corresponding labels. The researchers referred to this method as RVAE in this study.

The third method, termed soft classification, is treated as a mixture of discrete classification and linear regression. This method transforms ages into a vector size similar to the number of classes, but the classifier is not binary. Instead, a Gaussian distribution centred at the target age encodes the values in the vector.

This method is referred to as the LDAE in the remainder of the study. As for the CNN architecture, the researchers decided to take a slightly different approach and remove the fully connected portion of the network. The authors emphasised that the network's ability to classify ages comes from the convolutional layers and not the fully connected layers; therefore, these layers were wholly taken out. The authors present four different versions of their network, denoted as fast_CNN_2, fast_CNN_4, fast_CNN_6, and fast_CNN_8. The number that follows "CNN" in the network labels represents the number of convolutional layers of each network. For example, fast_CNN_2 consists of two convolutional layers, while fast_CNN_4 consists of four convolutional layers. The proposed fast_CNN networks follow the VGG16 design in several aspects. Firstly, all the convolutional layers consist of square feature maps with a kernel size similar to VGG16. Secondly, max-pooling layers are designed to follow the same design that was introduced in the original VGG16 architecture. Finally, the network layers are activated using the rectified linear activation (ReLU) function. The network contains a dropout layer between each convolutional layer with a rate of 0.5 and a batch normalisation layer as preventive measures to prevent overfitting. All variations in the fast_CNN network expect an RGB input facial image of size 64×64 and an output of either the exact age or the age class.

Two datasets were used in this study. IMDB-WIKI was used to train the gender detection and age estimation models. The second dataset was used only for testing, denoted as private balanced gender age (referred to as PBGA). This dataset was privately collected, and the authors claim it is more balanced regarding genders and samples per age group than various benchmark public datasets. The dataset contains 3540 images of subjects between 12 and 70 years old, where each age group consists of 30 images of males and 30 images of females.

For the 0/1-CAE and LDAE methods, two approaches were used to predict the age of an image. First, the class of the neuron with the highest activation was selected as the estimated age. This approach is denoted as ArgMax. Second, age is predicted based on the expected value of every output neuron.

Using the LDAE method with the expected-value approach, the lowest MAE was recorded at 6.05 years. The highest MAE was recorded from the pure regression RAVE method at 7.19 years. These results demonstrate how regression and a classification model can complement each other to produce as little error as possible.

The second assessment was based on the depth of the proposed fast_CNN, where the best scoring network for age estimation was the fast_CNN_6 with an MAE of 5.95 years. The worst-performing network was the fast_CNN_2, with an MAE of 6.65 years. The accuracy did not improve when the network was deeper than eight layers because fast_CNN_8 scored a classification score of 92.3% but a better MAE of 5.89 years.

The third assessment was conducted to determine whether a network would perform better if pre-trained on a complex task before age estimation. It turns out that the best performance is obtained when a network is pre-trained on face recognition and training is based on a single task. The MAE produced by this architecture was 5.96 years. The highest MAE (6.05) was obtained from a network trained on a single task without pre-training.

The final assessment was conducted to find the best network design, and VGG-19, VGG-16, and ResNet-50 were compared. The best-performing age estimation model was based on the VGG16 design with an MAE of 4.26, which is lower than the other

two designs. VGG16 was pre-trained on face recognition before performing age estimation tasks. The authors concluded their study by stating that LDAE is the best-performing network architecture when using pre-trained face recognition. The network was tested on popular datasets such as MORPH-II and FG-NET and scored an MAE of 2.99 and 2.84, respectively.

Besides transfer learning, it is evident from this study that combining classification and regression helps produce better results. In addition, this study demonstrates that fully connected layers do not affect the accuracy of age estimation; instead, it reduces the computation complexity. One of the main limitations of this study is that the models were not tested or trained on subjects under 12 years old since the youngest group in the testing dataset is 12 years old.

In a study by [7], the authors attempted to use transfer learning to build an age group classifier based on the Adience dataset. The age classifier, denoted as VGG-face, was pre-trained on facial recognition, making this model capable of extracting complex facial features from facial images taken in various circumstances.

The classifier consists of 11 layers in total, where eight of them are convolutional while the remaining three are fully connected layers. Each convolutional layer is activated using the rectified linear activation function (ReLU), followed by max-pooling and batch normalisation layers. In addition, padding and strides are added to each convolutional layer. The number of filters in the first layer is 64, which increases by the multiple of two in every subsequent layer resulting in the final layer having 512 filters. The fully connected portion of the classifier consists of two layers, each with 4096 neurons, and a dropout layer follows each with a rate of 0.5. The last output layer consists of N number of outputs where N is the number of classes.

The authors amended the network by changing the design of the fully connected layers. They changed the number of neurons from 4069 to 5000 in the second and third hidden layers while maintaining the number of neurons in the first hidden layer. As a preventive measure to avoid overfitting, a dropout layer is added after each layer with a rate of 0.5. The final output layer maps to eight age classes, which are: 0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, and 60+.

The network's training begins by firstly rescaling the images to 256 $\times$ 256, and then 224 $\times$ 224 patches are fed into the classifier. Stochastic gradient descent with mini-batches of 256 and a momentum of 0.9 is used for optimisation and a weight decay of 0.001. A dropout layer with a rate of 0.6 is used for regularisation and overfitting prevention. The learning rate is set to 0.1 and is decreased when the validation accuracy is not improving. The network weights are determined using a Gaussian distribution with a zero mean and a 10–2 standard deviation.

The classifier produced an overall accuracy of 59.90%, with the highest accuracy score obtained for the 0–2 age class at 93.17%. The second highest score was 86.17% for the 25–32 age class. In contrast, the lowest accuracy was 8.8% for the 38–43 age class. The second-lowest score was 24.23%, obtained for the 15–12 age class. In addition, the authors presented a one-off accuracy of 90.57%, which is higher than similar studies such as [4] and [5].

The main drawback of this method is that the model is computationally expensive and requires a lot of resources and time to train. In addition, it is evident from the confusion matrix that the model struggles to classify samples of subjects in the 38–43 years old category.

We theorise that the misclassification occurred because of the similarity in features between subjects in this class and subjects from the 25–32-year-old class. Some of the presented training samples are highly degraded, and essential ageing features were lost. Despite these limitations, the model has shown robustness in dealing with several test samples captured in real-life scenarios.

In a study by [64], the authors employed several pre-trained models based on the VGG16, ResNet50, and SENet50 [65] architectures. In addition, the authors employed K-fold cross-validation as a countermeasure against overfitting. For each network, the

authors used a pre-trained weight denoted as VGGFace, which has been trained to detect faces in images taken in unrestrained conditions. The fine-tuning of the networks consisted of adding five new layers to the existing architecture. The first layer is a flattening layer that converts the feature map into a vector. The three subsequent layers are fully connected; the first two dense layers consist of 1024 neurons, while the third dense layer consists of 512 neurons. The last dense layer maps to the new output layer, a softmax layer with eight output classes. Each layer is activated using the ReLU activation function. Each model was trained for 100 epochs in which each fold was trained for 20 epochs. The networks had a batch size of 32 and were optimised using the Adam optimiser with a learning rate of 0.001.

The authors in this study used the UTKFace dataset for training and testing and divided it into eight age groups: 0–2, 4–6, 8–12, 15–20, 25–32, 38–43, 48–53, and 60+. The number of folds for the K-fold validation split is five, where each fold allocates 80% of the images to training while the remaining 20% is used for testing. Out of the total number of images, the training set had more than 9300 samples, the testing set had more than 2300 samples, and the validation set had around 330 images.

The authors reported the average accuracy taken over the five folds for each network. The VGG16 network produced an average accuracy of 83.76%, while the ResNet50 network produced 88.03%. The SESNet50 produced the lowest average accuracy of 74.43%. In addition to the accuracies per fold, the authors presented the accuracy of each network compared to similar methods. The best-performing network is the ResNet50, with a testing accuracy of 71.84%, while the worst-performing model is the SENet50, with a testing accuracy of 61.96%.

The proposed method had successfully overcome the issue of overfitting and data disparity by introducing K-fold cross-validation. Despite the reported high accuracies, one of the drawbacks to this method is the age gap between each age class. Several samples were removed from training and testing due to how the images were discretised. In addition, the models were not evaluated on a different dataset to confirm the accuracy of the three models on unseen data.

In a study by [66], the authors attempted to find the optimum number of classes and the ideal age range between classes by performing several experiments on four pre-trained models with different age classes. The authors used the VGG16, ResNet18, GoogLeNet, and AlexNet. Each model has already been trained on the ImageNet dataset, which consists of more than 14 million labelled images of different objects. The authors fine-tuned each model by replacing the last output layer with a layer that maps to N number of age classes, which changes every experiment.

The authors used both FG-NET and MORPH datasets for training and testing, with 80% of the images dedicated to training while the remaining 20% are used for testing. Therefore, 44,909 images were used for training and 11,227 for testing.

This study's first set of tests focused mainly on finding the optimum number of age classes. The first experiment divided the age groups into six classes, each with a gap of five years: 0–5, 6–10, 11–19, 20–29, 30–39, and 40–77. The best-performing network was the GoogLeNet, with an accuracy of 74%, while the worst-performing network was the VGG16, scoring 68%. The AlexNet and ResNet18 networks scored 69% and 72%, respectively. During the second experiment, the number of age groups was reduced to 5, and the age gap increased to 10. The five age classes were: 0–9, 10–19, 20–29, 30–39, and 40–77. The GoogLeNet network was the best-performing architecture, with an accuracy of 85%, while VGG16 produced the lowest accuracy of 79%. ResNet 18 and AlexNet scored accuracies of 83% and 81%, respectively. The third trial had the number of classes reduced to 4, and the age gap increased to 15. The samples were divided into 0–14, 15–29, 30–44, and 45–77 classes. Similar to the previous two experiments, the GoogLeNet model scored the highest accuracy of 87%, while the ResNet18 produced the second-highest accuracy of 86%. The AlexNet model produced an accuracy of 83%, and the VGG16 model produced 81%.

The fourth and final experiment segregated the images into three classes: 0–19, 20–39, and 40–61, with a gap of 20 years in between. Like the previous experiments, the GoogLeNet

model produced the highest accuracy of 89%, followed by the ResNet18 model with an accuracy of 88%. The worst-performing models were the AlexNet and VGG16, with 87% and 86% accuracy, respectively.

The second set of experiments focused more on changing the age gap than the number of classes. The number of classes was fixed to two while the age gaps changed during each experiment. The first test had a vast age gap of 30 years. The first class was 0–29 years old, while the second was 30–77 years old. All four models produced accuracies of more than 90%, except for the VGG16 network, which produced an accuracy of 90%. The GoogLeNet network produced an accuracy of 94%, followed by the ResNet18 model with an accuracy of 93%. The second worst-performing network was the AlexNet, with an accuracy of 92%. The second experiment worked with age groups between 0 and 15 years old, where one class had samples of subjects between 0 and 5 years old, while the other class had samples of subjects between 10 and 15. All networks produced more than 90% accuracy, with an accuracy of 99% produced by GoogLeNet. The ResNet18 model produced the second-highest accuracy of 98%, while the AlexNet model produced 97%. The worst-performing network remains the VGG16, with an accuracy of 94%.

This study highlighted a crucial aspect that several pieces of works of literature have ignored, which is the different ways of classifying age groups. The study has shown that the overall accuracy tends to degrade with the introduction of more classes. In addition, the age gap between each class plays an essential role in defining the age estimator's accuracy. However, one of the drawbacks is the vast age gap of 30 years that the authors proposed. A wide age gap defeats the purpose of automatically estimating age from facial images since most systems would require a more specific model instead of a model trained on only two age groups.

In a study by [67], the authors attempted to solve the problem of low-quality images by reconstructing the original images to a better-quality equivalent. This objective is achieved using a conditional generative adversarial network (CGAN) that rebuilds low-resolution facial images before processing.

The authors then used pre-trained models such as ResNet, VGG, and DEX to estimate the age from the input reconstructed images. The datasets used to train and test the age estimation model were the PAL, MORPH, and FG-NET databases and the authors reported an MAE of 8.3 as their best result. Although the proposed method confirmed that other architectures, such as GANs, can be used to improve the rate of correctly estimating facial age, one of the main drawbacks of this method is the increased processing time caused by the GAN component.

More recently, a new model by [68] known as ShuffleNet was introduced. The proposed model is based on the mixed attention mechanism (MA-SFV2). The main highlight of the proposed model is that it transforms the output layer and merges classification and regression age estimation concepts. In addition, the authors claim that the model focuses only on the critical features extracted during the pre-processing stage and data augmentation. The proposed method consists of several image pre-processing steps to ease the training process and a data augmentation step such as filtering, sharpening, and stretching to overcome overfitting. The authors tested and trained their model on the MORPH-II and FG-NET datasets and were able to achieve an MAE of 2.68.

In [6], the authors introduced the concept of gender-based age classification, in which each input image is filtered by gender before estimating the age. The proposed method consists of three components: (1) Gender classifier, (2) Males-only age classifier, and (3) Females-only age classifier.

The proposed method begins with a pre-processing step in which all images are normalised and resized to a constant size. Next, a pre-trained VGG16 is modified and trained to estimate age class from the pre-processed images. During the training phase, each VGG16 network is trained on a group of images filtered by gender. Therefore, the males-only age classifier is trained on images of males, while the females-only age classifier is trained on images of females. The gender classifier is the first entry point when the

system is in use and is responsible for loading the appropriate age classifier based on the gender label. The authors used the UTKFace dataset to train their age classifiers and a gender dataset from Kaggle to train the gender classifier. The authors reported an accuracy of 80.5%; however, the main drawback of this system is that it does not consider non-binary individuals. Table 4 presents a comparison of all the review methods.

Table 4. Comparison of different proposed age estimation methods.

Method	Dataset	MAE	Accuracy	1-Off Accuracy	Error Rate
[4]	Adience	N/A	45.1%	79.5%	N/A
[3]	UTKFace	N/A	72.0%	N/A	N/A
[5]	Adience	N/A	50.7%	80.0%	N/A
[58]	ChaLearn	N/A	N/A	N/A	0.373
[60]	IMDB-WIKI	3.2	N/A	N/A	N/A
[61]	FGNET, Adience, CACD	1.84	N/A	N/A	N/A
[62]	MORPH, FGNET, FACES	4.43	N/A	N/A	N/A
[63]	IMDB-WIKI	5.96	N/A	N/A	N/A
[7]	Adience	N/A	59.9%	N/A	N/A
[64]	UTKFace	N/A	88.03%	N/A	N/A
[66]	FGNET, MORPH	N/A	87.0%	N/A	N/A
[67]	FGNET, MORPH, PAL	8.3	N/A	N/A	N/A
[68]	MORPH, FGNET	2.68	N/A	N/A	N/A
[6]	UTKFace	N/A	80.5%	N/A	N/A

8. Discussion

Our findings reveal that researchers mostly prefer using transfer learning over building a new model from scratch, and there are several reasons why pre-trained models are preferred. Firstly, pre-trained models are usually faster to train because they only require hyperparameters fine-tuning and, in some cases freezing or unfreezing the hidden layer. Moreover, based on the task, pre-trained models require modifying the output layer to produce the required outputs. For example, a regression task using a VGG16 network would require an output layer with one neuron, while a classification task would require an output layer with N neurons where N is the number of classes. The second benefit of choosing a pre-trained model over a custom model is that pre-trained models are initially trained on comparatively complex tasks with millions of images.

Several pre-trained models such as VGG16 or ResNet50 have been trained on more complex tasks using more enormous datasets such as ImageNet or VGGFaces. Therefore, when these models are employed for a less complex task such as age estimation, they require little to no modification to solve the given problem.

Although age estimation is flexible and can be treated as either a regression, a classification or a hybrid problem, most studies treat it as a regression problem to produce an actual age value. As age label is treated as a continuous variable, building a regression model could be more straightforward since deciding on the number of age classes and the age gap between classes becomes unnecessary. In addition, comparing the accuracy of age classifiers can become an imprecise process because there is no standard way to create age classes, and different models will have various outputs. In comparison, evaluating the performance of various regression models is more realistic since the age label is constant, and the models will always produce a single output; therefore, the accuracy will mainly be affected by the proposed architecture and the data quality.

The common gaps discovered in most of the reviewed methods lie mainly in the data aspect of each study. The first issue is that the training images in most benchmark datasets, such as the Adience dataset, are of low resolution in which distinctive ageing features such as wrinkles or skin texture are imperceptible, resulting in a drop in training performance. The second issue, which is more complex to solve, is the variation in ageing patterns, and it may depend on the subject's lifestyle, gender, or ethnicity. Since the lifestyle or ethnicity of a person influences facial ageing, it is not easy to collect enough data covering all the unique ageing patterns. Therefore, regardless of the accuracy or complexity of existing age estimation models, they are still far from perfect to use in real-life situations. Another critical gap has been primarily observed in classification-based methods in determining

suitable age classes with the appropriate age gap between each class. Based on the findings by [68], the more age classes exist, the worse the performance of a model is and vice versa. However, having fewer classes with more large age gaps will decrease the model's ability to produce more specific predictions. For example, if a model has only two age classes: 0–31 and 32–77, making predictions for a specific age group such as teenagers or adolescents is impossible since the model will always produce a more general prediction of 0–31.

Another common issue among most age estimation methods is data disparity, which usually leads to overfitting. Most current benchmark datasets are imbalanced in terms of age group, gender, or ethnicity, where there are more samples of a specific group of subjects than others. Usually, this issue is prevented by either adding more samples to the dataset or reducing the number of samples of the majority class. These two processes are denoted as oversampling and undersampling, respectively. Additionally, a different approach could be generating artificial images using models such as generative adversarial networks (GANs) to balance an existing dataset.

Based on our observations, we anticipate that future research into age estimation will not depend mainly on optimising the model due to the availability of numerous pre-trained networks. Instead, more focus will be on building data-centric age estimation systems. In addition, we foresee that with the advancement of architectures such as generative adversarial networks (GANs) where we can control the features, researchers will have the ability to train entire models on artificially synthesised images.

9. Conclusions

Facial age estimation is a hot area of research, yet a reasonably complex task for various reasons, such as insufficient training data or the lack of a model that fits all the different ageing patterns. This study examined the definition of age estimation from a machine learning perspective, the different methods to estimate age from facial images and the details of several benchmark datasets. Moreover, we presented several existing studies that have attempted to solve the problem of age estimation in addition to the pros and cons of each method. We conclude the study by discussing the common existing gaps and the current direction of research.

Author Contributions: K.E. reviewed the literature, wrote the sections of this manuscript, and carried out the analysis of the findings. V.R. and P.T. reviewed the grammar, structure, and content of this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Swinburne University of Technology (Sarawak Campus) for providing the necessary resources to carry out this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Coleman, S.R.; Grover, B.R. The anatomy of the aging face: Volume loss and changes in 3-dimensional topography. *Aesthetic Surg. J.* **2006**, *26*, S4–S9. [CrossRef] [PubMed]
2. Al-Shannaq, A.S.; Elrefaei, L.A. Comprehensive Analysis of the Literature for Age Estimation From Facial Images. *IEEE Access* **2019**, *7*, 93229–93249. [CrossRef]
3. Elkarazle, K.; Raman, V.; Then, P. Towards Accuracy Enhancement of Age Group Classification Using Generative Adversarial Networks. *J. Integr. Des. Process Sci.* **2022**, *25*, 8–24. [CrossRef]
4. Eidinger, E.; Enbar, R.; Hassner, T. Age and Gender Estimation of Unfiltered Faces. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 2170–2179. [CrossRef]

5. Levi, G.; Hassncer, T. Age and gender classification using convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 34–42. [CrossRef]
6. Raman, V.; Elkarazle, K.; Then, P. Gender-specific Facial Age Group Classification Using Deep Learning. *Intell. Autom. Soft Comput.* **2022**, *34*, 105–118. [CrossRef]
7. Qawaqneh, Z.; Abumallouh, A.; Barkana, B. Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model. *arXiv* **2017**, arXiv:1709.01664.
8. Rasmus, R.; Radu, T.; Luc Van, G. DEX: Deep EXpectation of apparent age from a single image. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015.
9. Softopia Japan Foundation. Human and Object Interaction Processing (HOIP) Face Database. Available online: http://www.hoip.jp/ (accessed on 18 September 2022).
10. Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; Hua, G. Ordinal Regression with Multiple Output CNN for Age. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 27–30 June 2016; pp. 4920–4928. [CrossRef]
11. Chen, B.-C.; Chen, C.-S.; Hsu, W.H. Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014.
12. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning Face Representation from Scratch. *arXiv* **2014**, arXiv:Abs/1411.7923.
13. Ricanek, K.; Tesafaye, T. MORPH: A longitudinal image database of normal adult age-progression. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 341–345. [CrossRef]
14. Afifi, M.; Abdelhamed, A. AFIF4: Deep Gender Classification based on AdaBoost-based Fusion of Isolated Facial Features and Foggy Faces. *J. Vis. Commun. Image Represent.* **2019**, *62*, 77–86. [CrossRef]
15. Zhang, Y.; Liu, L.; Li, C.; Loy, C.C. Quantifying Facial Age by Posterior of Age Comparisons. *arXiv* **2017**, arXiv:1708.09687.
16. Hassner, T.; Harel, S.; Paz, E.; Enbar, R. Effective face frontalization in unconstrained images. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015.
17. Zhang, Z.; Song, Y.; Qi, H. Age Progression/Regression by Conditional Adversarial Autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
18. Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; Zafeiriou, S. Agedb: The first manually collected, in-the-wild age database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, Honolulu, HI, USA, 21–26 July 2017; p. 5.
19. Han, H.; Jain, A.K.; Wang, F.; Shan, S.; Chen, X. Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2597–2609. [CrossRef]
20. Huang, G.B.; Jain, V.; Learned-Miller, E. Unsupervised Joint Alignment of Complex Images. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio De Janeiro, Brazil, 14–21 October 2007; pp. 1–8. [CrossRef]
21. Delac, K.; Grgic, M.; Grgic, S. Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set. *Int. J. Imaging Syst. Technol.* **2005**, *15*, 252–260. [CrossRef]
22. Fu, Y.; Guo, G.; Huang, T.S. Age Synthesis and Estimation via Faces: A Survey. IEEE Trans. *Pattern Anal. Mach. Intell.* **2010**, *32*, 1955–1976. [CrossRef]
23. Gallagher, A.C.; Chen, T. Understanding images of groups of people. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 256–263.
24. Bastanfard, A.; Nik, M.A.; Dehshibi, M.M. Iranian Face Database with age, pose and expression. In Proceedings of the 2007 International Conference on Machine Vision, Islamabad, Pakistan, 28–29 December 2007; pp. 50–55. [CrossRef]
25. Fu, Y.; Hospedales, T.M.; Xiang, T.; Gong, S.; Yao, Y. Interestingness Prediction by Robust Learning to Rank. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014. [CrossRef]
26. Mehrotra, R.; Namuduri, K.; Ranganathan, N. Gabor filter-based edge detection. *Pattern Recognit.* **1992**, *25*, 1479–1494. [CrossRef]
27. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 886–893. [CrossRef]
28. Kanopoulos, N.; Vasanthavada, N.; Baker, R. Design of an image edge detection filter using the Sobel operator. *IEEE J. SolidState Circuits* **1988**, *23*, 358–367. [CrossRef]
29. Pietikäinen, M. Local Binary Patterns. *Scholarpedia* **2010**, *5*, 9775. [CrossRef]
30. Meyers, E.; Wolf, L. Using Biologically Inspired Features for Face Processing. *Int. J. Comput. Vis.* **2007**, *76*, 93–104. [CrossRef]
31. Unnikrishnan, A.; Ajesh, F.; Kizhakkethottam, J.J. Texture-based Estimation of Age and Gender from Wild Conditions. *Procedia Technol.* **2016**, *24*, 1349–1357. [CrossRef]
32. Hayashi, J.; Yasumoto, M.; Ito, H.; Koshimizu, H. Age and gender estimation based on wrinkle texture and color of facial images. In Proceedings of the 2002 International Conference on Pattern Recognition, Quebec City, QC, Canada, 11–15 August 2022; Volume 1, pp. 405–408. [CrossRef]
33. Luu, K.; Ricanek, K.; Bui, T.D.; Suen, C.Y. Age estimation using Active Appearance Models and Support Vector Machine regression. In Proceedings of the IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, Washington, DC, USA, 28–30 September 2009; pp. 1–5. [CrossRef]

34. Kohli, P. Age Estimation Using Active Appearance Models and Ensemble of Classifiers with Dissimilarity-Based Classification. In *Advanced Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 327–334.
35. Geng, X.; Zhou, Z.H.; Zhang, Y.; Li, G.; Dai, H. Learning from Facial Aging Patterns for Automatic Age Estimation. In Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, 23–27 October 2006; pp. 307–316.
36. Angulu, R.; Tapamo, J.-R.; Adewumi, A.O. Age estimation via face images: A survey. *EURASIP J. Image Video Process.* **2018**, *2018*. [CrossRef]
37. Geng, X.; Zhou, Z.-H.; Smith-Miles, K. Automatic Age Estimation Based on Facial Aging Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2234–2240. [CrossRef]
38. Geng, X.; Smith-Miles, K.; Zhou, Z.H. Facial Age Estimation by Nonlinear Aging Pattern Subspace. In Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, BC, Canada, 26–31 October 2008; Association for Computing Machinery: New York, NY, USA; pp. 721–724.
39. Fu, Y.; Liu, M.; Huang, T.S. Conformal Embedding Analysis with Local Graph Modeling on the Unit Hypersphere. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–6. [CrossRef]
40. Wang, R.; Nie, F.; Hong, R.; Chang, X.; Yang, X.; Yu, W. Fast and Orthogonal Locality Preserving Projections for Dimensionality Reduction. *IEEE Trans. Image Process.* **2017**, *26*, 5019–5030. [CrossRef]
41. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights into Imaging* **2018**, *9*, 611–629. [CrossRef] [PubMed]
42. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
43. Xiao, J.; Wang, J.; Cao, S.; Li, B. Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks. *J. Phys. Conf. Ser.* **2020**, *1518*, 012041. [CrossRef] [PubMed]
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *NIPS* **2017**, *60*, 84–90. [CrossRef]
46. Cristianini, N.; Ricci, E. Support Vector Machines. In *Encyclopedia of Algorithms*; Kao, M.Y., Ed.; Springer: Boston, MA, USA, 2008. [CrossRef]
47. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001. [CrossRef]
48. Zhu, X.; Ramanan, D. Face detection, pose estimation, and landmark localisation in the wild. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2879–2886. Available online: www.ics.uci.edu/~{}xzhu/face/ (accessed on 18 September 2022).
49. Ahonen, T.; Hadid, A.; Pietikainen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef] [PubMed]
50. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
51. Ojala, T.; Pietikäinen, M.; Mäenpää, T. A Generalized Local Binary Pattern Operator for Multiresolution Gray Scale and Rotation Invariant Texture Classification. In *Advances in Pattern Recognition—ICAPR 2001*; Singh, S., Murshed, N., Kropatsch, W., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2001; Volume 2013. [CrossRef]
52. Wolf, L.; Hassner, T.; Taigman, Y. Descriptor based methods in the wild. In Proceedings of the Workshop on Faces In "Real-Life" Images: Detection, Alignment, and Recognition, Marseille, France, 17 October 2008.
53. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning transferable architectures for scalable image recognition. *arXiv* **2017**, arXiv:1707.07012.
54. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
55. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
56. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
57. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
58. Ranjan, R.; Zhou, S.; Chen, J.C.; Kumar, A.; Alavi, A.; Patel, V.M.; Chellappa, R. Unconstrained Age Estimation with Deep Convolutional Neural Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015; pp. 351–359. [CrossRef]
59. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *arXiv* **2016**, arXiv:1609.04802.
60. Rothe, R.; Timofte, R.; Van Gool, L. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *Int. J. Comput. Vis.* **2018**, *126*, 144–157. [CrossRef]

61. Fang, J.; Yuan, Y.; Lu, X.; Feng, Y. Muti-stage learning for gender and age prediction. *Neurocomputing* **2019**, *334*, 114–124. [CrossRef]
62. Guo, X.; Li, S.; Yu, J.; Zhang, J.; Ma, J.; Ma, L.; Liu, W.; Ling, H. PFLD: A Practical Facial Landmark Detector. *arXiv* **2019**, arXiv:1902.10859.
63. Antipov, G.; Baccouche, M.; Berrani, S.-A.; Dugelay, J.-L. Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognit.* **2017**, *72*, 15–26. [CrossRef]
64. Uddin, S.S.; Morshed, S.; Prottoy, M.I.; Rahman, A.A. Age Estimation from Facial Images using Transfer Learning and K-fold Cross-Validation. In Proceedings of the 3rd International Conference on Pattern Recognition and Intelligent Systems, Bangkok, Thailand, 28–30 July 2021; pp. 33–36. [CrossRef]
65. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *arXiv* **2017**, arXiv:1709.01507.
66. Dagher, I.; Barbara, D. Facial age estimation using pre-trained CNN and transfer learning. *Multimed. Tools Appl.* **2021**, *80*, 20369–20380. [CrossRef]
67. Nam, S.H.; Kim, Y.H.; Truong, N.Q.; Choi, J.; Park, K.R. Age estimation by super-resolution reconstruction based on adversarial networks. *IEEE Access* **2020**, *8*, 17103–17120. [CrossRef]
68. Liu, X.; Zou, Y.; Kuang, H.; Ma, X. Face image age estimation based on data augmentation and lightweight convolutional neural network. *Symmetry* **2020**, *12*, 146. [CrossRef]