Evaluating Automated Facial Age Estimation Techniques for Digital Forensics

Felix Anda, David Lillis, Nhien-An Le-Khac, Mark Scanlon
Forensics and Security Research Group
School of Computer Science
University College Dublin, Ireland
Email: felix.andabasabe@ucdconnect.ie, {david.lillis, an.lekhac, mark.scanlon}@ucd.ie

***Abstract*—In today's world, closed circuit television, cellphone photographs and videos, open-source intelligence (i.e., social media/web data mining), and other sources of photographic evidence are commonly used by police forces to identify suspects and victims of both online and offline crimes. Human characteristics, such as age, height, weight, gender, hair color, etc., are often used by police officers and witnesses in their description of unidentified suspects. In certain circumstances, the age of the victim can result in the determination of the crime's categorization, e.g., child abuse investigations. Various automated machine learning-based techniques have been implemented for the analysis of digital images to detect soft biometric traits, such as age and gender, and thus aid detectives and investigators in progressing their cases. This paper documents an evaluation of existing cognitive age prediction services. The evaluative and comparative analysis of the various services was conducted to identify trends and issues inherent to their performance. One significant contributing factor impeding the accurate development of the services investigated is the notable lack of sufficient sample images in specific age ranges, i.e., underage and elderly. To overcome this issue, a dataset generator was developed, which harnesses collections of several unbalanced datasets and forms a balanced, curated dataset of digital images annotated with their corresponding age and gender.**

1. Introduction

While human capabilities to detect and identify multiple facets, such as age, gender, ethnicity and facial expressions, can be accomplished by a quick glance at a digital image, machines are required to be trained intensively in order to understand traits present in photographs. Facial recognition has been the main attraction of several products in these last couple of years and has recently returned to the mainstream media with the release of Apple's iPhone X. This phone offers facial detection technology as its primary unlocking/authentication security mechanism that surpasses the traditional fingerprint authentication.

Hi-tech facial recognition is in active development around the world for a variety of applications. China has used facial recognition technology across multiple applications, e.g., ride hailing service driver identification, pay with a smile, jaywalker identification, etc. In the USA, it has been used in churches to track worshipers, and in the UK, it has been used to stop shoplifters. A myriad of facial identification applications have already reached the marketplace; often surprising consumers by the capabilities and reach that they offer. The use cases for accurate age estimation are not only limited to child abuse investigation but are useful across a range of crimes.

Age estimation commodities are becoming more common in our milieu. The appearance of ubiquitous smart face detectors will generate a considerable impact on the life of users and consumers. Migrating to an authentication factor, based on facial recognition, can strengthen system security to prevent impersonation attacks. Facial security checks could impede card cloning, fraudulent exam takers, and identity theft. Age estimation can also be used for adult entertainment venue access, purchase of age-restricted goods, such as alcohol and tobacco, age targeted advertising and recently, services such as `how-old.net` have been used for the recognition of children refugees from Syria. The aforementioned scenarios are just some examples of the variety of applications that can be achieved with "multi-layered deep learning technology for highly intelligent services" [1].

The accurate age estimation of a subject has always been a challenge for research across several fields. From counting the annual rings of wood growth to determine the age of a tree to measuring the skeletal maturity of a bone to obtain the age of a living being. Age estimation has been a complex task to achieve and several methods from measurement-driven anthropomorphic analysis to the application of machine learning algorithms have been applied and the accuracy of results obtained are constantly improving. Furthermore, the demand to distinguish the marginally under-age from the slightly overage is a matter of study where existing methods have been proven not be reliable enough to be able to perform the task.

[2] outlines a motivating facial recognition application regarding photo organization that is an application in digital forensics and cybercrime investigation. In order to process the considerable amount of data seized and processed in a typical case, a time-consuming, highly-skilled digital forensic analysis must be conducted. Furthermore, seized devices must often be processed immediately due to the urgent need

of evidence to progress an investigation that could be a matter of life or death.

Scanning the surface of a disk for data with probative value has long been a time-consuming task for forensic investigators. Nevertheless, innovative machine learning techniques and computer vision (core branches of artificial intelligence) can support digital forensic experts to carry out automated file classification, flag different types of content relevant to a case in court, and lessen the exposure of child abuse material in digital forensic laboratories.

Child abuse investigations are common occurrences in law enforcement agencies throughout the world. These investigations have become an arduous task due to the increasing usage of anonymization tools, private P2P networks [3], and cloud-based KVM systems [4]. Worldwide law enforcement and child protection communities have been struggling to diminish child abuse material (CAM[1]), and combat human trafficking. Organized cybercrime groups are operating in the deep web, which has become a hub for criminal black markets, where pedophiles are able to exchange vast amount of CAM; often to obtain acceptance within a group of pedophiles and ultimately gain access to other collections of illegal content [7].

Image classification and categorization according to age, gender, objects contained therein, and determining each image's location are often crucial for digital investigators. Similarly, grouping materials of the same person without specifically knowing their identity is a potentially valuable capability. Without an automated process, the procedure would resemble looking for a needle in a haystack. A major problem in machine learning is the lack of data for training and testing. According to [8], facial recognition is valuable for society but too intrusive on citizens' privacy. When minors are involved, privacy-enhanced age detection software should be enhanced by looking only at ages and not individual identities. Overall, a viable dataset should have a balanced number of faces at each age range. To satisfy an equally distributed dataset by age range and gender, we have created a multimedia dataset manager that allows the creation of a dataset on demand by selecting random pictures from various publicly available datasets. We have analyzed different online age estimation services and offline pre-trained models to determine their performance against a common dataset. The evaluation of these various pre-trained models has highlighted the lack of source child images used for the training phase. This has had an impact on several online and offline prediction services like Kairos and DEX [9]

1.1. Contribution of this Work

The contribution of this work can be summarized as:

- An overview of existing facial recognition techniques.
- Performance evaluation of offline and cloud-based facial recognition models with regards to their accuracy in the determination of ages and the influence of gender and the subject's actual age on the models' estimations.
- The release of a tool designed to generate unique, uniformly distributed random images by age and gender from several facial image datasets (such as FG-NET, FERET [10], IMDB-WIKI [9], MEDS [11], YFCC100M [12]) .

1. While other nomenclatures are interchangeably used in the literature for this illegal material, CAM is that adopted by [5] and [6].

2. Literature Review/State of the Art

The importance of sharing datasets in order to save time and money in the research community is fundamental. Facial image datasets annotated with both age and gender are needed to train machine learning models and predict further information from incoming data. [13] stresses the benefit of sharing datasets within the research community in favor of replicating results.

There have been previous studies on age estimation, where datasets have been shared. [14] noted the absence of facial data and offered a dataset that gathered images in the wild from public Flickr creative commons licensed albums to overcome this weakness. The age and gender were annotated by observation engendering the specification of 8 age groups.

In 2016, [9] crawled over half a million celebrity images from both Wikimedia and IMDB to produce a dataset of images annotated with actual age and gender. Their study, denoted as Deep Expectation (DEX), won the first place of the ChaLearn LAP 2015 challenge on apparent age prediction with a convolutional neural network (CNN) of 16 weight layers pre-trained on ImageNet (an image dataset that is organized according to the WordNet hierarchy) [15].

Published in the same year, [16] acknowledged that the accuracy in age estimation of child pictures were significantly lower in comparison to its other age group counter parts due to the lack of images for the mentioned age group. Strong legal and ethical issues arise with the use and distribution of child images. The work by [16] attempted to overcome this hurdle by compiling a total of 1,655 images for a 0-25 year old age range. Unfortunately, this dataset has remained private.

2.1. Digital Forensic Backlog

Storage capacities are growing exponentially and in combination with the growing needs for digital forensic analysis in a variety of cases, this results in a vastly increased volume of data requiring digital forensic expertise than current capabilities in law enforcement agencies throughout the world. This results in significant delays in the judicial process and can result in court cases being dismissed due to insufficient evidence [17]. According to [18], there is a less likelihood of prosecution due to the uncertainty in determining the age of a victim portrayed in a digital image. The backlog is growing due to both the lack of

relevant experts to analyze the data and a overly arduous digital forensic process [19]. Per [20], these factors will continuously influence the throughput of digital forensic laboratories and therefore, are likely to continue hindering digital forensic investigators in the future.

2.2. Human Facial Age Perception

Humans are quite accurate at estimating the age of other humans. The error rate has been measured to vary from as low as 2.07 years and as high as 8.62 years depending on a variety of factors including the age of the assessor, the age of the subject, and the difference between the two [21]. The age of young people tends to be consistently overestimated [22], [23], [24] and a tendency to assimilate the estimated age with one's own age is suggested [25], [26]. Moyse and Brédart [21] presented a study on own-age bias in the accurate estimation of faces. The authors found that their 114 participants were more accurate at estimating the ages of those within their own age-group (10-14, 20-30, and 65-75 years old). The accuracy of human age estimation of others can also be negatively impacted by a range of other factors including gender [26] and emotion/facial expressions [26], [27]. Neutral expressions results in the highest accuracy, whereas any other expression results in less accurate estimations [26].

2.3. Age Estimation

The age of the victim is vital to determine in a CAM subject in an era where much of digital investigators' time is taken up processing these cases. Age as a soft biometric trait is difficult to predict due to discrepancies between face and body features, absence of reliable cues, natural variation regarding variability across different ethnicities, and the environment where the victim appears [28]. The aforementioned research takes into account multiple factors that can lead to the classification of an image either if it is an indecent image and the respective age group. Countless studies on age estimation have been developed. In order to measure age estimation accuracy, we have considered the Mean Absolute Error (MAE), which is the difference between the estimated age and the actual age. In the past two decades, these error rates have been decreased remarkably. From early 2002 to date, the published MAE rates have been oscillating between 1 and 5 years. A MAE of 1.47 was achieved by [18] in 2014 by accomplishing an AdaBoost powered fusion of several state-of-the-art classifiers such as Fisher's LDA, Neural Networks, SVM, etc. Nevertheless, this study was executed over a limited private dataset of 50 female images with an age range from 10 to 19, which highlights the lack of data available as a consequence to the scarcity of images available of youngsters and the ethical implications required to use their pictures. In 2011, [29] was able to obtain a MAE of 4.1, which has been the predominant ratio amongst other algorithms that utilized the FG-NET dataset. The Contourlet appearance model used was more accurate and faster at localizing facial landmarks than Active Appearance Models. [30] acknowledged poor accuracy results for age estimation on juvenile faces by human observation. Moreover, female age estimation was more accurate in younger age groups and male age prediction were more precise after 11 years of age. However, the statements are based on a small sample stored in a private dataset.

2.4. Transfer Learning

Multiple researchers have published pre-trained models to avoid the tedious task of training data and optimizing the cost of running algorithms on hardware. Transfer learning is a new learning framework that allows us to use pre-trained models from other researchers. [31] exploited the transfer learning strategy to train deep convolutional neural networks due to the lack of age labeled face images. They state that transfer learning includes pre-train and fine tune where in the former, the randomly initialized networks are first trained with a fair amount of labeled data and in the latter, learned parameters in the mentioned former process are used as an initialization for a new task. Pre-trained models are simply a model created to solve a specific problem and are prone to re-usability.

Well documented pre-trained models for age estimation are communal in the Caffe Model Zoo. [32] shared a deep convolutional neural network for age and gender classification in 2015. Their model was trained with the Flickr dataset of facial images in the wild [14], to raise performance in learning representations when limited data is available. Training each network required about 4 hours using a robust GPU. Similarly, [33] proposed a ranking CNN-based framework for age estimation also trained over the Adience dataset used by [14]. Finally, we take into consideration a pre-trained model external to the model zoo but compatible with the Caffe framework. The Deep Expectation (DEX) model approaches the automated estimation of facial ages with a CNN [9].

2.5. Facial Age Datasets

In a recent study by [34], the multiple algorithms compiled in their work have been evaluated under the public domain FG-NET dataset. Moreover, they state that such dataset has been biased towards young children. This is a motivating fact for our study. However, the numbers shared by FG-NET for underage images were less than our expectations. The MORPH dataset consists of approximately 78,000 images of subjects with age ranges between 15 and 77. This dataset is useful for facial recognition and relevant to our work when we consider age estimation for the teenager faction. Age and gender labels are well documented unlike broad datasets where tagging features is unusual. Analogously, MEDS [11] is a mugshot dataset of deceased subjects with the oldness feature annotated but the age range irrelevant to underage individuals. The FERET dataset contains approximately 14,000 images and is pertinent to face detection [10]. The age labeling is based solely on observation; therefore, our

research cannot rely on conjectures due to the considerable MAE values for age prediction produced by the state-of-the-art age estimation algorithms. Therefore, we have considered omitting the use of this dataset. The largest dataset available to our knowledge is the IMDB-WIKI dataset shared by [9]. This dataset consists of over half a million photos from celebrities with an ample age range and considering that the images were obtained by scraping images over the Internet, we have been cautious on using such images due to the copyright restrictions. For our study, we have encountered noise in the source provided by DEX therefore we had to implement a filter to overcome this issue. Furthermore, our solution reduced the quantity of images available per age cluster and the lack of underage images was inescapable.

The OUI-Adience set is a public collection of labeled images obtained by online facial images of Flickr "in the wild". Although [14] has stated that they use Creative Commons license for their images, we have detected from a sample of 10,842 images, that the 89.55% are associated to images with copyright. We have opted to omit these images yet we acquire alternative photos associated to consistent tags and titles that can support the information over the age of the child and that are subject to creative commons licenses.

By creating a combined dataset from a variety of constituent datasets (IMDB, WIKI, FG-NET, MEDS) using the age dataset generator software, it was recognized that the volume of images was insufficient for particular age ranges, i.e., for both the underage and the elderly. As an example, throughout the aforementioned combined dataset, there are ¡100 images available for each of the ages between 0 and 6 inclusive for males and 0 and 7 inclusive for females. The program randomly selects unique images from different datasets within a pool greater than half a million pictures. With the constant operation of the software solution, the values will tend to reduce when a user filters noisy images.

2.5.1. Yahoo Flickr Creative Commons 100M. For the benefit of the research community, the Yahoo Flickr Creative Commons 100M (YFCC100M) was released in 2014 [12]. To the best of our knowledge, this is the biggest dataset of images and videos liberated for scientific purpose. Due to the size of the collection, the dataset is volatile; however, an updated response from a query to the Flickr API is possible considering identification keys stored in the set. Initially, the sheer records were stored in a single text plain file. To manage such volume of information, a script was executed and iterated the repository line by line, copying each record to a NoSQL MongoDB collection. Relevant indexes had to be created so the queries could process with a prompt response. Once the database was set, it was possible to retrieve data from the collection and craft URLS so the image could be accessed from the public domain, filter images by tags, and chose the adequate Creative Commons license. This dataset is useful for our study as we can acquire creative common licensed images of individuals of particular underrepresented ages.

3. Existing Tools and Models

Both online and offline tools are considered for the evaluation of the performance of age estimation. The main advantage of using cloud-based biometric services is that the results obtained are processed by state-of-the-art classifiers developed by experienced companies in the space, such as Amazon, Microsoft, and IBM. The main disadvantage of online tools is the ongoing costs associated with their usage. Most of the service responses are configured in JSON, which allows us to easily integrate our performance evaluators.

In 2010, Amazon acquired "Rekognition" from an Artificial Intelligence start-up company, Orbeus [35]. The company had developed a facial recognition software that detected traits on images with Artificial Neural Networks (ANNs). ANNs are systems that learn to accomplish tasks by observing examples rather than executing a specific algorithm. They are structured by an initial input layer of neurons, one or more hidden layers, and a final layer of output neurons.Machine Learning as a Service was introduced to facilitate non-experts in the training of models without expertise in the topic. The service is a deep learning-based image analyzer that is able to detect age with a minimum and maximum value as a dual class output. We have evaluated the most suitable results and in order to normalize the output to a single rate, the tests were conducted assessing the mean average error with the minimum, maximum, and mean value of the output range. Our investigation led to the use of the minimum value, which is also an acceptable threshold for the procedures in a digital forensic case where the cost of inaccuracy is potentially high.

Deep Expectation (DEX), as mentioned in Section 2.4, is a pre-trained Convolutional Neural Network model that achieved a satisfying ratio amongst the other online services. The principal preference on using offline pre-trained models is the freedom to execute as much estimation as it is needed. The proposed method by [9] used the architecture of the aforementioned model and was the winner of the Chalearn LAP 2015 challenge [36]. The response requires normalization but differs from Rekognition due to improved results over the mean value between the minimum and maximum age. Kairos (a free online service) uses a Support Vector Machine (SVM) algorithm for the model to help isolate different types of faces into the corresponding age class. Nevertheless, the performance is low compared to the rest of the age estimation services. Finally, Microsoft Azure Cognitive Service uses a Multi-layered deep learning methodology [1].

4. Overview of the software solution for performance evaluation

To evaluate the state-of-the-art cloud-based biometric services, a significantly robust set of labeled digital images was required. A non-biased collection of images was generated by selecting random unique photos from the different datasets mentioned in Section 2.5. The query criteria applied

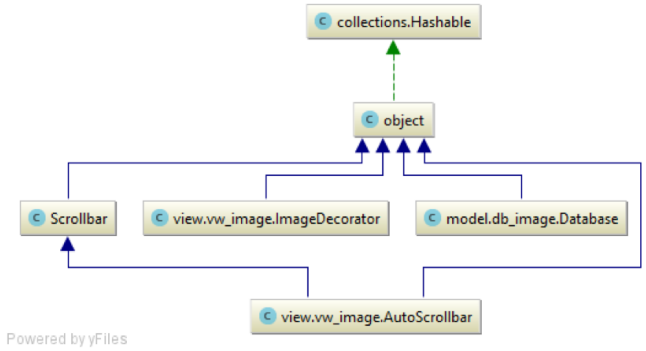


Figure 1. Decorator Pattern UML Class diagram

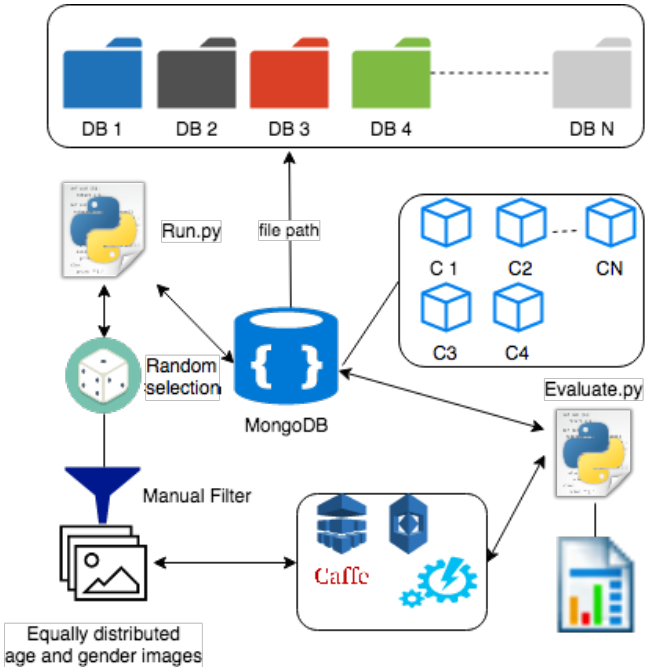


Figure 2. System Architecture

to obtain results are: minimum age, maximum age, and number of images. The software had to be scalable to fulfill future adaptations to more datasets and services. Therefore, an Model-View-Controller (MVC) software architecture was proposed that enabled code re-utilization and parallel development. Multiple design patterns and inheritance made scalability permissible. Refer to Figure 1 for a UML class diagram example of a decorator pattern used in our work. The services included for evaluation performance are: Amazon Rekognition[2], Microsoft Azure Cognitive Services[3], KAIROS[4], and DEX [5]. Inheritance enables the seamless addition of new services for future evaluations.

2. https://aws.amazon.com/rekognition/
3. https://azure.microsoft.com/en-us/services/cognitive-services/
4. http://kairos.com
5. https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

The system architecture is composed of multiple file sources (FG-NET, FERET, IMDB, WIKI, MEDS, YFCC100M, etc.) that correspond to each image dataset. The metadata was stored in different collections belonging to each file set. Python scripts were developed to assist the task and in some cases, Comma Separated Value (CSV) files were used in order to interoperate with the NoSQL repository and the scripts. Once a randomly selected dataset of images that is equally distributed by age and gender, both offline and cloud-based biometric services were evaluated. Figure 2 outlines the system architecture.

A key component of the system architecture is the manual filtering step. The amalgamation of various random images per age class generates a dataset with noise that can only be effectively filtered through user interaction. Users are presented with an interface whereby they are able to discard images that are not useful and randomly generate new images by clicking on each image button, as illustrated in Figure 3.

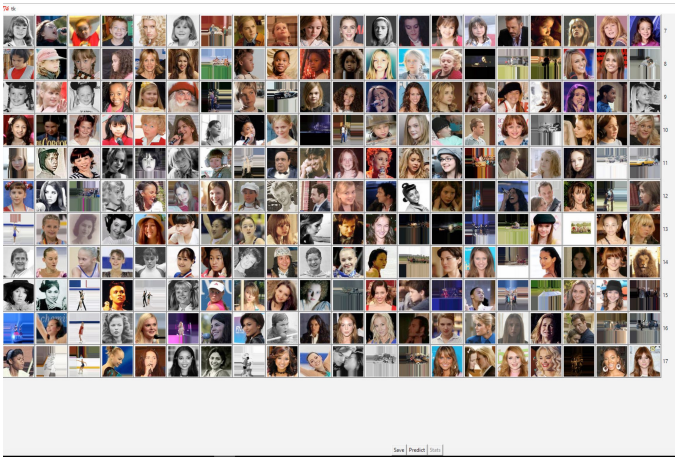


Figure 3. Dataset Generator Software

For version control, Bitbucket has been used and our repository is publicly available from https://bitbucket.org/4nd4/image_database.git

5. Evaluation Methodology

The type of evaluation used in the research is an empirical evaluation based on observation. The purpose of the evaluation is to find the least MAE within the different cloud-based biometric services and pre-trained models. The results of the evaluation would be helpful in selecting which one is most effective, or indicate what combination of different services might aid in creating a data fusion/ensemble approach. Our research exploits the pre-trained Caffe model produced by DEX in order to foretell the age and gender of digital images. With this tool and other state-of-the-art online age predictors, comparative analyses were performed to evaluate both the efficiency and accuracy of the different predictions.

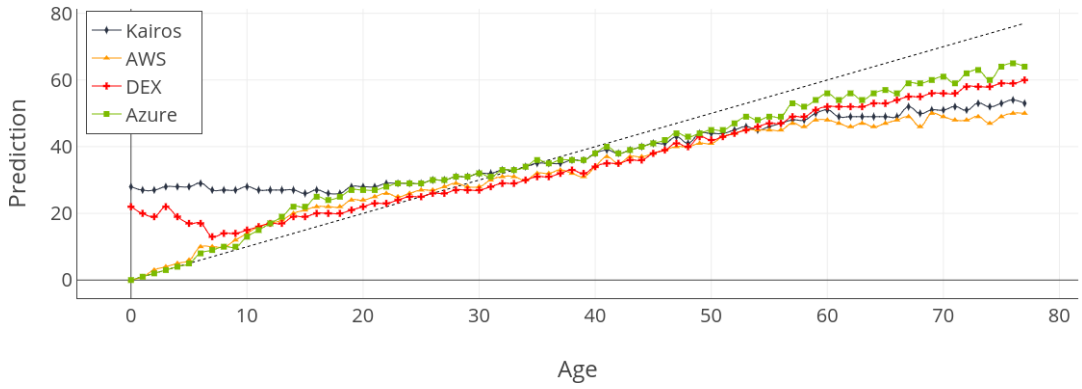


Figure 4. Average Estimated Age Compared with Actual Age across Entire Dataset.

Each of the cloud-based biometric services mentioned for age estimation were evaluated by generating a random equally distributed dataset in order to avoid unbalanced analysis. The equal distribution required that the same number of images for each age were used. This had a limiting effect on the total size of the dataset, as there was a lack of images for certain ages. Initially, we had intended to consider the whole age range from 0 to 100; however due to lack of images, an age range from 0 to 77 was studied. Due to a scarcity of images of youngsters between the ages of 0 and 14 and the applicability of this age range to CAM investigation, additional images were manually collected. Creative commons licensed pictures with accurate age and gender were gathered from Flickr. Photos were manually labeled with the age and gender of the individual based on the descriptions, title of the photo, the respective tags or any visual clues. This process ultimately ensured that 65 images per age per gender could be included; resulting in a total dataset size of 10,140 images.

Each image was passed through each of the four systems, and the results recorded. These were then evaluated under three influencing factors, the results of which are discussed in Section 6. Initially, the four systems were compared across the entire age range by analyzing the error rates that each exhibited, i.e. the difference between the predicted age of each subject and their actual age. Next the dataset was divided by gender to investigate whether this had any effect on the accuracy of the age predictions. Finally, in the third test the dataset was subdivided into a number of age ranges (i.e. 0-9, 10-19, 20-29, etc.). The goal here was to find whether certain systems performed better in different age ranges, or whether one system could be said to be the most accurate over the entire dataset.

6. Results

This section outlines the results from our evaluation of the current online and offline age estimating options. Firstly, we present the results across the entire age range of our dataset (0-77 years old) in Section 6.1. Next, Section 6.2 presents the results of subdividing the dataset by gender. Finally, in Section 6.3 we compare the performance of the four systems within different age ranges.

6.1. Entire Age Range Estimation

The first analysis that was conducted was to measure the accuracy associated with each of the four systems across the entire dataset, with a view to discovering which services are most effective. Firstly, the MAE was calculated for each system across all the subjects. The results of this are shown in Table 1.

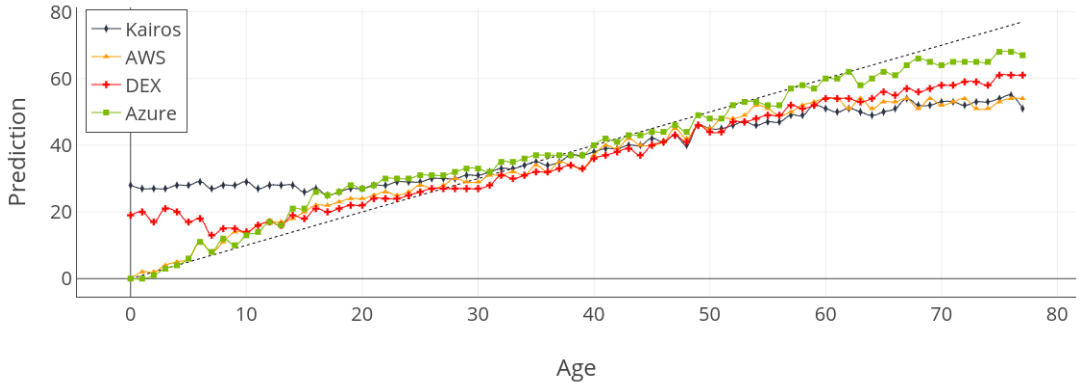
TABLE 1. MEAN ABSOLUTE ERROR PER SERVICE.

Service	MAE
Kairos	11.236
AWS	9.286
DEX	8.079
Azure	7.614

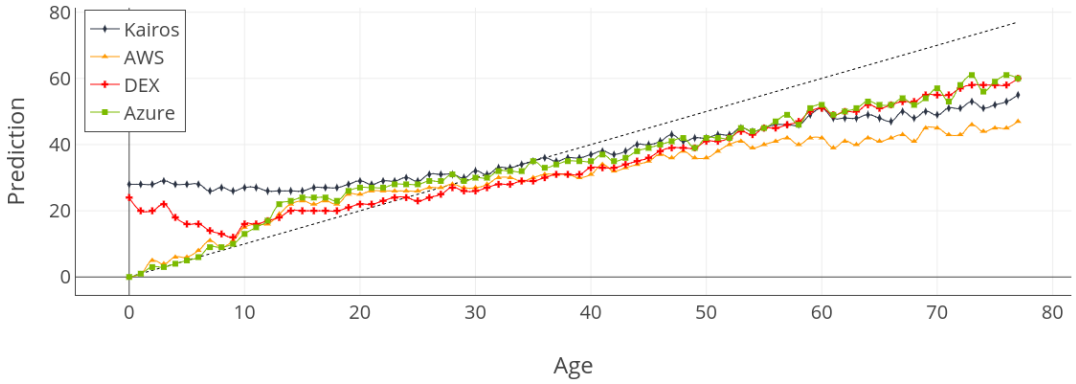
Using this metric, Microsoft Azure was found to achieve the best results, with the lowest MAE. This can be interpreted to mean that across the entire range of all subjects, the average difference between the predicted age and the actual age was 7.614 years. The error rates for the other services were higher, with DEX achieving better overall results compared to AWS, which in turn outperformed Kairos. While this is a useful finding in itself, a more in-depth view is required to examine the characteristics of each service further.

Figure 4 illustrates the performance of the four systems more clearly across the dataset. The X-axis indicates the actual age of the subjects. A line is plotted for each service, which indicates the average age it predicted for subjects in each age class. Each point therefore represents the average predicted age for 65 subjects that have the same age. The dotted line is used to indicate where correct predictions should lie.

A number of interesting observations can be made from this figure. Both DEX and Kairos have a tendency to substantially over-estimate the age of young children. In the case of Kairos, this over-estimation continues well into the



(a)



(b)

Figure 5. Average estimated age compared with actual age for (a) male subjects (b) female subjects.

late teenage years, before its predictions become closer to those of its competitors after this point.

In all cases, a tendency to underestimate the age of subjects begins to emerge from approximately the age of 40, though this is more pronounced for some services. Kairos, in addition to being the least accurate at early age ranges, also has the second-highest error rate for older subjects, behind only AWS. DEX, from having a high error rate for young children, becomes the second most accurate (behind Azure) at later ages. Its tendency to underestimate ages becomes apparent earliest, from the late-20s onward. The line for AWS is very close to the true age line in the early stages, but exhibits the highest level of underestimation for the later ages.

When bearing digital forensic use-cases in mind, it is worthwhile focusing on the late teenage years in particular, around the boundaries where people cease to be minors in various jurisdictions. The very accurate performance of AWS and Azure begins to diverge from the correct prediction line around the age of 10. DEX, which performs poorly on young children, is closest to this ideal line in the mid-to-late teenage years and continues into the early 20s. A further examination of the relative performance of the systems in various age ranges is contained in Section 6.3.

6.2. Influence of Gender on Estimation

To further explore the characteristics of the four services, we also divided the dataset according to gender, and conducted a similar analysis to the previous section. The overall MAE for each service is shown in Table 2. The main interesting insight that can be gained from this table is that uniformly, all four services exhibit a higher rate of error for female subjects than for males. Of these, Kairos is the only one for which the difference in error rates by gender is less than 1 year on average. The difference is most pronounced for AWS, for which the error rate for male subjects is only marginally greater than for DEX, but whose predictions for female subjects are more comparable to Kairos. The relative ranking of the four systems remains the same for both genders, however.

TABLE 2. MEAN ABSOLUTE ERROR PER GENDER PER SERVICE.

Service	Male	Female
Kairos	10.6838	11.7960
AWS	7.2192	11.4057
DEX	7.1975	8.9613
Azure	6.4205	8.8092

As with the previous section, a more in-depth view is re-

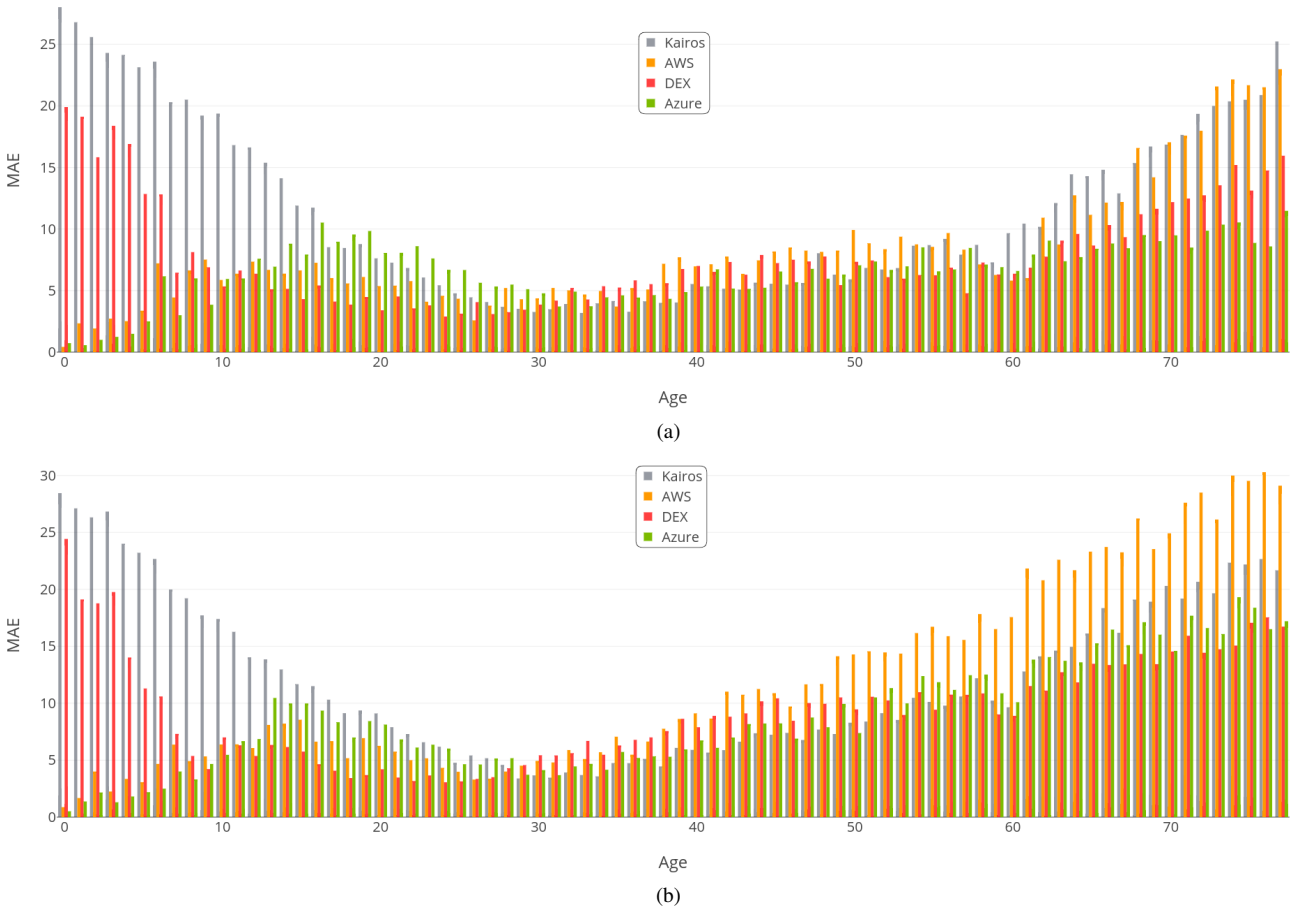


Figure 6. Mean Average Error Rate for (a) Male and (b) Female Subjects.

quired beyond the overall error rates. Figure 5 is constructed in the same way as for Figure 4, with the exception that it shows separate line graphs for each gender. Figure 5(a) refers only to the results for male subjects, and Figure 5(b) relates only female subjects.

The overall patterns observed in the previous section are generally apparent in both graphs: Kairos and DEX overestimate at young ages and all four services tend to underestimate the age of older subjects. However, the rate at which the latter effect occurs is far more pronounced for female subjects. By the age of 36, all four services underestimate on average, and this gap becomes more pronounced with increasing age. In contrast, Azure in particular remains much closer to the ideal line for male subjects.

Figure 6 illustrates this data by displaying the MAE rate for each of the systems at each age. Error rates generally increase towards older ages, with this being more pronounced for female subjects, due to the age underestimation common to all services. As previously observed, Kairos and DEX exhibit relatively high error rates for young subjects. However, it is notable that although Kairos is clearly the least accurate for young subjects, it achieves better error rates than the other systems towards the middle of the ages evaluated. This is the focus of the following section, where this data is viewed within a range of age brackets.

A local peak is observed in the teenage years, before error rates decline into the 20s and 30s. This suggests that more focus is required on this area in the future, especially due to the use cases that require accuracy within this borderline adulthood age range.

6.3. Age Range Analysis

Previous observations of the data indicate that although the Azure performed better than the other four on average, performance was affected not only be gender by also according to age. This motivated a deeper analysis of the relative performance of the four systems across different age ranges. For this, the dataset was subdivided into 10-year age ranges (0-9, 10-19, 20-29, etc.). The only exception was the final range, which was from 70 to 77 due to the lack of available older subjects in constituent datasets.

Figure 7 was generated to provide insights into the data. In this figure, a box is plotted for each service within each age range. For each box, the data used was the average predicted age for each actual age. The boxes show the mean, median, interquartile range, with the whiskers representing

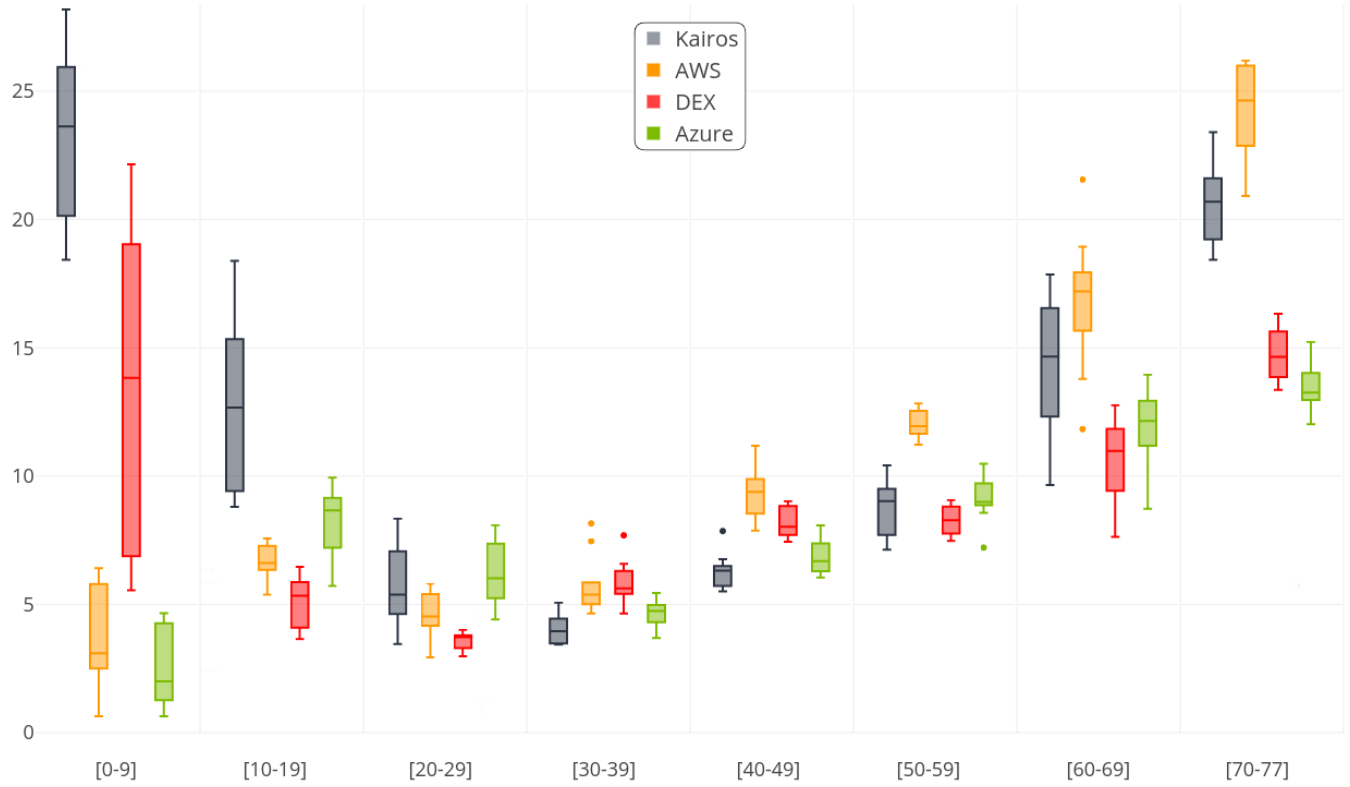


Figure 7. Mean Average Error Rate for Each System in Different Age Ranges.

the maximum and minimum values within 1.5 times of the interquartile range. Outliers are shown as individual points where relevant.

The primary object of this study is to ascertain which system(s) offer the most accurate performance for the age prediction task. Although Azure has been shown to the lowest overall error, this figure indicates that this does not reflect an overall superior performance across all age ranges. The best-performing system, on average, in each age range is summarized in Table 3. Azure is the most precise only for the youngest and oldest age ranges. A somewhat surprising result here is that although Kairos has the highest overall error rate, it is the most precise on average in the 30-39 and 40-49 age ranges. The high rate of error for DEX for the youngest children is much reduced by the teenage years, and this system has the lowest error rates for the 10-19 and 20-29 age ranges. Indeed, it is again the most precise for 50-59 and 60-69. In total, DEX has the lowest error rate for four of the age ranges, with Azure and Kairos having the lowest error for two range apiece. AWS is not the most accurate in any of the ranges, but is the second-best average performer for all ranges up to the age of 29.

TABLE 3. PERFORMANCE PER AGE RANGE.

Age Range	Lowest Mean Absolute Error
0-9	Azure
10-19	DEX
20-29	DEX
30-39	Kairos
40-49	Kairos
50-59	DEX
60-69	DEX
70-77	Azure

7. Concluding Remarks

There are numerous machine learning-based methods that are focused on easing the digital evidence backlog. As part of this effort, it is crucial to invest into improving the accuracy of automated age estimation in photographic and video content. This study evaluates four different age prediction services (both cloud-based services and offline sources). The outcome of this evaluation emphasizes a higher error rate for female subjects; from which we can deduce that gender is a soft biometric trait that significantly impacts the overall accuracy of the age prediction models. Male age estimation was more accurate and, on average, had a MAE of approximately 2.1 years better than the female subjects.

It was important to determine the behavior of the estimation services in different age bands. Thus, 3 different evaluations were performed; individual age estimation, influence of gender on age estimation, and several grouped age ranges. Our research proves that although the Microsoft Azure online evaluation service was predominant with the lowest overall error, the other services performed with better results within a number of specific age ranges. This conclusion

encourages further work in this area and indicates that relying on a single age estimation service might not be apt.

The amount of images accurately labeled with age and gender available to researchers is limited. We proposed a random, balanced dataset generator to overcome this hindrance by combining existing datasets. Furthermore, we have included the collection of underage digital images, which helped us fairly evaluated performance over males and females. This dataset will be made available for the use in age prediction research and other aspects, such as an asset for which pre-trained models can be enhanced.

7.1. Future Work

It is clear that the task of selecting a best-performing facial age estimation technique is more complex than merely choosing one with a lower overall error rate, and that gender and age are significant factors in influencing the effectiveness of all four systems considered.

The fact that varying systems perform best in different age ranges motivates further investigation as to how the results of a variety of systems could be combined together to improve the overall accuracy of predictions. Numerous machine learning regression techniques are available that have the potential to use the system predictions as inputs and to provide a prediction that is hopefully closer to the subjects' real ages than the individual systems.

The ultimate aim of this work is to automate the arduous task of analyzing digital data from seized devices. Therefore, our ambition is to investigate how to aid digital forensic cases with automated machine learning-based techniques. As future work, our objective is to expand this study further through comparative analysis of further services.

Moreover, the weakness of the current tools is presented where the supplied photograph is not particularly clear. Because of the angle at which it was taken and/or poor quality lighting. These are standard problems in all forms of facial recognition that we wish to address in the future.

References

[1] H. Weber, A. Cruz Rodrigues, and A. Mateus, "Emotion and Mood in Design Thinking," *Design Doctoral Conference'16: TRANSversality - Proceedings of the DDC 3rd Conference*, no. July, pp. 65–72, 2016.

[2] G. Guo, "Human age estimation and sex classification," *Studies in Computational Intelligence*, vol. 409, pp. 101–131, 2012.

[3] R. Hurley, S. Prusty, H. Soroush, R. J. Walls, J. Albrecht, E. Cecchet, B. N. Levine, M. Liberatore, B. Lynn, and J. Wolak, "Measurement and analysis of child pornography trafficking on p2p networks," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 631–642.

[4] J. Farina, M. Scanlon, N.-A. Le-Khac, and M.-T. Kechadi, "Overview of the Forensic Investigation of Cloud Services," in *10th International Conference on Availability, Reliability and Security (ARES 2015)*. Toulouse, France: IEEE, 08 2015, pp. 556–565.

[5] M. Aiken, M. Moran, and M. J. Berry, "Child abuse material and the internet: Cyberpsychology of online child related sex offending," in *29th meeting of the INTERPOL Specialist Group on Crimes against Children, Lyons, France, September*, 2011, pp. 5–7.

[6] B. Jones, S. Pleno, and M. Wilkinson, "The use of random sampling in investigations involving child abuse material," *Digital Investigation*, vol. 9, pp. S99–S107, 2012.

[7] T. Krone, "A typology of online child pornography offending," *Trends & issues in crime and criminal justice*, no. 279, pp. 1–6, 2004.

[8] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft, "Privacy-preserving face recognition," in *Proceedings of the 9th International Symposium on Privacy Enhancing Technologies*, ser. PETS '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 235–253.

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

[10] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[11] A. P. Founds, N. Orlans, W. Genevieve, and C. I. Watson, "Nist special databse 32-multiple encounter dataset ii (meds-ii)," *NIST Interagency/Internal Report (NISTIR)-7807*, 2011.

[12] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[13] C. Grajeda, F. Breitinger, and I. Baggili, "Availability of datasets for digital forensics–and what is missing," *Digital Investigation*, vol. 22, pp. S94–S105, 2017.

[14] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] P. Grd and M. Bača, "Creating a face database for age estimation and classification," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016 39th International Convention on*. IEEE, 2016, pp. 1371–1374.

[17] E. Casey, M. Ferraro, and L. Nguyen, "Investigation delayed is justice denied: proposals for expediting forensic examinations of digital evidence," *Journal of forensic sciences*, vol. 54, no. 6, pp. 1353–1364, 2009.

[18] M. Ratnayake, Z. Obertová, M. Dose, P. Gabriel, H. Bröker, M. Brauckmann, A. Barkus, R. Rizgeliene, J. Tutkuviene, S. Ritz-Timme *et al.*, "The juvenile face as a suitable age indicator in child pornography cases: a pilot study on the reliability of automated and visual estimation approaches," *International journal of legal medicine*, vol. 128, no. 5, pp. 803–808, 2014.

[19] D. Lillis, B. Becker, T. O'Sullivan, and M. Scanlon, "Current Challenges and Future Research Areas for Digital Forensic Investigation," in *The 11th ADFSL Conference on Digital Forensics, Security and Law (CDFSL 2016)*. Daytona Beach, FL, USA: ADFSL, 05 2016.

[20] M. Scanlon, "Battling the digital forensic backlog through data deduplication," in *Innovative Computing Technology (INTECH), 2016 Sixth International Conference on*. IEEE, 2016, pp. 10–14.

[21] E. Moyse and S. Brédart, "An own-age bias in age estimation of faces," *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, vol. 62, no. 1, pp. 3 – 7, 2012.

[22] J. B. Pittenger and R. E. Shaw, "Perception of relative and absolute age in facial photographs," *Attention, Perception, & Psychophysics*, vol. 18, no. 2, pp. 137–143, 1975.

[23] R. Henss, "Perceiving age and attractiveness in facial photographs," *Journal of Applied Social Psychology*, vol. 21, no. 11, pp. 933–946, 1991.

[24] G. Willner and P. Rowe, "Alcohol servers' estimates of young people's ages," *Drugs: education, prevention and policy*, vol. 8, no. 4, pp. 375–383, 2001.

[25] J. Vestlund, L. Langeborg, P. Sörqvist, and M. Eriksson, "Experts on age estimation," *Scandinavian Journal of Psychology*, vol. 50, no. 4, pp. 301–307, 2009.

[26] M. C. Voelkle, N. C. Ebner, U. Lindenberger, and M. Riediger, "Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age." *Psychology and Aging*, vol. 27, no. 2, p. 265, 2012.

[27] T. Ganel, "Smiling makes you look older," *Psychonomic bulletin & review*, vol. 22, no. 6, pp. 1671–1677, 2015.

[28] J. A. Kloess, J. Woodhams, H. Whittle, T. Grant, and C. E. Hamilton-Giachritsis, "The challenges of identifying and classifying child sexual abuse material," *Sexual Abuse*, p. 1079063217724768, 2017.

[29] K. Luu, K. Seshadri, M. Savvides, T. D. Bui, and C. Y. Suen, "Contourlet appearance model for facial age estimation," in *Biometrics (ijcb), 2011 international joint conference on*. IEEE, 2011, pp. 1–8.

[30] E. Ferguson and C. Wilkinson, "Juvenile age estimation from facial images," *Science & Justice*, vol. 57, no. 1, pp. 58–62, 2017.

[31] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," *Neurocomputing*, vol. 187, pp. 4–10, 2016.

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

[33] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

[35] K. Rajesh and K. Ramesh, "Artificial intelligence–fact or fiction," *Computing NaNo*, 2012.

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