July 2021

Innovative Technology
POWERING TRANSACTIONS AND INTERACTIONS

White Paper

Age estimation: determining age using facial features

In this paper we will outline the challenges in anonymously estimating a person's actual age from physical cues in their appearance. We will present age estimation technology, an automated system which estimates a person's age and tackles the known weaknesses which humans exhibit performing the same task. We argue that the age range in which estimation should be most accurate is early-stage adulthood, and present both reasoning and data to reinforce this idea.

We will present results from an independent testing body, which has concluded that our technology has performed to a level of accuracy which allows it to be deployed in a Challenge 25 scenario to control access to age restricted goods. We also introduce factors which influence true bias (differences in measurements related solely by gender or skin tone) and apparent bias (differences in measurements caused by other factors but may be attributed to gender or skin tone) and will present some preliminary results which demonstrate very promising performance of the system in tackling both true bias in gender and skin tone.

Key Findings

1. Humans misjudged **38% of 16 yr old boys and 56% of 16 year old girls** to be **over 18**
2. Underestimates **18-year-olds** on average by only **0.19 years**
3. Tackles **gender and skin tone bias** in age estimation technology
4. Helps restrict access to age restricted goods
5. Total offline solution with **no recurring costs**

Introduction

Chronological or actual age is the time passed since date of birth, and due to recording of the event can be known with very high accuracy (minutes or even seconds). The question we will tackle in this paper is, can we accurately estimate a person's age without referring to personal documents? To what accuracy can we do this, and what tools can we use?

Age estimation is a simple concept – is it possible to determine a person's actual age solely on examination of some physical characteristics. Methods employed range from the very intrusive - dental X-rays, examination of bone structure, inspecting blood or tissue samples, to the very passive analysis of a person's face. Age estimation depends on interpretation of some physical cues, which can be used as a substitute measurement in the absence of personal documents.

The challenge in using physical characteristics as a determination of age is due to the differing rates at which people's physical attributes change with time, and consequently the divergence in the appearance of those physical cues across the whole population. Simply stated, different people age at different rates, people of the same chronological age will present with differing physical traits. It is quite evident, there exists a disconnect between a person's chronological age (actual time from date of birth) and their physiological age (visible effects which are used to reflect age).

As we will highlight in this paper, employment of physical traits to determine true actual age will not be as accurate as documentation stating a person's birth date. However, for certain applications we can ask, do we really need to know a person's age in years, days and minutes? What approach will give us the best results and how can we implement such a tool to benefit society. We will also present results both obtained internally, and results obtained independently in the evaluation of our age estimation technology.

Challenges for the ageing face

From birth our face will undergo changes which can be both used as physical cues relating to age, but also affect the perception of our true age. Children especially all grow and change at different rates so this will limit the accuracy of age estimation[1]. The journey from childhood to adulthood can be referred to as the first stage of physical changes, where the development of the bones in the skull are the primary contributary factor to the changing face. Beyond this as we enter the second stage of ageing (mid 20s.), other factors become the primary contributors to the appearance of the ageing face in particular skin appearance and texture which form the biggest cues for age perception. Genetics, lifestyle choices (poor diet, smoking, drug abuse) and illness will have a large influence on the ageing process and presents a big challenge in accurately estimating a person's age as they approach middle age and progress into old age. Many of these factors become more influential with time, and as such the contribution can be more pronounced over time and differ from individual to individual.

The age of early adulthood (17-21), where there is less divergence across the population in the physical cues of ageing, may offer the best opportunity in using the facial characteristics as a good estimate of actual age. The diversity in rates of changes in childhood will add complications in achieving high accuracies. Similarly for older cohorts (>30), the influence of external factors (lifestyle, health etc) will play a larger part in the rates of ageing and as such it would be expected that there is a bigger challenge in estimating the true chronological age.

The challenge for humans in estimating age

Humans use the various facial cues to estimate a person's age either consciously or sub-consciously. If you are responsible for controlling access to age restricted products, then determining the actual age is crucial to protect children from potential harm. In particular, correctly identifying subjects under the legal age is vital.

There are known factors which influence a human's ability to accurately estimate age from a subject's appearance. Moyse et al suggested that humans tend to be most accurate in estimating ages of individuals closer to their own age[2]. Fujisawa et al also concluded that the bias is driven by a comparison with the estimators own age[3]. Additional factors will also play a

role in a person's ability to perform decision making – these include fatigue, hunger, and the mood of the estimator.

Danziger et al[4] presented some interesting findings relating to judicial decisions which did suggest that judicial rulings can be swayed by irrelevant variables that should have no bearing on legal decisions. It cannot be ruled out that incorrect decisions can also be made by a check-out operator due to intimidation by a customer, desire to avoid public disorder and pressure to ensure quick transactions. A study by Willner[5], found that shopkeepers misjudged 38% of 16-year-old boys and 56% of 16-year-old girls to be of legal drinking age, i.e. at least 18 years old. In-house at Innovative Technology Ltd (ITL) a sample study was performed where staff were asked whether a subject was under 18 or not. On average, the human observer correctly identified 69% of the underage subjects. This means 31% of the subjects were incorrectly identified as over 18.

Human observers are prone to inherent bias and can be further influenced in their decision making by other factors. To protect children from accessing age restricted goods and services tools should be provided to the decision maker to help reach maximum compliance. An ideal tool would be an unbiased device, free from fatigue and pressure from the end-customer.

Test Results Summary

A full presentation and explanation of the results can be found in the Appendix.

1. Internal testing

Figure 1 below displays the Mean Absolute Error (MAE) per age. As expected, as we move away from the core age of focus (17,18 years) the MAE increases. Factors which contribute to this are both the challenges in estimating actual age from physical cues but also the training data utilised.

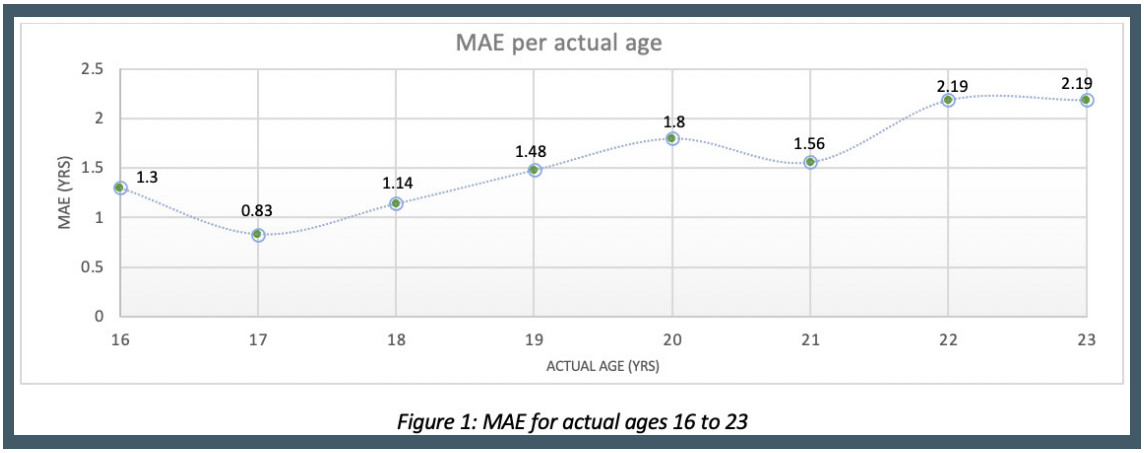

Figure 1: MAE for actual ages 16 to 23

These results are further discussed and broken down by gender and skin tone in the Appendix.

1 Rhodes, Matthew. (2009). Age Estimation of Faces: A Review. Applied Cognitive Psychology. 23. 1 - 12. 10.1002/acp.1442.
2 Moyse, E., Bredart, S. (2012) An own-age bias in age estimation of faces. European Review of Applied Psychology, Vol 62, Issue 1, 2-7 (https://doi.org/10.1016/j.erap.2011.12.002)
3 Fujisawa, T. , Azuma, Y. , Konishi, M. , Miyamoto, N. and Nagata, N. (2016) Age-Related Bias in Age Estimation Based on Facial Images of Others. Psychology, 7, 459-468. doi: 10.4236 psych.2016.74047.
4 Danziger, Levav, Avnaim-Passo (2011) Extraneous factors in judicial decisions, PNAS April 26, 2011 108 (17) 6889-6892; https://doi.org/10.1073/pnas.1018033108
5 Willner P, Rowe G. (2001) Alcohol Servers' Estimates of Young People's Ages. Drugs: Education, Prevention and Policy, 8:4, 375-383, DOI: 10.1080/09687630010019299

2. Independent certification

The algorithms which were used to produce the data presented in the previous section were also submitted for independent evaluation with the Age Check Certification Scheme (ACCS)[6].

Actual Age (yrs)	Estimated Age (yrs)	Delta (yrs)
18.21	18.02	-0.19

Table 1: Actual age vs estimated age

Actual Age (yrs)	ACCS MAE (yrs)	ITL MAE (yrs)
18.21	1.22	1.14

Table 2: MAE for actual age of 18.21 years

The Mean Predicted Age for the whole test crew is 18.02 against a Mean Actual Age of 18.21. It is, therefore, on average, underestimating age by 0.19 years.

The MAE is 1.22 years which is in broad agreement with the MAE of 1.14 years from our internal testing.

These results are further discussed in the Appendix.

Biometric Solutions

The core objective when developing our biometric solutions was to produce a low cost, effective device to aid workers in assessing the age of customers attempting to access age restricted goods. One such biometric solution for age estimation is ICU Lite, a total off-line solution with no recurring costs, which can be deployed in a fully or semi-automated setting. We believe that this device, will aid greatly in reducing the risk of children accessing age restricted goods, and protect both children and also the staff themselves.

ICU Lite estimates a person's age based solely on the facial features. ICU Lite utilises proprietary algorithms (based on artificial intelligence) which have been specifically trained to target the 15-25 age range to ensure maximum accuracy in this age cohort. The training data has been specifically and ethically sourced to be of sufficient diversity to combat any gender or ethically driven bias.

The quality and performance of any machine learning algorithm depends on several different factors. The choice of framework, the quality of the training data, the tuning of the algorithm and appreciating and mitigating other external factors. Age estimation has several challenging parameters so a multi-layered approach to the age check process is required. This is the approach we have taken for our biometric products – we employ our own proprietary algorithms, with our own focussed training datasets with a unique approach to tackle bias and/or overfitting of data.

6 https://www.accscheme.com/

Our training data has focused on the core range of 15 to 25-year-olds with an even distribution of gender and skin tone (as defined by the Fitzpatrick scale[7]). The core range was chosen, as a key application for ICU Lite is access to age restricted goods, which focus on the 18-year-old threshold in the UK. Coincidentally, this age range will also give the best opportunity to accurately estimate age, as the diversity of age cues across the wider population should be relatively far less than that of a growing child or an ageing adult.

The following appendix outlines the detailed analysis of both an independent test and certification (Age Check Certification Scheme – ACCS[8]) of our age estimation technology and includes results complied in ITL covering performance in ages around consent age (18 yrs. old) as well as examining gender and ethnic performance.

Conclusions & Discussion

In this paper we have outlined the challenges in anonymously estimating a person's actual age from physical cues in their appearance. Due to the disparity in the onset of these physical cues and dependence of those cues on external factors over time, we argue that with current approaches, the age range within early stage of adulthood presents the minimal divergence of these age cues the population as a whole. As a result, early-stage adulthood is the age at which using these physical cues would be most accurate.

The difficulty for a human in estimating age was also presented. The inherent bias for humans, as well as external factors such as fatigue and pressure, all contribute to poor performance in humans consistently and accuratly estimating age from physical appearance. To combat the frailties in humans estimating age, we have introduced ICU Lite – which is a device that uses artificial intelligence to estimate age from a subject's physical appearance. To minimise gender and skin tone bias, the algorithms were trained using a diverse and clean dataset. However, while the training set forms the foundation of a well performing algorithm it is not the only consideration to make. Bias can be present in every step of the estimation pipeline, from face detection, landmarking and classification. Each of these processes need to be addressed and it is vital that a deep understanding of the function and outcome of each step is gained. It is also important to understand other factors, and not to misinterpret this as a true inherent bias. These factors include lighting, expression, and extreme pose angles. Steps must be implemented to reduce the influence these factors have on the final classification.

Addressing the many factors which influence true bias and apparent bias, we presented some preliminary results. While further work is needed, we have shown very promising performance of the system across both gender and skin tone. Independent evaluation shows high accuracy of the system in estimating subjects of 18 years of age.

The core objective was to produce a low cost, effective device to aid staff when assessing the age of customers attempting to access age restricted goods. This can be deployed in a fully automated setting or semi-automated. We believe that this device, will aid greatly in reducing the risk of children accessing age restricted goods, and protect both children and the staff themselves.

7 Fitzpatrick, T. (1988) The Validity and Practicality of Sun-Reactive Skin Types I Through VI. Archives of Dermatology 1988; 124 (6): 869–871
8 https://www.accscheme.com/

Appendix

Methodology & Results

Two sets of results are presented in this paper. Firstly, the results of an independent evaluation of the system for deployment in a Challenge 25 scenario, which also includes some preliminary indications of potential bias. We also present the preparatory testing performed internally to ensure the system was ready for independent evaluation.

To minimise gender and skin tone bias, the algorithms were trained using a diverse and clean dataset. However, while the training set forms the foundation of a well performing algorithm it is not the only consideration to make. Bias can be present in every step of the estimation pipeline, from face detection, landmarking and classification. Each of these processes need to be addressed and it is vital that a deep understanding of the function and outcome of each step is gained. It is also important to understand other factors which influence the algorithm, and not to misinterpret this as a true bias. These factors include lighting, expression, and extreme pose angles. Steps must be implemented to reduce the influence these factors have on the final classification.

1. Internal testing

The core foundation for improved performance, regarding accuracy and combatting bias, is a diverse and clean training dataset. However, this is not the only requirement. Several layers of intelligence also need to be deployed to minimise errors in age accuracy across gender and skin tone. The test images presented are not included in the training data.

Figure 2: Illustration of the Fitzpatrick scale of skin tones

The core application for age estimation is controlling access to age restricted goods. For this reason, we concentrate our training and testing in the age range age 15 to 25 year-olds. The test datasets are equally distributed for gender and skin tone. The Fitzpatrick scale and definition can be seen in figure 2 above.

Each test image was manually labelled according to age, gender, skin tone, pose, lighting, glasses, beard, and any additional comments.

Skin Tone was divided into 3 categories

- Skintone I = Type 1 & 2
- Skintone II = Type 3 & 4
- Skintone III = Type 5 & 6

This ranges from lightest skin tone to darkest skin tone. This definition was chosen to maintain consistency with the independent validation by the ACCS.
Approximately 2K images constitute the test set with an even distribution of gender and skin tones.

Some example images are presented below...

Figure 3: Example of test images presented to ICU

The data we present will use the metric of the Mean Absolute Error (MAE) as an indicator of performance. This is the absolute difference between the estimated age and the actual age, averaged over the entire test set.

We will also present the MAE per gender and skin tone. For example:

1) MAE (All Males)
= Average {MAE (MALE Skintone I), MAE (MALE Skintone II), MAE (MALE Skintone III)}

2) MAE (All Females)
= Average {MAE (FEMALE Skintone I), MAE (FEMALE Skintone II), MAE (FEMALE Skintone III)}

3) MAE (ALL)
= Average {MAE (All Males) , MAE (All Females)}

The images are in the wild, which means there is a cross section of pose angle, face sizes, lighting conditions, facial expressions and facial occlusions and image quality (resolution and contrast). The effects of other factors such as pose, lighting and expression will not be explicitly presented in this paper but its influence may be alluded to.

All subjects and gender

Table 3 presents the results for all subjects and is subdivided for all males and all females across the age range of interest.

There is a tendency for the MAE to increase as the actual age moves away from the core age of 17/18. This can be both down to the points raised earlier regarding the divergence of age cues, but also down to the distribution of the training data.

The difference between the genders is minimal, which reflects the training data been equally distributed between male and female. In fact, preliminary data indicates that the disparity may be due to differences in image quality of the test data (i.e., facial occlusion, shadow across face, extreme expression) rather than a fundamental bias between the genders.

Actual Age	All Subjects	Male	Female
16	1.35	1.40	1.30
17	0.86	0.88	0.83
18	1.14	1.14	1.14
19	1.48	1.47	1.48
20	1.74	1.68	1.80
21	1.46	1.37	1.56
22	2.19	2.18	2.19
23	2.21	2.23	2.19

Table 3: MAE for gender

The data can be easily visualised in figure 4 and figure 5 below.

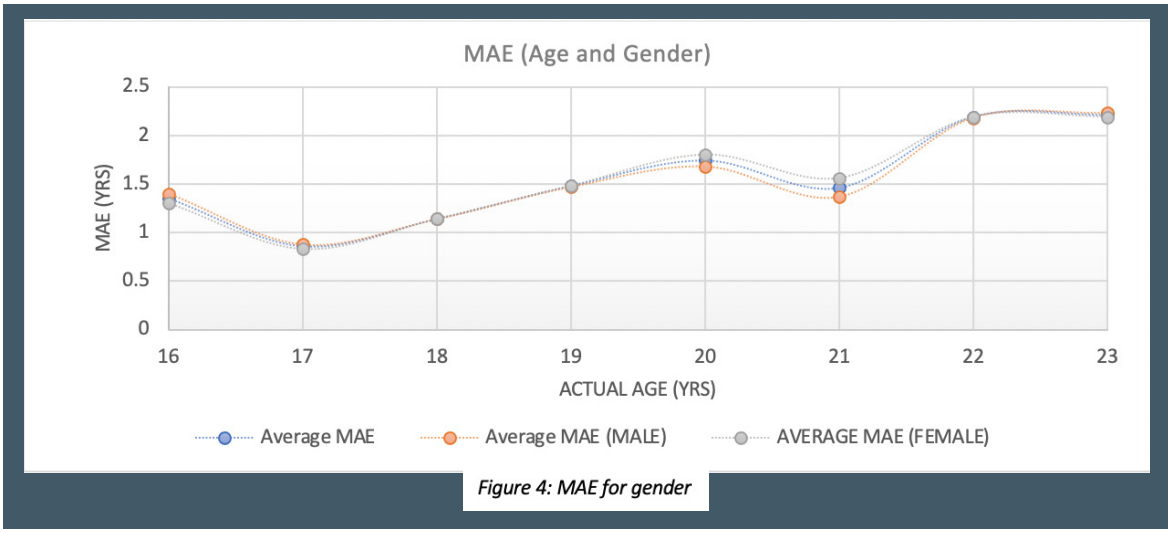

Figure 4: MAE for gender

Figure 5 presents the data for gender only across the age range. While a slight disparity (0.12yrs, 0.19yrs) occurs at both 20 and 21 years old we do not believe this is because of fundamental bias, but rather the image quality presented to the camera. It should also be noted, that for the test data we only know the users age in years. We do not know the age down to the resolution of months. This may also add to the disparity – for example we treat a subject of 18ys 1 month and 18yrs 6 months as both being an actual age of 18 yrs.

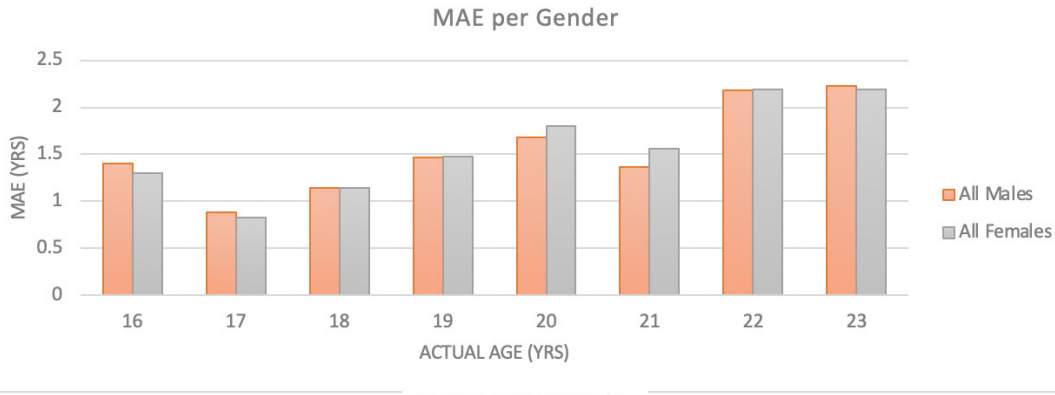

Figure 5: MAE per gender

All subjects and skin tone

In this section we will present the MAE for skin tone. This data is presented in table 6 and illustrated in figures 6 and 7.

Actual Age	All Subjects	I	II	III
16	1.35	1.28	1.4	1.38
17	0.86	1.19	0.8	0.58
18	1.14	1.16	1.2	1.01
19	1.48	1.27	1.77	1.38
20	1.74	1.8	1.48	1.95
21	1.46	1.31	1.38	1.71
22	2.19	2.12	2.2	2.2
23	2.21	2.35	1.98	2.31

Table 6: MAE per Skin Tone

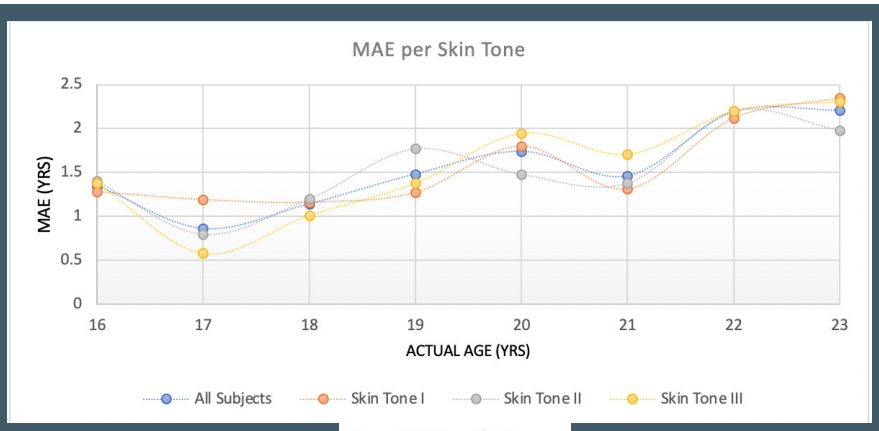

Figure 6: MAE per Skin Tone

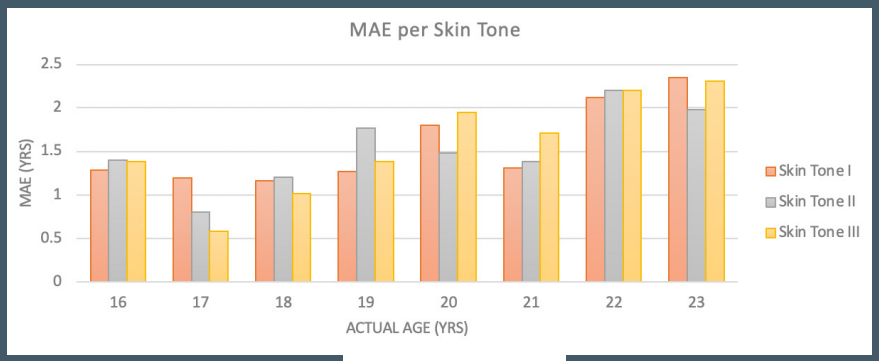

Figure 7: MAE per Skin Tone

The data shows some more variation than in the gender alone – this is most apparent for 17 year olds, where there is a maximum difference of 0.61 years between light skinned (1.19 yrs) and dark skinned (0.58 years). While this is still less than the 1-year resolution of the test data further investigation is required to determine the disparity.

However, the good performance on darker skin tone does reflect the increased darker skinned subjects in the training data.

MAE for gender and skin tone

In this section we present the results for genders for all skin tones. The data is presented in table 5 and illustrated in figure 8.

Skin Tone		I		II		III	
Gender		M	F	M	F	M	F
Age	**16**	1.24	1.31	1.53	1.26	1.42	1.33
	17	1.08	1.31	0.73	0.86	0.82	0.33
	18	1.3	1.02	1.00	1.39	1.11	1.00
	19	1.44	1.1	1.5	2.05	1.48	1.28
	20	1.95	1.65	1.48	1.47	1.6	2.29
	21	1.62	1.00	1.25	1.5	1.25	2.17
	22	2.33	2.00	2.00	2.40	2.22	2.17
	23	2.44	2.25	2.00	1.96	2.26	2.36

Table 5: MAE for gender and skin tone

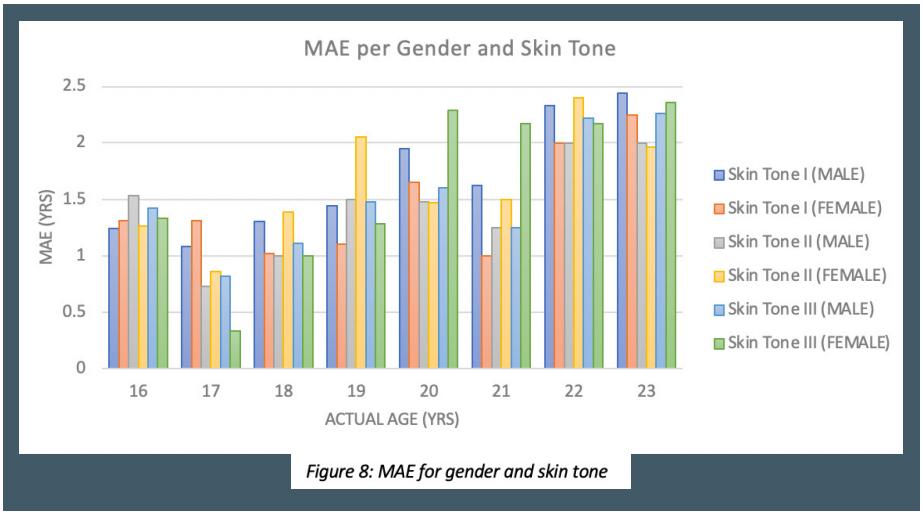

Figure 8: MAE for gender and skin tone

The data shows some more variation than in the gender alone – this is most apparent for 17 year olds, where there is a maximum difference of 0.98 years between light skinned female (1.31 yrs) and dark skinned females (0.33 years). While this is still less than the 1-year resolution of the test data further investigation is required to determine the disparity.

Independent certification

The algorithms which were used to produce the data presented in the previous section were also submitted for independent evaluation with the Age Check Certification Scheme (ACCS)[9] . The datasets were not known or seen by age estimation device prior to the test. There are approximately 300 images, but mostly presented as passport style presentations – i.e., no extreme pose angles and mostly neutral expressions. All images were of high quality and under a controlled lighting environment. The results are summarised in the following tables.

Actual Age (yrs)	Estimated Age (yrs)	Delta (yrs)
18.21	18.02	-0.19

Actual Age (yrs)	MAE (yrs)
18.21	1.22

Table 6: Key findings from ACCS evaluation

The Mean Predicted Age for the whole test crew is 18.02 against a Mean Actual Age of 18.21. Therefore, on average, underestimates age by 0.19 years.

The MAE for the actual age of 18 yrs is 1.22 yrs. For our internal testing we calculated a MAE of 1.14 yrs. The difference in the results may be accounted for by the difference in the cohort of the test data. In our internal testing, we had a greater proportion of skin tone III and as shown in table 4 have a greater performance for skin tone III. This will contribute to a better overhaul performance. Additionally, the resolution of the ages in our test data (in years and not months) may also contribute to some disparity. However, in general the MAE for both tests seem to be in agreement.

None of the results were above the absolute tolerance level (Age 25).

9 https://www.accscheme.com/

Gender Bias

The sample size for this test was insufficient to determine the absence of gender bias to an acceptable level of reliability, however it may give an indication of performance of the system. Further independent testing, with sufficient sample size will be performed in the future. The sample consisted of 260 test crew images of which 139 (53.5%) were Female and 121 (46.5%) were Male.

The results based on the sample size of this test show:

	AGE (yrs)	MAE (yrs)
Mean Predicted Age (Whole Sample)	18.02	1.22
Mean Predicted Age (Female)	17.71	1.26
Mean Predicted Age (Male)	18.38	1.17

Table 7: Preliminary gender performance

While the test results are based on gender sample sizes that are too small to extract statistically valid analysis there is close agreement with internal testing (1.17 vs 1.14 and 1.26 vs 1.14). The trend indicates that males are predicted as older than females, however males have a slightly better MAE. While improvements can be made, there is little evidence to suggest any clear and definite bias in gender. Further independent testing must be carried out to verify.

Skin Tone Bias

The sample size for this test was insufficient to determine the absence of skin tone bias to an acceptable level of reliability, however in combination with internal testing it can be used to determine if any trends appear.

The current sample consisted of 260 test crew images of which 182 (70%) were Skin Tone I; 55 (21.2%) were Skin Time II; 23 (8.8%) were Skin Tone III.

	AGE (yrs)	MAE (yrs)
Mean Predicted Age (Whole Sample)	18.02	1.22
Mean Predicted Age (Skin Tone Type I)	18.00	1.27
Mean Predicted Age (Skin Tone Type II)	18.21	1.08
Mean Predicted Age (Skin Tone Type III)	17.79	1.13

Table 8: Preliminary skin tone performance

The informal results from the independent test have similar results as the internal testing (1.16 vs 1.27, 1.2 vs 1.08, 1.01 vs 1.13). Disparities may be linked to sample size and resolution in actual age. While further testing is required, the results look promising.