From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation

Albert Clapés
Computer Vision Centre and
University of Barcelona, Barcelona, Spain
aclapes@cvc.uab.cat

Ozan Bilici, Dariia Temirova, and Egils Avots
iCV Lab, University of Tartu
Tartu, Estonia
{ozan,dariia,ea}@icv.tuit.ut.ee

Gholamreza Anbarjafari
iCV Lab, Univ. of Tartu, and GoSwift Inc., Estonia
Hasan Kalyoncu University, Turkey
shb@icv.tuit.ut.ee

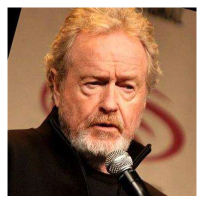
Sergio Escalera
Computer Vision Centre and
University of Barcelona, Barcelona, Spain
sergio@maia.ub.es

Abstract

Real age estimation in still images of faces is an active area of research in the computer vision community. However, very few works attempted to analyse the apparent age as perceived by observers. Apparent age estimation is a subjective task, which is affected by many factors present in the image as well as by observer's characteristics. In this work, we enhance the APPA-REAL dataset, containing around 8K images with real and apparent ages, with new annotated attributes, namely gender, ethnic, makeup, and expression. Age and gender from a subset of guessers is also provided. We show there exists some consistent bias for a subset of these attributes when relating apparent to real age. In addition we run simple experiments with a basic Convolutional Neural Network (CNN) showing that considering apparent labels for training improves real age estimation rather than training with real ages. We also perform bias correction on CNN predictions, showing that it further enhance final age recognition performance.

1. Introduction

Automatic age estimation is a challenging computer vision problem [14, 15, 5] with applications in biometrics [31], human-robot interaction [39, 6], personalised advertisement [28], and personality analysis [40], just to mention a few. However, ageing is a variable-paced process depending on each person's genetics and other physiological factors [23]. Even for humans it is a difficult task to precisely determine other people's chronological age from observed visual ageing signs. Our best guess will be an estimate of others' *apparent age*, which in turn is likely to be biased by differences in gender, ethnicity, culture, and



(a) A: 55.00, R: 75, Diff: -19.98

(b) A: 21.28, R: 30, Diff: -8.72

(c) A: 27.69, R: 19, Diff: +8.69

(d) A: 37.46, R: 53, Diff: -15.60

(e) A: 44.28, R: 32, Diff: +18.28

(f) A: 71.00, R: 55, Diff: +16.00

Figure 1: Examples of real-apparent age biases on APPA-REAL dataset [10, 11]. Apparent (A), real age (R), and difference A-R (Diff) are shown for each face image.

age, among others. Despite those biases, apparent age better correlates with physical appearance and hence it is easier to estimate from visual information [1].

From a computer vision perspective, age estimation is often posed as a feature representation and regression problem. While earlier works just focused on real age prediction [21, 45, 8, 25], many recent ones shifted to apparent age estimation [36, 22, 47, 26, 1], especially after the apparition of APPA-REAL dataset[1] [10, 11]. From the work of [1] on this dataset, several conclusions are drawn on apparent age:

[1]ChaLearn LAP (2015-2016) provided a dataset of faces with both real and apparent labels annotated by human observers. http://chalearnlap.cvc.uab.es/dataset/26/description/

(1) it is easier to predict than real age, and (2) it enhances real age estimation. Hence, improving the apparent labels by taking care of biases would potentially improve both real and apparent age prediction.

Age predictions can greatly differ from their true label (see Fig. 1). Two main categories of bias are identified in apparent age estimation: the ones inherent to the target subject (target-bias) and those introduced by the apparent age guessers (guess-bias). Among the first category one may consider, for instance, the bias introduced by makeup covering age signs such as age spots or wrinkles [17]. Within the second, one might find that apparent age guessers perform worse on estimating the age of target subjects from the opposite gender. The graphic distinction between the two biases is depicted in Fig. 2.

In this work, we provide additional attributes to the APPA-REAL dataset: gender, ethnicity, level of makeup, time of the photo, and facial expression. We analyse the bias these attributes introduce when relating apparent to real age. We run a baseline CNN showing that apparent labels enhance real age recognition performance rather than training with real age labels. Furthermore, we perform bias correction on CNN predictions based on the modelled analysed biases. As a result, we show that there exist some consistent bias introduced by those attributes and that their correction further enhance age recognition performance.

The rest of the paper is organised as follows: Section 2 discusses related work. Section 3 discusses the details of the provided dataset. Section 4 explains the analysed biases. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

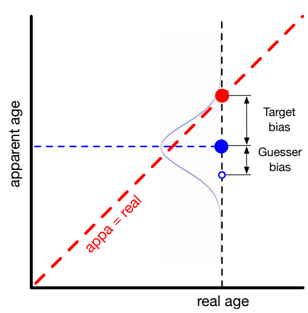


Figure 2: Target- and guess-bias. Target-bias is the difference between apparent age (i.e. the mean value of age guesses) and real age, whereas guess-bias is the disagreement of one guesser's guess and average guess value.

2. Related work

Here, we review state-of-the-art methods for real and apparent age estimation emphasising those that take care of some kind of bias. Then we review some relevant studies on bias analysis for age estimation. Finally, we summarise public available datasets for age estimation.

2.1. Real age estimation methods

Most current approaches for both real and apparent age estimation rely on deep learning based methods. The early work of [21] proposed a relatively shallow CNN architecture to classify age into rough age groups from OUI-Adience [9]. The work of [30] addressed the non-stationary property of ageing by casting it to an ordinal regression problem that they transformed into a series of binary classification problems solved by a Multiple Output CNN. [16] performed age estimation via age difference.

Real age estimation, as in the case of apparent age estimation, involves target biases: those that are intrinsic to target's visual face appearance; those that depend on gender, ethnicity, face expression, and so on. The work of [46] deals with the biases introduced by gender and ethnicity by posing the age estimation problem as a multi-task classification problem. In [24], expression-invariant age is estimated using structured learning.

2.2. Apparent age estimation methods

Even though many studies have focused on real age estimation, apparent age estimation is still in its infancy stage. Deep EXpectation of Apparent Age From a Single Image(DEX) [36], which uses the CNN VGG-16 [42], was the winner of ChaLearn LAP 2015 apparent age estimation challenge [10]. They considered the problem as a classification problem between 0 to 100 years old. The IMDB-WIKI [36] dataset was created with images crawled from IMDB and Wikipedia, and these data were used to fine-tune a VGG-16 model pre-trained on ImageNet [38]. Then, they split the ChaLearn LAP 2015 [10] dataset into 20 different groups, and fine-tuned 20 models using 90% of each group for training and for validation.

AgeNet [22] considered the problem as both classification and regression. They trained real value-based regression and Gaussian label distribution-based classification models. Both used large-scale deep CNN. First, they pre-trained the network using a face dataset collected from the Worldwide Web with identity labels. Afterwards, they fine-tuned it with a real age dataset with noisy age labels, and with an apparent age dataset which was provided by ChaLearn LAP 2015 [10]. Although Zhu et al. [47] applied CNNs as well, their purpose was different - CNNs was employed for feature extraction. Then support vector machine (SVM), support vector regressor, and random forests were used for final apparent age estimation.

In the second round of the ChaLearn competition [11], the number of images was augmented from 5K to around 8K face images. The age distribution of the dataset was also changed, especially, the percentage of the children images included was significantly increased. The winners [11, 3] fine-tuned two separate CNNs, one for all age labels applying label distribution encoding [11], and other just for children between 0 to 12 years old. First, the test data were used in the first CNN. The second CNN was only used in the case the first prediction was not above 12 years old.

Refik et al. [26] adopted the method proposed in [36]. However, instead of using a single label, they split data into three age groups, and created three different models accordingly. The average prediction from the three models was used for estimating final apparent age.

2.3. Studies on bias in age recognition

Target biases involved in age prediction from face images have been studied in different computer vision works. For real age prediction, gender and ethnicity bias was analysed in [46] and age bias in [24]. However, the utilisation of apparent age labels demands visiting this fresh concept that is the guess-bias. To the extent of our knowledge, there are no apparent-age previous works on this subject, yet we can find that it has been discussed in other areas. In psychology, [44] studied the determinants and biases in age estimation across the adult life span. Their investigation on more than 2,000 face images revealed age estimation ability decreases with age. The study also showed nonetheless older people are more accurate guessing ages from older adults than younger adults on that same age range. In contrast, they found the gender of the guesser did not make any significant difference. They also analysed target biases: older people faces are more difficult to estimate, and facial expression influences the guess (neutral faces are more easily estimated, whereas age of happy faces tend to be underestimated).

One of our goals is, then, to present a preliminary study on dealing with guess-bias and demonstrate the influence in age estimation also in relation to various target-biases.

2.4. Age recognition datasets

There are just few available age databases with a substantial number of labelled face images. Table 1 shows their summary [1]. In this work we augment APPA-REAL database annotations, the only one available containing a large number of both real and apparent ages by introducing an additional set of attributes (see Section 3).

3. Dataset

The APPA-REAL database [1] contains 7,591 images with real and apparent age labels collected based on the opinion of many subjects using a crowd-sourcing data collection and labelling application based on Facebook API, data from the AgeGuess platform[2] and Amazon Mechanical Turk (AMT) workers. The total number of apparent votes is around 250,000. On average it contains around 38 votes per image, which makes the average apparent age very stable (0.3 standard error of the mean). The images are split into $4,113$ train, $1,500$ valid and $1,978$ test images.

Table 1: Age-estimation related datasets [1].

Database	# of faces	# of subjects	Age range	Age type	Environment
FG-NET [20]	1,002	82	0-69	Real Age	Uncontrolled
GROUPS [13]	28,231	28,231	0-66+	Age Group	Uncontrolled
PAL [27]	580	580	19-93	Age Group	Uncontrolled
FRGC [33]	44,278	568	18-70	Real Age	Partly Controlled
MORPH2 [35]	55,134	13,618	16-77	Real Age	Controlled
YGA [12]	8,000	1,600	0-93	Real Age	Uncontrolled
FERET [34]	14,126	1,199	-	Real Age	Partly Controlled
Iranian face [4]	3,600	616	2-85	Real Age	Uncontrolled
PIE [41]	41,638	68	-	Real Age	Controlled
WIT-BD [43]	26,222	5,500	3-85	Age Group	Uncontrolled
Caucasian Face Database [7]	147	-	20-62	Real Age	Controlled
LHI [32]	8,000	8,000	9-89	Real Age	Controlled
HOIP [18]	306,600	300	15-64	Age Group	Controlled
Nis Web-Collected Database [29]	219,892	-	1-80	Real Age	Uncontrolled
OUI-Adience [9]	26,580	2,284	0-60+	Age Group	Uncontrolled
IMDBWIKI [37]	523,051	20,284+	0-100	Real Age	Uncontrolled
APPA-REAL [1]	7,591	7,000+	0-95	Real and Apparent Age	Uncontrolled

In this work, the database has been enriched by adding further attributes: ethnicity (namely caucasian, asian, african/afro-american), age of the image (namely old photo or modern photo), existence of makeup (namely very subtle, no makeup, makeup, and not clear), and facial expression (namely neutral, slighty happy smile, happy, and other). Fig. 3, 4, 5, 6, and 7 show some visual examples of those new categories. Table 2 shows statistics of the intersection pairs of the new attributes for the APPA-REAL dataset.

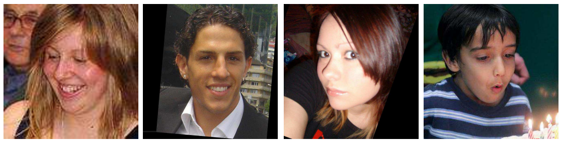
While all images were labelled by one person and most of the categories are non-subjective, there are still some that are difficult to determine (i.e. makeup and expression). Nonetheless, the fact that the annotation was done by only one person ensures labelling consistency.

4. Bias analysis

In this section we show the apparent-real relations in the dataset based on the different new attributes we provide and the meta-information of apparent age guessers. For each real-age value there can be several subjects with different guessed apparent age. This sample of subjects is a distribution we represent by their mean and standard deviation.

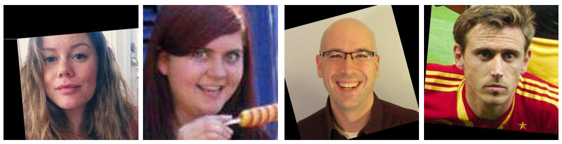
Fig. 8(a) shows the correlation between real and apparent age estimates along x- and y-axis. We can observe there is a tendency of overestimating apparent age with respect to real age in the range [10,30) years, in contrast to the underestimation in the range [30,100]. Another trend is the smaller variance at younger ages of [8,25]. From there on, the variance keeps increasing. Although at older ages the lack of data causes small samples and hence distributions to be poorly estimated.

[2] http://www.ageguess.org/



(a) (b) (c) (d)

Figure 3: The happiness attribute categories: (a) happy; (b) slightly happy; (c) neutral; and, (d) other.

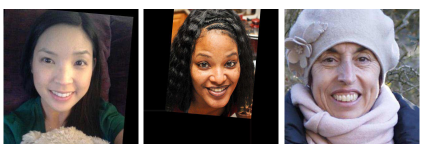


(a) (b) (c) (d)

Figure 4: Examples of the gender attribute categories: (a) and (b) show female ; (c) and (d) show male gender.

(a) (b) (c) (d)

Figure 5: The makeup attribute categories: (a) makeup; (b) no makeup; (c) not clear; and, (d) very subtle makeup.



(a) (b) (c)

Figure 6: The ethnicity attribute categories: (a) asian; (b) afroamerican; (c) caucasian.

(a) (b)

Figure 7: Time of photo category: (a) old; (b) modern.

Next, we analyse the target-bias in the apparent-real age relation for the new introduced attributes. Then, we analyse the same behaviour from the point of view of the age and gender of the guessers on the subset of data providing this information. Finally we show how apparent age groundtruth and target-bias correction over predictions further improves real age prediction performance.

4.1. Target-bias

We next analyse target biases introduced by the new attributes in augmented annotation on APPA-REAL. We discuss how gender, ethnicity, makeup, the time the photo was taken, or expressed happiness affects age guessing in humans.

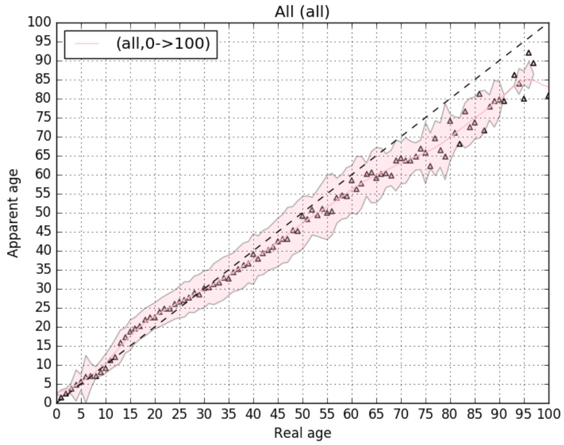
We first consider the target gender influence on guessing apparent age. In Fig. 8(b), we see ”female” category presents a considerably larger bias between real and apparent ages guesses than ”male”. Males' apparent age in that range is always closer to real age. One can clearly observe an overestimation-to-underestimation point shifts from 25 when not distinguishing gender (see Fig. 8(a)) to 20 in the case of female and 35 in male. That is, females apparent age is overestimated in the range [0,20] and later underestimated, while in the case of males the overestimation lasts until 35. In the case of females, there is the interval [13-18] in which their age is consistently overestimated +5 years. Interestingly it is only for 77 years old people (and older) that males' apparent age is more biased (underestimated) than females'. Some visual examples showing large biases for male and female in the data set are shown in Fig. 1.

Besides gender, ethnicity also plays an important role on apparent age. In the most populated range, i.e. [15-55], we notice the more pronounced and constant underestimation of apparent age in Asian population in relation to caucasians and afroamericans (shown in Fig. 8(c)) up until a very short interval ranging between [57,63]. The latter two also present differences. Afroamericans apparent age is generally more biased than caucasians'. In the ages ranging from 0 to 25, afromericans' age is overestimated w.r.t. Caucasians. From 25+ years, Caucasians are less biased than afroamericans. For all ethnicity the apparent age is overestimated on younger and underestimated on older ages. This seems to be a trend independently from gender and ethnicity categories. Unfortunately, a more rigorous analysis on this category is unfeasible since we do not have information on guessers' ethnicity; if the sample of guesses is not balanced, we cannot decouple target- from guess-bias easily.

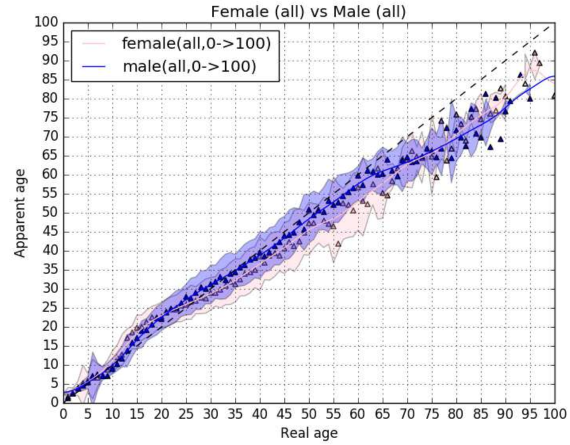
It is often said makeup makes people look younger. In Fig. 8(d), we can see this is true only for people older than 27. At that age is when the first age signs, e.g. wrinkles, start to appear. The masking effect of makeup makes that features less obvious, so apparent ages tend to be underestimated respect to real age. This is also true for subtle makeup, which has a similar effect until the age of 52. Subjects with no makeup instead, present a much smaller bias. The fact that makeup is worn more often by younger adults than teenagers, causes people younger than 24 to appear older. Since makeup is not normally worn by children, we consider those highly deviated points [0,10] outliers. Fig. 1(c) shows one example of a large makeup bias.

The time of the photo was taken also introduces a clear bias. Guessers tend to overestimate apparent age of people on old photos, as seen in Fig. 8(e). One visual example is shown in Fig. 1(f). Unfortunately, the very few photos available in the ”old photo” category in contrast to ”modern photo” impedes us to develop further analyses on this.

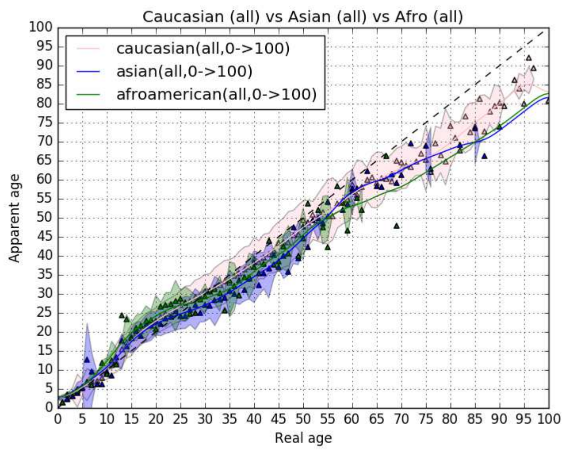
Last but not least, we show how exhibiting happiness can



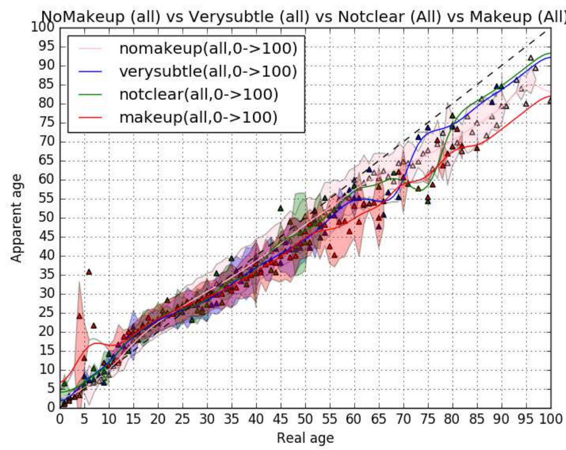
(a)



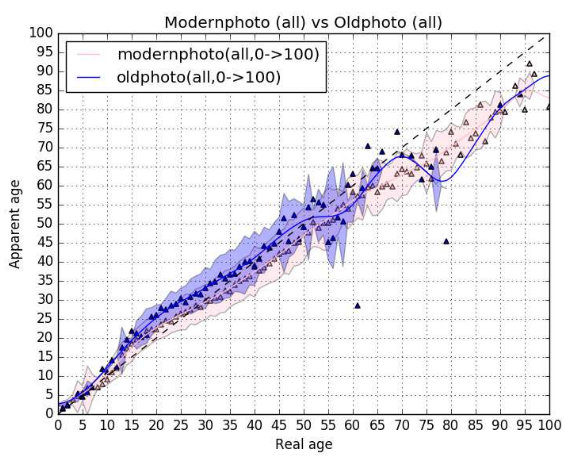
(b)



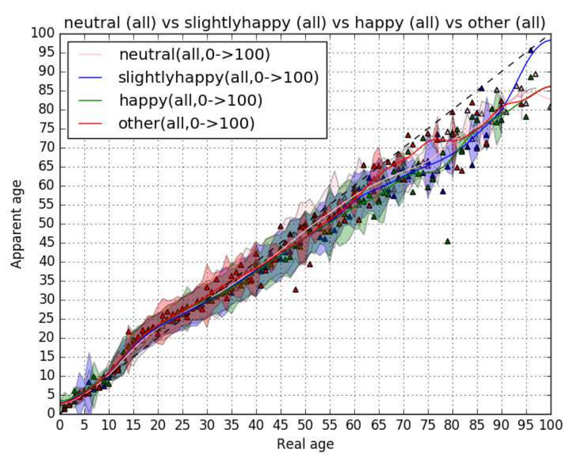
(c)



(d)



(e)



(f)

Figure 8: Analysis on target-biases on APPA-REAL dataset. We illustrate the relation among mean apparent ages of subjects and real age (a) and also the analysis of 5 target biases (b)-(f). The dotted diagonal is the "apparent = real" age line. Triangles (△) indicate the mean apparent age across subjects of a particular real age; curved lines (−) are a linear interpolation of mean apparent values (triangles) smoothed after convolving a 3-year mean kernel; and shadowed areas illustrate the standard deviation across subjects' apparent age.

Table 2: Counts on attribute categories of the augmented APPA-REAL grountruth.

	female	male	caucasian	asian	afroamer..	nomakeup	verysubtle	notclear	makeup	modernph..	oldphoto	neutral	slightly..	happy	other
female	3773	0	3285	366	122	1345	658	126	1644	3464	309	931	1562	1128	152
male	0	3818	3401	308	109	3721	41	39	17	3526	292	1749	1221	589	259
caucasian	3285	3401	6686	0	0	4437	596	144	1509	6116	570	2365	2449	1499	373
asian	366	308	0	674	0	479	87	13	95	665	9	234	276	144	20
afroamer..	122	109	0	0	231	150	16	8	57	209	22	81	58	74	18
nomakeup	1345	3721	4437	479	150	5066	0	0	0	4756	310	2044	1775	931	316
verysubtle	658	41	596	87	16	0	699	0	0	665	34	147	279	247	26
notclear	126	39	144	13	8	0	0	165	0	142	23	53	57	46	9
makeup	1644	17	1509	95	57	0	0	0	1661	1427	234	436	672	493	60
modernph..	3464	3526	6116	665	209	4756	665	142	1427	6990	0	2377	2638	1605	370
oldphoto	309	292	570	9	22	310	34	23	234	0	601	303	145	112	41
neutral	931	1749	2365	234	81	2044	147	53	436	2377	303	2680	0	0	0
slightly..	1562	1221	2449	276	58	1775	279	57	672	2638	145	0	2783	0	0
happy	1128	589	1499	144	74	931	247	46	493	1605	112	0	0	1717	0
other	152	259	373	20	18	316	26	9	60	370	41	0	0	0	411

affect age guessing. Fig. 8(f) supports the fact that neutral faces introduce less bias than other facial expressions. We hypothesise this might be due to the fact that it prevents the apparition of wrinkles that might be confused with age-caused wrinkles on older people faces, as discussed in [2].

4.2. Guess-bias

We use the gender of the guessers provided along with apparent age guesses in APPA-REAL. The dataset consists of exactly 260,656 guesses, from which 49,652 are from males, 60,827 from females, and the remaining 150,177 ones are undefined.

Yet it is true that females introduce more target-bias than males, they are also more accurate at guessing other people's real age (see Fig. 9(a)). In particular, they are better than male in all cases: estimating ages from other females (Fig. 9(b)) and males (Fig. 9(c)).

We also considered exploiting the age information of guessers provided in APPA-REAL. Psychology studies suggest age perception is affected with age [44]. However, a comprehensive study on this area is not feasible due to the little number of guesses from 60-year old people and above (less than 3% of total apparent age guesses in the dataset). In this studies it is also stated there is not a significant difference among genders in age guessing. Yet it is true there is no statistically significant differences between gender, we see there's a trend observable from the plots that suggest females could be better at the task.

4.3. Correction of biases

The key idea behind our bias correction is to shift apparent age towards their corresponding real age value. The illustrative examples from Fig. 1 show how some subjects' apparent age greatly differs from their true age. Those examples, that can be considered as outliers, may harm the age prediction performance. For instance, Fig. 1(a) is underestimated by almost -20 years respect to his true age. Fig. 8(b) shows how older males age tends to be underestimated. We discuss here how can we perform these corrections to shift predicted apparent age to approximate better the real one.

For target bias correction, we first divide the subjects into the set of mutually-exclusive attribute categories defined by our augmented ground-truth (see Table 2). Then for a fixed partition we re-compute the mean apparent age for each real age. In order to deal with missing values in very small partitions or not having big shift differences on subsequent mean apparent ages, we apply 1D-linear interpolation and an iterative mean filter of size 3. The result of this are the fitted curves illustrated in Fig. 8-9. Each of them is a 100-dimensional vector that we subtract to the vector $[1, 2, \ldots, 100]$. We finally obtain the corrected apparent ages applying this shift to the subjects that fall in that partition. Following this strategy, different biases are decoupled and addressed.

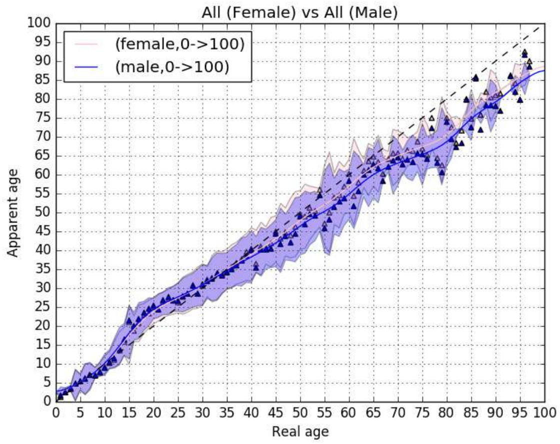
Similarly, we perform correction of biases introduced by guessers, i.e. guess-bias. In this case, the correction would be applied not on apparent ages of subjects but on guesses when the guesser gender information is available, prior to mean apparent age label for each subject. Those biases are shown in Fig. 2. Fig. 10 shows how after applying bias correction in Fig. 9(a) the distribution of age guesses from annotators presents less bias and variance. In the next section, we show the effect on applying these bias corrections on apparent age for predicting real age.

5. Experimental evaluation of biases

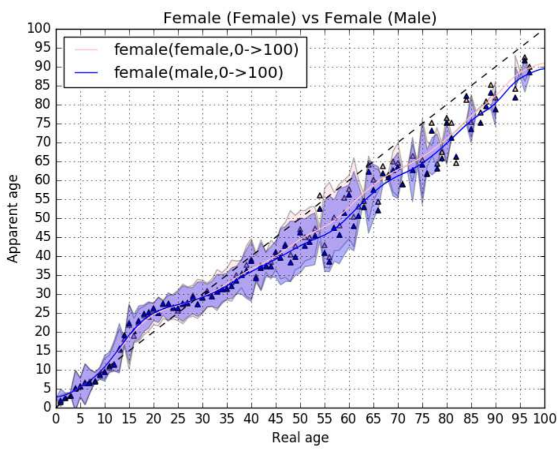
In this section, we first describe the experimental setup used for the experiments. Then we run initial baseline results with a simple CNN model for age estimation using both real and apparent labels for training. Then we perform a series of bias corrections on the obtained CNN predictions before computing final real age estimation.

5.1. Experimental setup

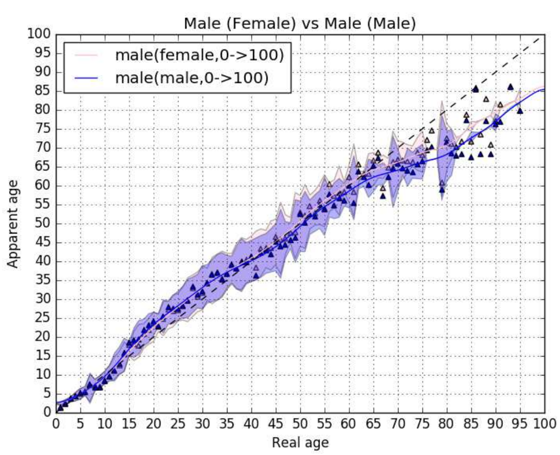
For experiments we join the train and validation data of the APPA-REAL dataset for training and use the test set for testing. We use the public images provided in [1] with cropped and non-rigid aligned faces for training and testing in our experiments. No additional pre-processing is performed. For each experiment, we either train on real or apparent age ground-truth, whereas during prediction we always evaluate by comparing to real age labels. We compute the error metric as the mean absolute error between the predicted and the real age. We use the AlexNet CNN model for baseline experiments [19]. We train it from scratch and without any data augmentation. For training we defined 101



(a)

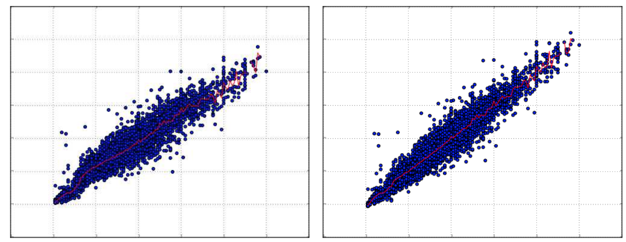


(b)



(c)

Figure 9: Guesser biases on APPA-REAL. Using gender of guessers: (a) who are better guessing the age of others (either females or males)?; (b) who is better guessing females' age?; (c) who is better guessing males' age?



(a) Uncorrected guess-bias (b) Corrected guess-bias

Figure 10: Correction of apparent age guesses. Bias and variance of age guesses is reduced and better adjusted to the diagonal real = apparent.

classes as output in range between 0 and 100. Learning rate was fixed to 0.005, batch size to 32, and number of iterations to 1000. Here, we are not taking care of obtaining accurate regression ages but on experimentally testing the bias effect on a simple deep model (given the limited amount of data) at the same conditions for all experiments. All the training were run on a Nvidia Titan XP GPU.

5.2. CNN apparent-real for real age estimation

For baselines (B#1) and (B#2), we trained the network on real and apparent ages labels respectively. For B#3, we trained the network with apparent ages of males and real ages of females and vice versa in B#4. Table 3 shows the results of those baselines in terms of mean absolute error (MAE) between predicted and real-age groundtruth. B#2 is the best performing baseline followed by B#4. In the case of B#4, apparent ages are used for the case of female (that often present larger biases than males 8(b)). B#1 presents the worst performance, being only better in "not clear" (make up). B#2 is the one doing better in highly biased categories, such as "makeup", "happy", or "oldphoto". The decrease of roughly 1 point in MAE in B#2 with respect to B#1 confirms apparent age serves to predict real age better: when trained on apparent ages, the network observes lower variability on visual appearance for people with similar age label. Also the presence of outliers is reduced using apparent ages: a large difference between how someone looks and apparent age is less likely than with true age. These two factors will lead a classifier to less confusion, which means better generalisation and, hence, better performance.

5.3. Bias correction on predictions

From previous baselines, we choose B#2 to perform target-bias correction on its output predictions at test time. Applying the correction as previously explained in 4.3, we obtain more accurate estimations. From Table 4), we see corrections on "Makeup" and "Gender" biases showed better results, reducing the error to less than 12 points. The best correction ("Makeup lvl") reduced the prediction error 1.77 points, whereas the worst correction achieved a still

(a) A: 32.51, R: 36, P: 32, C: $32 + 3.16 = 35.16$

(b) A: 52.07, R: 60, P: 27, C: $27 + 10.14 = 37.14$

Figure 11: Network predictions and corrections. A = apparent age, R = real age, P = predicted age by the network, C = P + correction factor (= corrected prediction).

significant reduction of 1.35 points. One example of successful bias correction is shown in Fig. 11(a). Predicted age (32) is quite close to the apparent ground-truth (32.51) from which the network has learnt. However, the APPA-REAL bias would cause this prediction to contribute to a larger MAE. After 3.16 years correction, the corrected final prediction (35.16) is much closer to the real age (36).

However, although bias corrections show to improve in most of the cases, there are some other cases in which the network prediction is too far from being right. One visual example is shown in Fig. 11(b). In this case the correction is futile simply because of a bad pre-correction prediction (27) compared to either apparent (52.07) or real age (60) labels; in here, the +10.14 correction is insufficient.

Despite we discussed about both target- and guess-bias, the guess-bias correction cannot be applied to correct the predictions of the network as for target-bias, since it needs to de-bias individual guesses. Yet to validate our hypothesis, we measured MAE between apparent age and real age ground-truth, which we found to be 5.980. Then, we corrected apparent age ground-truth and measured MAE to find that it is reduced. Specifically, we tried the three kinds of corrections illustrated in Fig. 9 and found a slightly reduction of error: 5.660 (-0.320), 95.8632 (-0.117), and 5.8453 (-0.135), respectively.

6. Conclusion

In this paper, we augmented APPA-REAL with an additional set of attributes: gender, ethnic, makeup, time of the photo, and expression. We analysed the apparent to real age differences and found biases were introduced by the new attributes when relating apparent to real ages. We also introduced the guess-bias concept in terms of gender and age. For the experimental part, we ran simple CNN experiments that showed apparent labels can achieve better performance for predicting real age than directly training with real age. We also performed bias correction on the CNN predictions when trained with apparent labels before computing final real age estimation. We showed the modelled attribute biases when applied to correct the predictions improves final

Table 3: Results of four different age estimation baselines (all using the same AlexNet-CNN model). MAE is evaluated as a measure performance. '*' is best result per attribute category.

	B#1	B#2	B#3	B#4
female	15.3980	14.2843	14.4275	14.2814*
male	13.6942	12.8246*	13.3340	12.9280
caucasian	14.5215	13.6386	13.9710	13.623*
asian	15.1085	12.4729*	13.0388	13.3178
afro	14.9821	14.1607	13.5357*	14.4286
nomakeup	14.4652	13.9992	14.3415	13.7854*
verysubtle	13.8659	12.7439	11.5488*	12.8659
notclear	13.6333*	14.0667	14.8000	14.8000
makeup	14.9101	12.862*	13.3147	13.3659
modernphoto	14.9649	14.1520	14.3996	14.1235*
oldphoto	13.1982	11.5626*	12.1390	11.8815
happy	14.7334	13.3447*	14.1273	13.5263
slightlyhappy	14.5881	14.1248	13.5802*	13.8673
neutral	14.4868	13.6283*	13.8651	13.7249
others	14.2813	12.1875*	14.2891	12.5469
ALL	14.5728	13.5774*	13.8979	13.6259

Table 4: Target-bias correction on B#2 test predictions when no correction or when correcting predictions from B#2 using either the global bias or attributes from the augmented groundtruth, as explained in 4.3. '*' is best result per attribute category.

	Global corr.	Gender	Ethnicity	Makeup lvl.	Time photo	Emotion
female	12.6436	12.1350	12.6750	11.9454*	12.7733	12.5576
male	11.4599*	11.6807	11.4703	11.6602	11.6533	11.4723
caucasian	12.1299	11.9729	12.1932	11.8626*	12.3042	12.1019
asian	10.9157	10.7842	10.7286	10.6955*	10.9552	10.7954
afro	12.8236	12.6636	11.9774*	12.5957	12.8205	12.6415
nomakeup	12.4768	12.5908	12.4891	12.6840	12.6164	12.4663*
verysubtle	10.9865	10.5608*	10.9269	10.6721	11.0208	10.8752
notclear	12.5676	12.3789	12.5528	12.2298*	12.4960	12.5609
makeup	11.4157	10.7861	11.4656	10.2724*	11.6424	11.3321
modernph..	12.5117	12.3182	12.5254	12.2334*	12.4674	12.4442
oldphoto	10.5234	10.5012	10.5707	10.3134*	11.4017	10.5867
happy	12.2262	12.1982	12.2775	12.0841*	12.4886	12.4442
slightly..	12.4890	12.2241	12.4960	12.1136*	12.6133	12.3358
neutral	11.7822	11.5183	11.7787	11.4299*	11.8655	11.4722
others	10.8248	10.8472	10.8367	10.7001*	10.8811	10.9739
ALL	12.0703	11.9149	12.0915	11.8072*	12.2308	12.0319

real age estimation. Despite our initial analyses on target- and guess-bias and their utility to enhance age recognition, still several issues remain open for future work. Further attributes from guessers should be considered for guess-bias analysis. Additional data with more homogeneous age distribution will be also desired to check for statistical significance of the biases. More interestingly, future work may involve prior-to-learning bias correction as well as automatic learning of different attribute bias correction combinations.

Acknowledgements Work supported by Spanish project TIN2016-74946-P (MINECO/FEDER, UE), CERCA Programme (Generalitat de Catalunya), Estonian Research Council Grants (PUT638), Scientific and Technological Research Council of Turkey (TÜBTAK) (Proje 1001 - 116E097), and Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund. We gratefully acknowledge the support of NVIDIA Corporation with the donated GPU used for this research.

References

[1] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 87–94. IEEE, 2017.

[2] F. Alnajar, Z. Lou, J. M. Álvarez, T. Gevers, et al. Expression-invariant age estimation. In *BMVC*, 2014.

[3] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay. Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 96–104, 2016.

[4] A. Bastanfard, M. A. Nik, and M. M. Dehshibi. Iranian face database with age, pose and expression. In *Machine Vision, 2007. ICMV 2007. International Conference on*, pages 50–55. IEEE, 2007.

[5] O. Bilici. *Local Phase Quantization Feature Extraction based Age and Gender Estimation Using Convolutional Neural Network*. PhD thesis, Tartu Ülikool, 2017.

[6] A. Bolotnikova, H. Demirel, and G. Anbarjafari. Real-time ensemble based face recognition system for nao humanoids using local binary pattern. *Analog Integrated Circuits and Signal Processing*, pages 1–8, 2017.

[7] D. M. Burt and D. I. Perrett. Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information. *Proc. R. Soc. Lond. B*, 259(1355):137–143, 1995.

[8] H. Dibeklioğlu, F. Alnajar, A. A. Salah, and T. Gevers. Combining facial dynamics with appearance for age estimation. *IEEE Transactions on Image Processing*, 24(6):1928–1943, 2015.

[9] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

[10] 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*, pages 1–9, 2015.

[11] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

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

[13] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 256–263. IEEE, 2009.

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

[15] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A study on automatic age estimation using a large database. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1986–1991. IEEE, 2009.

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

[17] R. Jana, D. Datta, and R. Saha. Age estimation from face image using wrinkle features. *Procedia Computer Science*, 46:1754–1761, 2015.

[18] Y. KAZUHIKO. Human and object interaction processing (hoip) project. *Joho Shori Gakkai Shinpojiumu Ronbunshu*, pages 379–384, 2002.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[21] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[22] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24, 2015.

[23] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013.

[24] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu, and T. Gevers. Expression-invariant age estimation using structured learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):365–375, 2018.

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

[26] R. C. Malli, M. Aygün, and H. K. Ekenel. Apparent age estimation using ensemble of deep learning models. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2016 IEEE Conference on*, pages 714–721. IEEE, 2016.

[27] M. Minear and D. C. Park. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36(4):630–633, 2004.

[28] L. Nguyen Bao, D.-N. Le, L. Van Chung, and G. N. Nguyen. Performance evaluation of video-based face recognition approaches for online video contextual advertisement user-oriented system. In *Information Systems Design and Intelligent Applications*, pages 287–295. Springer, 2016.

[29] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 85–94. ACM, 2009.

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

[31] E. Patterson, A. Sethuram, M. Albert, K. Ricanek, and M. King. Aspects of age variation in facial morphology affecting biometrics. In *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1–6. IEEE, 2007.

[32] S. Peng, L. Liu, X. Yang, and N. Sang. A database schema for large scale annotated image dataset. In *Image and Signal Processing, 2008. CISP'08. Congress on*, volume 3, pages 57–62. IEEE, 2008.

[33] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005.

[34] 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*, 16(5):295–306, 1998.

[35] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.

[36] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15, 2015.

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

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[39] A. Sandygulova, M. Dragone, and G. M. O'Hare. Real-time adaptive child-robot interaction: Age and gender determination of children based on 3d body metrics. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 826–831. IEEE, 2014.

[40] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[41] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 53–58. IEEE, 2002.

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

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

[44] 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*, 27(2):265, 2012.

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

[46] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling. Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, 66:106–116, 2017.

[47] Y. Zhu, Y. Li, G. Mu, and G. Guo. A study on apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 25–31, 2015.