Forensic Science International: Digital Investigation 32 (2020) 300921

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS 2020 EU – Proceedings of the Seventh Annual DFRWS Europe

DeepUAge: Improving Underage Age Estimation Accuracy to Aid CSEM Investigation

Felix Anda*, Nhien-An Le-Khac, Mark Scanlon

Forensics and Security Research Group, University College Dublin, Ireland

ARTICLE INFO

Article history:

Keywords:
Child Sexual Exploitation Material (CSEM)
Age estimation
Underage facial age dataset
Child sexual abuse investigations
Deep learning

ABSTRACT

Age is a soft biometric trait that can aid law enforcement in the identification of victims of Child Sexual Exploitation Material (CSEM) creation/distribution. Accurate age estimation of subjects can classify explicit content possession as illegal during an investigation. Automation of this age classification has the potential to expedite content discovery and focus the investigation of digital evidence through the prioritisation of evidence containing CSEM. In recent years, artificial intelligence based approaches for automated age estimation have been created, and many public cloud service providers offer this service on their platforms. The accuracy of these algorithms have been improving over recent years. These existing approaches perform satisfactorily for adult subjects, but perform wholly inadequately for underage subjects.

To this end, the largest underage facial age dataset, VisAGe, has been used in this work to train a ResNet50 based deep learning model, DeepUAge, that achieved state-of-the-art beating performance for age estimation of minors. This paper describes the design and implementation of this model. An evaluation, validation and comparison of the proposed model is performed against existing facial age classifiers resulting in the best overall performance for underage subjects.

1. Introduction

The technological advancement of modern society encompasses a myriad of interconnected digital devices. Petabytes of information about our lives are consciously uploaded daily to various social media platforms and communications via mobile messaging applications. Nevertheless, personal data is also collected surreptitiously from location based services, facial detection by private or municipal CCTV cameras and omnipresent internet-connected sensors.

The prevalence of multimedia content has necessitated the storing of increasing volumes of information – resulting in an increased local storage capacity on mobile device and the cloud servers where the data is backed-up. Also, consumers continuously demand higher quality photographs that require an increase on the number of pixels in digital cameras. Moore's law has negatively affected digital forensic investigations due to the amount of devices that are seized in modern crime scenes coupled with the increasing storage per device (Lillis et al., 2016). After the evidence acquisition phase, the devices are stored for further analysis. An exponential accumulation of devices in digital forensic laboratories is an ongoing issue that has contributed to backlogs in Law Enforcement Agencies (LEAs) over the past years throughout the globe (Scanlon, 2016). Intelligent automation is needed to expedite digital investigations that are hampered by lack of resources, such as time and skilled expertise. Moreover, Sanchez et al. (2019) verified that digital forensic practitioners demand automated tools to detect CSEM, age estimation and skin tone detection and intelligent artefact prioritisation can expedite digital investigation (Du and Scanlon, 2019). Our research tackles the age estimation problem for minors. Our age estimation model discards any unnecessary details, i.e., background and other noise. Based on the *No-Free-Lunch* theorem, i.e., there is no single model that works best for every problem, our focus was to find a model that suits best for the underage age estimation. The approach may fail in other situations, such as adult age ranges.

The application area for this work is to execute facial age estimation over images that have already been flagged with nudity and be further used in conjunction with an ensemble of models that endeavour to solve the bigger problem, i.e., the automated categorisation of this content as CSEM. The element of study is depicted

* Corresponding author.
E-mail addresses: felix.andabasabe@ucdconnect.ie (F. Anda), an.lekhac@ucd.ie (N.-A. Le-Khac), mark.scanlon@ucd.ie (M. Scanlon).

https://doi.org/10.1016/j.fsidi.2020.300921

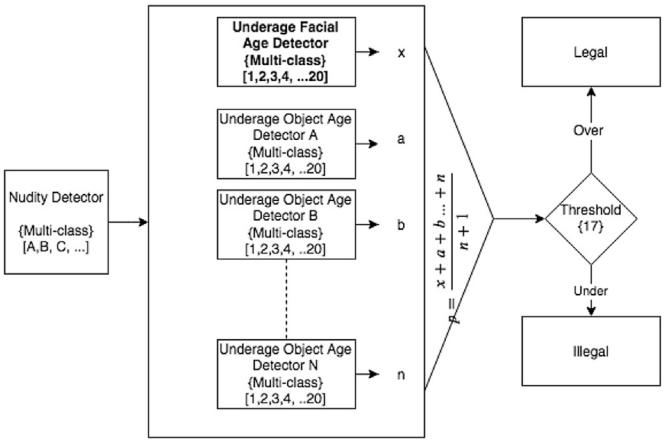

Fig. 1. Pipeline of designed ensembles to detect CSEM.

as the Underage Facial Age Detector in Fig. 1. Although the pipeline is comprised of several components, our goal is to achieve the highest accuracy possible for underage subjects between the age of 1 to 18 years old. In the aforementioned pipeline, an initial image filter will decrease the input of images based on nudity/pornography detection, such as the technique developed by Sae-Bae et al. (2014). A nudity/pornography detector would serve as a primary filter to aid the ensemble and facilitate the encapsulation of our age estimation methodology.

Challenges arise due to the lack of underage facial images for training models, the validation of age labels, and the use of estimated or simply guessed values as ground truths in existing datasets. Moreover, performance within this age group has been previously evaluated with state-of-the-art age estimators and the inadequate results are a motivation for our research. Anda et al. (2019) assessed several online services such as Amazon Rekognition, Microsoft Azure Cognitive Services, How-Old.net, and Deep EXpectation (DEX), an offline Caffe[1] model and winner of the ChaLearn LAP 2015 challenge. To address the existing challenges, this work created a balanced and validated underage dataset (VisAGe), a non-binary age estimation model was trained, and the accuracy of underage age estimation was improved.

The paper is organised as follows. In Section 2, an overview of the related work is presented. Section 3 provides an overview of the design and methodology of the developed model and its derivation from the VisAGe dataset, and the development of the DCA contouring pre-processing technique. Section 4 describes the performance of the DeepUAge model in comparison with other state-of-the-art facial age estimators as well as the evaluation of the pre-processing techniques used. Section 4.1 explains the evaluation for the age estimation framework. Finally, the last section outlines the conclusions and discusses future work.

Contribution of this work:

- Creation of a dataset consisting of an underage facial age dataset with validated, accurate age and gender labels.
- Creation of DeepUAge an age estimation state-of-the-art model with a mean absolute error (MAE) rate of 2.73 years for underage subjects.
- Evaluation of several facial image pre-processing techniques and the impact of facial landmark points for age estimations on subjects from the age range 1 to 18 years old.
- Creation of DCA, an encapsulation of an offline pre-processing method suitable for law enforcement agencies (LEA) used for the automatic age prediction of unlawful images. Available at https://github.com/4ND4/DeepUAge.

2. Related work

Many facial age estimation methods are built on datasets with estimated or guessed age label, low sample counts, highly noisy data, unbalanced/biased data, or age bins (e.g., binary groups that separate underage subjects from adults, and/or small groups such as child, teen, adult, etc.). Usually, the usage of groups hinder the filtering of illegal content due to its non-dynamic structure. Some of the aforementioned methods affect the reliability of the machine learning assistance and hence decrease the credibility of the approach. Furthermore, any introduction of reasonable doubt may dismiss a case. To attempt to alleviate this issue, the Daubert standard is suggested. It was introduced in 1993 and has been used by most state courts in the USA as a rule of evidence to assess the reliability of scientific evidence through the following factors: (1) the method can be and has been tested, (2) subject to peer review, (3) error rates are acceptable, (4) general acceptance in the scientific community (Nawara, 2010). In regards to these factors, Nutter (2018) states that “machine learning easily satisfies three of the four Daubert factors without extensive discussion”.

2.1. Facial age estimation

Automated facial age estimation requires three main properties: validated facial age labelled datasets, robust face detection, and machine learning. Many existing approached reuse the same dataset(s) either for bench-marking, validation, training or testing purposes. In the early stages, facial anthropometric[2] models were suggested for age prediction. In 2006, Ramanathan and Chellappa (2006) proposed a cranio-facial growth model that classifies growth related shape variations observed in underage faces. Their model was capable of face recognition across age progression. The dataset for the aforementioned model was the FG-Net ageing dataset (82 subjects with ages ranging from 0 to 69 years old and over 50% juvenile subjects[3]) and a separate dataset containing 233 images (Ramanathan and Chellappa, 2006).

In 2009, Guo et al. (2009) found that age estimation performance was able to improve when manifold learning uses biologically inspired features (BIF). The accuracy achieved is positively influenced by a known gender; therefore, their approach consisted in two different MAEs for each gender. Furthermore, the model was a combination of “BIF locality sensitive discriminant analysis” and “BIF marginal fisher analysis”, reaching MAE rates of 2.58 years for males and 2.61 years for females. The data used was the large Yamaha gender and age (YGA) database that contains 8,000 outdoor facial images of Asian ethnicity subjects. The dataset was distributed equally by gender and divided in age ranges from 0 to 93 with intervals of 9 classes per age group until the age of 70 and a group containing the rest due to the lack of images available for subjects over 70.

Eidinger et al. (2014) presented an approach on age and gender estimation using standard linear support vector machines (SVM) with their own dropout-SVM scheme. The dataset employed was Adience, which is a collection of Creative Commons (CC) images

[1] https://caffe.berkeleyvision.org/.

[2] The science of measuring sizes and proportions on human faces (Ramanathan and Chellappa, 2006).

[3] https://yanweifu.github.io/FG_NET_data/index.html.

sourced from Flickr. Furthermore, the ages were labelled in heterogeneous categories in an unknown manner that contained in average 2,205 images per age group, but were unbalanced.

More recently in 2018, Rothe et al. (2018) proposed a deep learning solution based on a VGG16 convolutional neural network architecture pre-trained on ImageNet.[4] This achieved a MAE of 3.252 years. The data was trained on a dataset named IMDB-WIKI, consisting in over half a million facial images of celebrities (currently the largest public face image dataset annotated with age and gender labels) that had been crawled from IMDB and cross referenced with the age denoted in Wikipedia.

2.2. Underage facial age estimation

Work specialised in underage facial age estimation has been limited due to the challenges of collecting data which is understandably subject to ethical implications, lack of underage datasets, and scarcity of reliably annotated images. Nevertheless, apparent age estimation on children was studied by Antipov et al. (2016) in 2016. Antipov et al. used a fine-tuned VGG-16 (a very deep convolutional network of 16 weight layers used for large-scale image classification) to train a model of minors from the age range 0 to 12 years old. HeadHunter, a detector based on rigid templates, was their choice for face detection and the alignment technique was based on a multi-view facial landmark detection tool. The error rate achieved for validation was 0.2609, a metric ε defined as the size of the tail of the normal distribution with the mean μ and the standard deviation σ with respect to the predicted value $\hat{x}$.

In the same year, Ferguson and Wilkinson (2017) determined that manual human visual age estimation of childrens' faces reveals poor accuracy, confirming the difficultly to precisely predict age. They also suggested that black and white images were classified with less accuracy. The latter suggestion was confirmed with several experiments that are mentioned in Section 4.1. In 2019, Anda et al. improved facial age estimation with an ensemble technique approach that was fine-tuned on DEX for the age group 16 to 17 years old, which falls in borderline adulthood age in many jurisdictions. Their work also evaluated the state-of-the-art cloud based facial age estimators, such as Amazon AWS Rekognition, and Microsoft Azure Cognitive Services.

2.3. Facial pre-processing for age estimation

Facial image pre-processing may not be necessary if the source is akin to a standard passport photograph. However, facial images in-the-wild may have characteristics such as various pitch/roll/yaw angles, multiple subjects per image, background noise, varying image size and quality, etc. Such photos require image pre-processing and normalisation to align and remove unnecessary features. Reisfeld and Yeshurun (1998) suggest that the knowledge of the location and the scale of a face impacts positively on the speed and reliability of face recognition systems.

In 2015, Han et al. (2015) designed a face pre-processing procedure to overcome image variations due to external factors. Their approach entails: (1) converting a colour facial image into greyscale, (2) rectifying the face based on the two eyes and cropping to 60x60 pixels with a 32-pixel interpupillary distance (IPD), (3) detecting the face and the eyes using Cognitec's commercial FaceVACS SDK, and (4) applying the Difference of Gaussians filtering. In the same year, Liu et al. (2015) proposed a deeply learned regressor and classifier for robust apparent age estimation. A three step preprocessing procedure was implemented: face detection, facial landmark localisation, and facial normalisation. For the first, a face detection toolkit developed by VIPL lab of CAS was used; for the second, 5 facial landmarks were detected with a Coarse-to-Fine Auto-Encoder Network. For the last step, external and internal normalisation approaches were considered.

2.4. Facial age dataset

Large datasets for underage subjects with accurate labels are rare. Accurately labelled age and gender datasets are preferred over apparent age estimation and guessed metadata. In 2013, Dalrymple et al. created a set of images with the following variations: 8 facial expressions, 5 angles and 2 lightning conditions. The collection consisted of combinations of these variations for 40 male and 40 female Caucasian children between 6 and 16 years-old. The real age was documented and also estimated by external raters with a 79.7% accuracy. Later in 2014, a 50 image dataset of female subjects aged 10 to 19 years from Germany, Italy, and Lithuania was created. In 2015, the In-The-Wild Child Celebrity (ITWCC) dataset was created by Ricanek et al. and the set was composed of 304 subjects with a total of 1,715 images (876 female and 839 male) from the age range 5 months to 32 years. Next in 2016, the Boys2Men collection was released as a private dataset mainly focused on male images from the age range 12 to 21 years-old (Castrillón-Santana et al., 2016). In 2018, Deb et al. released a dataset containing 3,682 face images of 919 subjects, in the age group 2 to 18. It is notable that in the past 5 years, the number of underage datasets has grown but still requires validation, accurate age labels, and balance.

3. Design/methodology

This work presents an underage facial age estimation deep convolutional neural network (DCNN) based on a residual neural network of 50 layers (ResNet50). Our proposed method is pre-trained on the ImageNet Dataset.[5] The last Fully-Connected (FC) Softmax layer with 1,000 outputs has been replaced by a FC Softmax activation function layer of 20 outputs that correspond to the age classes studied (1 to 20 years old) to suit our needs. Subsequently, the parameters of the convolutional layers during the training process have been frozen. The ResNet50 architecture employed for facial age estimation and the replacement in the last layer with 20 outputs can be seen in Fig. 2. The age estimation problem was treated as a classification task and therefore, a categorical cross-entropy logarithmic loss function is used.

The batch size is a hyper-parameter that defines the number of training samples used in one iteration and is directly proportional to the memory space required. A batch size of 64 was chosen due to RAM limitations with the development server. One forward pass and one backward pass of all the training examples are referred to as epochs. A reference of 100 epochs was chosen; however, the training process was monitored and an early stopping implementation to prevent over-fitting was accomplished. The metric used for accuracy was the MAE, i.e., the average of the absolute mean error between the ground truth and the predicted values. The formula used to evaluate the model is described in Equation (1).

$$MAE = \sum_{i=1}^{n} \frac{|predicted_i - real_i|}{n} \tag{1}$$

The optimiser chosen was stochastic gradient descent (SGD) that includes support for momentum, learning rate decay, Nesterov

[4] http://www.image-net.org/.

[5] ImageNet is a large scale ontology of images organised according to the WordNet hierarchy (Deng et al., 2009).

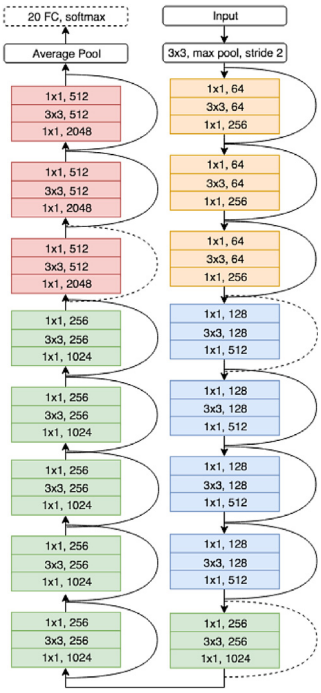

Fig. 2. ResNet50 Pre-trained on ImageNet with 20 Outputs in the FC Softmax layer. Grouping of Convolution Layers are Denoted by Colour.

momentum. SGD demonstrates excellent performance for large-scale problems (Bottou et al., 2010). The learning rate is a hyper-parameter that controls the number of changes affected by the model in response to the estimated error each time the model weights are updated. The selected value for the initialiser was 0.1 and the momentum was 0.9. The latter is a parameter that accelerates SGD in the relevant direction and reduces oscillations. The input image size for width and height chosen was 224 with the depth set to 3; values chosen due to the hardware limitations and the decreased performance experienced with images of smaller dimensions. Data augmentation techniques such as flip, rotation, zoom, distortion, colour, contrast, brightness and random erasing[6] were used. Moreover, augmentation in data-space provides a greater benefit for improving performance and reducing overfitting (Wong et al., 2016).

3.1. Proposed facial age and gender dataset

Age and gender estimation models require a large number of images with real age and gender labels; moreover, the data must be balanced within each class and this represents a challenge. Nevertheless, it was possible to build a balanced set divided with images from two sources; the VisAGe dataset supplemented by the Anda et al. (2018) dataset generator.

The creation of VisAGe involved the annotating of the largest set of underage images to date. These images are creative commons licensed with initially indicative age gathered from Flickr. Each of these photos were processed by face and gender detection algorithms, and other associated metadata were compiled such as dimensions, title, tags, comments, dates, etc. Given the level of error rates in automated facial age estimation and gender identification, each of these images were subjected to human age and gender verification. Each photo was voted on by three human assessors and if the decisions on age and gender were unanimous, the photo was added to the dataset. The set used in this work consists of 19,446 images from the age range 1 to 18. Further detail regarding age and gender per class is depicted in Table 1. It is notable that the age ranges contain an unbalanced amount of images within each age and/or gender group. The average number of male images per age is 521 versus 557 females. This unbalance will be addressed in future work.

In order to present a sample of the dataset, an average image per class has been calculated. Average faces from age 1 to 18 can be seen in Fig. 3. The face cropping technique used in this paper is also applied and visible on each face.

The age estimation model was trained with the majority of images from the VisAGe dataset and was prepared in a balanced fashion. Due to the size of the dataset available, 800 photos were selected from each class. When enough images were available for each class, a balanced amount of images were selected for both male and females. In the age classed of 8 and higher, there were insufficient images to fulfil the 800 training/testing images. As a result, the remainder of the 800 images used were filled with the Anda et al. (2018) facial age dataset generator with randomly obtained images from different datasets including FGNET (Lanitis and Cootes, 2002), IMDB-WIKI (Rothe et al., 2018), FERET (Phillips et al., 1998), MEDS (Founds et al., 2011).

Although the model application is for underage images, it was necessary to consider 2 additional years over the maximum year limit (18 years old). Since the best performance of existing approaches are approximately 2−3 MAE in years, the chosen limit was 20. Furthermore, the additional age classes were also completed with the aforementioned dataset generator.

3.2. Facial image preprocessing

An important step for an image classification task is to filter unnecessary features that would affect the learning process of a machine learning algorithm. Initially, the approach implemented used the `Face++`[7] API to obtain the face landmarks rather than the open source machine learning library `dlib` (King, 2009). Once the landmarks were collected, the left and right eye centre values were processed to further compute the angle between the eye centroids. Next, the median point between the two eyes in the input image was computed and subsequently rotated. Finally an affine transformation was applied to the image with warping using the specified matrix, as per Equation (2):

$$dst(X, Y) = src(M_{11}x + M_{12}y + M_{13}, M_{21}x + M_{22}y + M_{23}) \quad (2)$$

Another approach would be to solve the *procrustes problem* (Gower, 1975) by subtracting centroids, scaling by the standard deviation, and then using the singular value decomposition to calculate the rotation. Once the face was aligned, the new facial landmark positions were detected and a mask for the 273 contour points provided by Face++ was created. The mask is overlaid and the face is cropped. An example of a `dlib` cropped face vs a Face++ cropped face can be seen in Fig. 4. The `dlib` landmark tool extracts 27 relevant contour points from the 68 points in total against 273/1000 from Face++.

The major drawback of using the Face++ API is that the images must be sent to a remote cloud service and the accessibility is

[6] "Random erasing randomly selects a rectangle region in an image and erases its pixels with random values. In this process, training images with various levels of occlusion are generated, which reduces the risk of over-fitting and makes the model robust to occlusion" (Zhong et al., 2017).

[7] https://www.faceplusplus.com/

Table 1
VisAGe dataset - facial images per class per gender from 1 to 18 Years old.

Age	Combined	Male	Female
1	4,236	2,292	1,944
2	2,722	1,485	1,237
3	2,280	1,071	1,209
4	2,434	1,110	1,324
5	1,227	515	712
6	984	462	522
7	974	418	556
8	686	315	371
9	453	256	197
10	401	217	184
11	371	154	217
12	211	103	108
13	354	171	183
14	217	142	75
15	337	91	246
16	589	184	405
17	285	204	81
18	660	193	467
Total	19,446		

limited to this cloud environment. In dealing with CSEM investigations, LEAs cannot transmit this sensitive information to a third party service. To overcome this issue, a customised facial cropping technique was implemented using `dlib` as a base and extending it to predict the hairline with a facial proportion artistic approach by Loomis (2017).

3.2.1. DCA facial proportion artistic approach

Initially, the Face++ and `dlib` pre-processing techniques were analysed to decide which approach was most ideal to utilise for the validation of the model. Exploration of the Face++ 1,000 landmark points detection tool has 273 contour points (refer to the image on the middle of Fig. 4). Unfortunately, due to the problem with remote cloud services stated in Section 3.2 likely being insurmountable for CSEM investigation, the `dlib` tool was selected. The novel pre-processing technique, DCA, was implemented instead based on `dlib`. The `dlib library` returns 68 landmark points from which 27 correspond to the face contour. In the left of Fig. 4, the `dlib` jawline contour highlighted in green from point number 1 to 27 can be seen. Portions of the head, such as the forehead or wrinkles, are important features that impact the age estimation of a subject; these features are not supported with the 68 landmark detector. The DCA approach was addresses this limitation, as can be seen on the right of Fig. 4). It uses facial proportionality to reconstruct the face and obtain landmarks that are close to the hairline.

Fig. 5 depicts the proportionality between the nose, eyes and hairline contour. To predict the hairline landmarks, the following steps that emulate the face drawing methodology was carried out as follows:

1. Use the `dlib` landmark detector to obtain the coordinates x, y of the lowest point of the nose which corresponds to the point 34.
2. Compute the average distance between the point 34 and the intersecting points that lie close to a perpendicular drawn from the nose point towards the contours. The square side is twice this value.

$$d_1 = \sqrt{\left(p_{left_x} - p_{34_x}\right)^2 + \left(p_{left_y} - p_{34_y}\right)^2}$$

$$d_2 = \sqrt{\left(p_{right_x} - p_{34_x}\right)^2 + \left(p_{right_y} - p_{34_y}\right)^2}$$

$$square_{side} = \frac{d_1 + d_2}{2} * 2$$

$$square_{side} = d_1 + d_2$$

3. Locate both vertexes v_1, v_2 of the square and use those values to draw the shape within the circle as shown in Fig. 5.
4. From the centre points c_x, c_y, draw a regular polygon of size $N = 20$ (Icosagon). Notice that the circle drawn in Fig. 5 corresponds to the half icosagon drawn in Fig. 6.

Our method uses the foundations of the Loomis face proportion approach. In Fig. 6, the output of a digital sketch of our approach to detect hairline landmark points is depicted. Last, a cropped mask can be used to filter noise from facial images.

4. Results

As mentioned in Section 3.1, a balanced dataset of 16,000 images was prepared. 80% of the images were used for training and 20% were used for validation. The training stopped on 87 epochs, maintained a loss under 1.799 with a favourable MAE of 1.57 years and 2.73 years for validation and testing respectively. The model was further tested with 1,000 additional images that were gathered from the UTKFace dataset (Zhang et al., 2017) and the aforementioned dataset generator. These images amounted to 50 images per class. The testing achieved a MAE of 2.73 years.

In the process of creating the DeepUAge model, the DCA facial proportion artistic approach was developed. Its efficiency in pre-processing has been evaluated in comparison to other pre-processing techniques. It produced a MAE of 2.73 years, which was the best performing of all the approaches evaluated. The results for this experiment can be found in Table 2. The accuracy of the DeepUAge model in both validation and testing in comparison with other facial age estimators for underage subjects is outlined in Section 4.1.3.

4.1. Evaluation

4.1.1. DCA and other pre-processing techniques

By separating the age estimation problem into a smaller scope fo age range, it was possible to validate the input data of juveniles used to train the model. As the data used was not only frontal "passport photo style" facial images, pre-processing procedures have been adopted to minimise any negative impact of of noise on the final results.

Several types of existing pre-processing techniques were evaluated; `dlib` contour aligned, `dlib` contour non-aligned, `dlib` cropped, and Face++ contour. Variations of the `dlib` approaches all produced a MAE greater than 4 years. In particular, the `dlib` cropped technique was found to perform almost equally as not using a pre-processing filter at all. Additionally, even the best performing `dlib` contour aligned, was still 43.72% greater in MAE versus the Face++ contour. Unfortunately, due to the unsuitability of Face++ for CSEM investigation, the technique was deemed unusable for the DeepUAge model. This motivated the development of the DCA facial cropping techniques, which nonetheless achieved better results than all other pre-processing techniques evaluated.

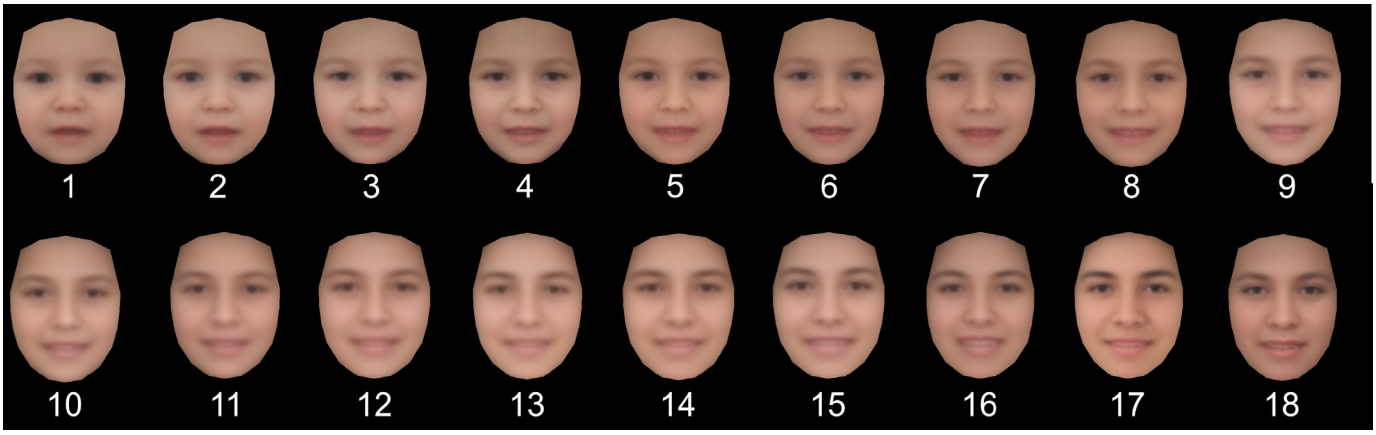

Fig. 3. VisAGe dataset - average face per age from 1 Year old (top-left) to 18 Years old (bottom-right).

Fig. 4. Image Taken from the FG-NET Aging Database (Wallhoff, 2006) with 64 `dlib` Landmarks (left), 1000 Face++ landmarks (middle), and the DCA approach (right).

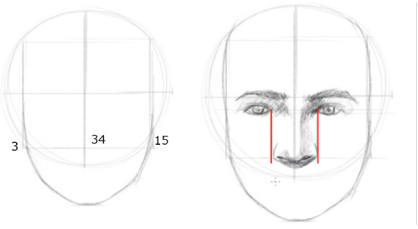

Fig. 5. `dlib` landmark points (3, 34, 15) superimposed over a Loomis face proportion approach sketch (Fussell, 2019) (image reproduced with permission).

4.1.2. DeepUAge performance

Fig. 7 illustrates a box plot of real age vs. DeepUAge predicted age where accuracy and precision for the age range 1–20 can be observed. Ages 18 and 20 have the largest range of predicted age. Whilst the model manages to obtain correct classification for both ages, the performance score for these classification age ranges is low. Instead, the correct age prediction is found at the maximum whiskers of the data for the age of 20 and at the 3rd quartile of the data distribution for the age of 18.

Similar results can be seen across the age range 14 to 20 where the correct age lies between the maximum and 3rd quartile of the classification distribution. Conversely, the reverse can be observed on the opposite end of the age spectrum where younger ages are often overestimated. It can be observed that for ages 1, 5, 8 and 10 the minimum of the classification distribution contained the correct catalogue of the images. Furthermore, for ages 2, 3, 9 and 11 the correct predicted age is between the minimum and the 1st quartile of the classification distribution. There are several outliers across all ages, predominately in ages 7, 13, 16, 17 and 19, where 4 or more outliers can be observed.

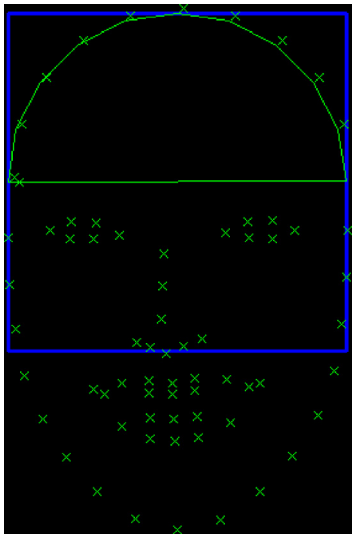

Fig. 6. Contour reconstruction from Dlib.

Table 2
Results of different preprocessing techniques.

Approach	MAE
DCA	**2.73**
Face++ contour	2.79
Dlib contour aligned	4.01
Dlib contour non-aligned	4.28
Dlib cropped	5.31
Non-processed	5.71

Table 3
Evaluation of facial age estimators for underage subjects.

Approach	MAE
DeepUAge Test	**2.73**
Microsoft Azure Face API	3.60
Amazon Rekognition	3.74
Dummy estimator (All assumed 10 y/o)	5.00
Face++	18.21
IMDB-WIKI WideResNet (Rothe et al., 2018)	20.43

4.1.3. Comparison of DeepUAge and other estimation techniques

State-of-the-art age estimators were compared against DeepUAge, with the same testing set. The test performance achieved is inline with the state-of-the art age estimation classifiers. The testing MAE of 2.73 for DeepUAge indicates that the average magnitude of the model error was significantly lower than that of its alternatives. The top 3 performers found in descending order were DeepUAge, Microsoft Azure Face API, followed by Amazon Rekognition, as can be seen in Table 3. Both Microsoft's and Amazon's approaches were side by side in performance with only being 0.14 MAE apart. DeepUAge stood at 0.87 MAE better than Microsoft Azure Face API and exceeding Amazon Rekognition by a MAE of 1.01.

Despite obtaining the best results of the age estimators evaluated, DeepUAge has a logarithmic loss (the variation of the actual label from the machine learning model predicted value) of 1.799. Our goal is to decrease this value further to as close to 0 as possible and is addressed as our future work in Section 6. Although IMDB-WIKI WideResNet was trained on a large celebrity dataset (over half a million images), it had the lowest performance. Along with Face++, both estimators had the poorest performance overall. The dummy estimator (an approach which classifies images to a fixed predicted age of 10) managed to surpass the performance of two intricate algorithms. This result can be due to the failure to validate age labels and the lack of images in the underage age group.

As shown in Table 3, the overall performance of DeepUAge was found to be best compared with the other models evaluated. This is demonstrated at a better granularity in Fig. 8 where DeepUAge is shown to catalogue age 9 to 18 with a higher accuracy than those of the other age predicting services. In reference to Fig. 7, it can be further concluded that the precision of DeepUAge is at its peak for ages 7, 8, 10, 13, 16 and 17. In particular, it is most accurate at age classification of 13 and 17 year olds; therefore providing better age classification techniques for "early adolescent (age 10 to 14) and late adolescent (age 15 to 19) years" (Signorielli, 1987). This age range is important in CSEM investigations because it covers subjects within the borderline of adulthood.

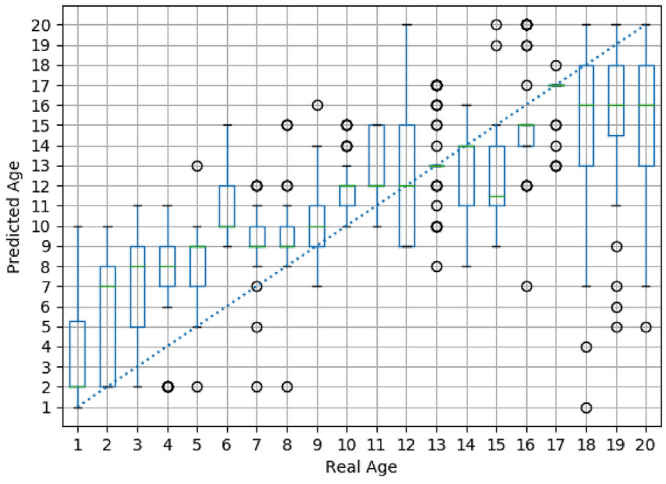

Fig. 7. DeepUAge - Real Age vs Predicted Age.

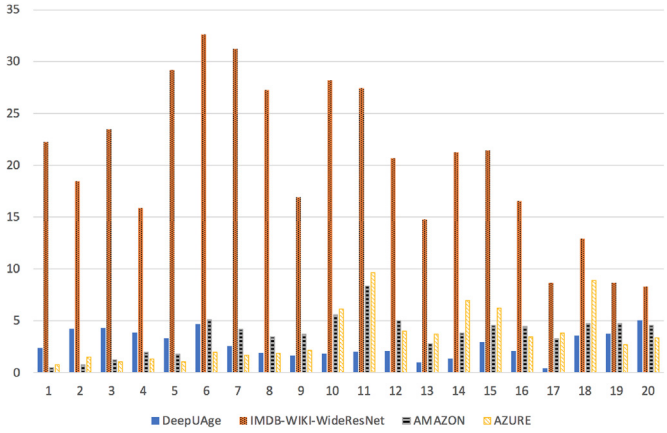

Fig. 8. Average difference per age per facial age predictor.

5. Discussion

According to the No-Free-Lunch theorem, the age scope for our age estimation problem was established. Although binary classification models are not used, numeric ages are used and the problem was treated as a classification problem as opposed to a regression problem. Several factors including environment, habits, makeup, and ethnicity can have an impact on our perceived age and therefore make this problem more complex. Until now, the lack of underage datasets and scarce validation has had an influence in the performance of age estimation models. This work has been able to improve the performance of age estimation for underage subjects.

One weakness of our model is the data itself – demographic balance in regards to ethnicity is lacking. Part one of the dataset used relies on input labels from Flickr users. Naturally, it is not often that a person classifies/tags images with these factors online or in social media. To improve facial age estimation, the problem should be further divided by gender and ethnicity and once a model is generated for each type, create an ensemble that could work in a hierarchical manner. The algorithm should be able to chose the model based on the age range, gender, ethnicity, and other relevant factors.

6. Conclusion and future work

It is true that the validation of data, the reduction of unnecessary features, and the implementation of structured design reduced the problem to a smaller one. Underage facial age estimation is a challenging problem that can be assisted by the use of pre-trained models. Such models hinge on accurate age labels. The performance of these models can be improved further by; 1) obtaining a

large dataset with balanced, validated class labelling, and 2) ensuring optimal pre-processing of the data, e.g., subject alignment and the elimination of unnecessary noise, such as photo backgrounds or hair.

Given that expert human age estimation of childrens' faces often achieves poor accuracy (Ferguson and Wilkinson, 2017), the model proposed with a predictable error rate makes DeepUAge potentially a valuable aid for law enforcement. Finally, it can be concluded that ensemble pre-trained deep learning methods can have a positive impact on CSEM investigation, e.g., as a triage technique or evidence analysis prioritisation tool.

6.1. Future work

As future work, the accuracy of our current age estimation algorithm can be improved, i.e., decreasing both the MAE and logarithmic loss. The main goal of this work is to aid law enforcement in the detection and investigation of CSEM. From a victim identification standpoint, we would like to analyse other components that are present in a digital forensic CSEM crime scene including garments, visual geolocation clues, object detection, etc. We also plan to make the VisAGe dataset available and encourage others to contribute, e.g., to improve the demographic balance that is currently lacking. Lastly, a safe framework will be created to interact with LEA and evaluate the accuracy of our approach with real cases.

Acknowledgements

The authors wish to acknowledge the contribution of colleagues in the UCD Forensics and Security Research Group and in the MITRE Corporation in the curation of the VisAGe dataset. This work has been supported in part by the Google Cloud Platform Research Credits Program.

References

Anda, F., Lillis, D., Le-Khac, N.A., Scanlon, M., 2018. Evaluating automated facial age estimation techniques for digital forensics. In: 2018 IEEE Security and Privacy Workshops (SPW). IEEE, pp. 129–139.

Anda, F., Lillis, D., Kanta, A., Becker, B.A., Bou-Harb, E., Le-Khac, N.A., Scanlon, M., 2019. Improving borderline adulthood facial age estimation through ensemble learning. In: Proceedings of the 14th International Conference on Availability, Reliability and Security. ARES '19. ACM, New York, NY, USA, ISBN 978-1-4503-7164-3. https://doi.org/10.1145/3339252.3341491, 57:1–57:8. https://search.crossref.org/?q=Improving+borderline+adulthood+facial+age+estimation+through+ensemble+learning.

Antipov, G., Baccouche, M., Berrani, S., Dugelay, J., 2016. Apparent age estimation from face images combining general and children-specialized deep learning models. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 801–809. https://doi.org/10.1109/CVPRW.2016.105.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (Eds.), Proceedings of COMPSTAT'2010. Physica-Verlag HD, Heidelberg, ISBN 978-3-7908-2604-3, pp. 177–186.

Castrillón-Santana, M., Lorenzo Navarro, J.J., Obregón, Freire, Boys2Men, C., 2016. An age estimation dataset with applications to detect infants in pornography content. In: First International Workshop on Biometrics and Image Forensics. IEEE.

Dalrymple, K.A., Gomez, J., Duchaine, B., 2013. The dartmouth database of children's faces: acquisition and validation of a new face stimulus set. PloS One 8 (11), e79131.

Deb, D., Nain, N., Jain, A.K., 2018. Longitudinal study of child face recognition. In: 2018 International Conference on Biometrics (ICB). IEEE, pp. 225–232.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, Imagenet, L., 2009. A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255.

Du, X., Scanlon, M., 2019. Methodology for the automated metadata-based classification of incriminating digital forensic artefacts. In: Proceedings of the 14th International Conference on Availability, Reliability and Security. ARES '19. ACM, New York, NY, USA, ISBN 978-1-4503-7164-3. https://doi.org/10.1145/3339252.3340517, 43:1–43:8. https://search.crossref.org/?q=Methodology+for+the+automated+metadata-based+classification+of+incriminating+digital+forensic+artefacts.

Eidinger, E., Enbar, R., Hassner, T., 2014. Age and gender estimation of unfiltered faces. IEEE Trans. Inf. Forensics Secur. 9 (12), 2170–2179. https://doi.org/10.1109/TIFS.2014.2359646.

Ferguson, E., Wilkinson, C., 2017. Juvenile age estimation from facial images. Sci. Justice 57 (1), 58–62. https://doi.org/10.1016/j.scijus.2016.08.005. http://www.sciencedirect.com/science/article/pii/S1355030616300739.

Founds, A.P., Orlans, N., Genevieve, W., Watson, C.I., 2011. .. In: NIST Special Database 32-Multiple Encounter Dataset II (MEDS-II). Tech. Rep. NIST.

Fussell, M., 2019. Facial Proportions - How to Draw a Face. https://thevirtualinstructor.com/facialproportions.html. (Accessed 4 October 2019).

Gower, J.C., 1975. Generalized procrustes analysis. Psychometrika 40 (1), 33–51.

Guo, G., Mu, G., Fu, Y., Dyer, C., Huang, T., 2009. A study on automatic age estimation using a large database. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1986–1991. https://doi.org/10.1109/ICCV.2009.5459438.

Han, H., Otto, C., Liu, X., Jain, A.K., 2015. Demographic estimation from face images: human vs. Machine performance. IEEE Trans. Pattern Anal. Mach. Intell. 37 (6), 1148–1161. https://doi.org/10.1109/TPAMI.2014.2362759.

King, D.E., 2009. Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. 10, 1755–1758.

Lanitis, A., Cootes, T., 2002. FG-NET aging data base. Cyprus College 2 (3), 5.

Lillis, D., Becker, B., O'Sullivan, T., Scanlon, M., 2016. Current challenges and future research areas for digital forensic investigation. In:). The 11th ADFSL Conference on Digital Forensics, Security and Law (CDFSL 2016, vols. 9–20. ADFSL, Daytona Beach, FL, USA.

Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., Chen, X., 2015. AgeNet: deeply learned regressor and classifier for robust apparent age estimation. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 258–266. https://doi.org/10.1109/ICCVW.2015.42.

Loomis, A., 2017. Figure Drawing for All It's Worth. Editoria Bibliomundi Serviçis Digitais LTDA.

Nawara, J., 2010. Machine learning: face recognition technology evidence in criminal trials. U Louisville L Rev 49, 601.

Nutter, P.W., 2018. Machine learning evidence: admissibility and weight. U Pa J Const L 21, 919.

Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J., 1998. The FERET database and evaluation procedure for face-recognition algorithms. Image Vis Comput. 16 (5), 295–306.

Ramanathan, N., Chellappa, R., 2006. Modeling age progression in young faces. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol 1, pp. 387–394. https://doi.org/10.1109/CVPR.2006.187.

Reisfeld, D., Yeshurun, Y., 1998. Preprocessing of face images: detection of features and pose normalization. Computer Cision and Image Understanding 71 (3), 413–430.

Ricanek, K., Bhardwaj, S., Sodomsky, M., 2015. A review of face recognition against longitudinal child faces. In: Brömme, A., Busch, C., Rathgeb, C., Uhl, A. (Eds.), Proceedings of the 14th International Conference of the Biometrics Special Interest Group (BIOSIG). Gesellschaft für Informatik e.V., Bonn, pp. 15–26.

Rothe, R., Timofte, R., Van Gool, L., 2018. Deep expectation of real and apparent age from a single image without facial landmarks. Int. J. Comput. Vis. 126 (2), 144–157. https://doi.org/10.1007/s11263-016-0940-3.

Sae-Bae, N., Sun, X., Sencar, H.T., Memon, N.D., 2014. Towards automatic detection of child pornography. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 5332–5336. https://doi.org/10.1109/ICIP.2014.7026079.

Sanchez, L., Grajeda, C., Baggili, I., Hall, C., 2019. A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). Digit. Invest. 29, S124–S142. https://doi.org/10.1016/j.diin.2019.04.005. http://www.sciencedirect.com/science/article/pii/S1742287619301549.

Scanlon, M., 2016. Battling the digital forensic backlog. In: Proceedings of the 2nd International Workshop on Cloud Security and Forensics (WCSF 2016), vols 10–14. IEEE, Dublin, Ireland.

Signorielli, N., 1987. Children and adolescents on television: a consistent pattern of devaluation. J. Early Adolesc. 7 (3), 255–268.

Wallhoff, F., 2006. Facial Expressions and Emotions Database. http://www-prima.inrialpes.fr/FGnet/html/home.html.

Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D., 2016. Understanding data augmentation for classification: when to warp?. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1–6. https://doi.org/10.1109/DICTA.2016.7797091.

Zhang, Z., Song, Y., Qi, H., 2017. Age progression/regression by conditional adversarial autoencoder. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4352–4360. https://doi.org/10.1109/CVPR.2017.463.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.. Random erasing data augmentation. CoRR 2017;abs/1708.04896. http://arxiv.org/abs/1708.04896.