16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Region-wise Modeling of Facial Skin Age using Deep CNNs

Matthew Shreve, Raja Bala
Palo Alto Research Center
Palo Alto, CA, USA
mshreve@parc.com

Wencheng Wu, Beilei Xu
University of Rochester
Rochester, NY, USA

Ankur Purwar, Paul Matts
Proctor & Gamble
purwar.a@pg.com

Abstract

We propose a deep learning approach for predicting the apparent age of a person's skin. Our method works by first normalizing a frontal image of a face and cropping rectangular-shaped skin patches that are each normalized and fed into separately trained region-specific CNNs. Each regional CNN model is fine tuned using a novel data augmentation technique that artificially reduces the apparent age of the skin through a series of smoothing operations that act as a proxy for subjects with younger looking skin. The deep features extracted from each of these regions are then used to train a separate set of regression models that predict the skin age. We evaluate our method using two strategies: the first looks at how well the predicted regional skin age clusters around the true biological age of the subject, for which we achieve a 1-off accuracy of approximately 83%. In the second strategy, we validate that our models predict apparent skin age based on a user study that asked over 15 judges to compare image pairs of subjects with the same chronological age, but with different skin age predictions. For this second study, we achieve an average 66% accuracy based on consensus rating across all human raters, and as high as 76% for some age groups.

1 Introduction

The automatic assessment and understanding of facial skin health has many research applications including the early detection of underlying health problems [9], suggested lifestyle and dietary changes such as less sun exposure or more hydration [1], as well as identification of recommended skin-care products that can improve the overall health of facial skin [14]. One of the strongest indicators of skin health that has been identified in the literature is the difference between the *apparent age* (or perceived age) of an individual's skin and their actual chronological age [16]. Research has shown that different parts of the face age differently and that for some individuals, the apparent age can be significantly different from their chronological age [6]. Although one cannot stop the natural aging process, the ability to quantitatively and objectively predict apparent skin age remains a useful proxy for skin health that could provide many benefits to many research areas and applications.

In this paper, we propose deep visual models for predicting apparent skin age (referred to herein as skin age for brevity). Specifically, our models predict the skin age of various facial skin regions based purely on micro-features such as wrinkle, spots, sagging, etc, and with minimal influence from macro-feature such as the shape of eyes, distance between eyes, nose, etc. Potential applications for our work include the assessment of skin quality and health, and recommendations for skin-care products.

Due to the lack of standards and ground truth for skin health or apparent skin age, we assume that it can be approximated by chronological age over a large population. This assumption can be rationalized by the observation that over a large ensemble of human faces, skin age and chronological age are strongly correlated *on average*, while for any individual subject the two age measures may deviate. Models trained on large datasets with skin micro-features as inputs and chronological age as ground truth labels would thus be expected to predict skin age (even if deviated from chronological age) for any given individual. Furthermore, our goal is to locally predict skin age in different regions of the face (e.g. cheeks, chin, under eye, etc.). To this end, we introduce a data augmentation scheme that utilizes the fact that smoother skin is typically perceived to be younger [2]. In all our experiments, we select only faces that are free of makeup.

The following are the key contributions of this paper: (1) To the best of our knowledge, this is the first paper to address the estimation of region-wise apparent skin age using CNNs explicitly designed to focus on micro-features of the skin; (2) We show that it is feasible to predict apparent skin age without macro-cues such as the eyes or nose, which have been commonly used in prior art; (3) We propose a novel two-stage training process for the age regressors, using data augmentation to ensure that our method captures region-specific characteristics of facial skin age; (4) We utilize an early fusion approach to aggregate features from regional skin age estimations to estimate the overall facial skin age.

A secondary goal of this work is to extend our skin age prediction approach to smartphone selfie images captured in the wild. While our initial models are trained on standardized high-quality photographs captured under near-ideal acquisition conditions (hair pulled back, even lighting, neutral expression, no

makeup, etc.), we extend our training and evaluation to roughly 1.5K images collected using a variety of frontal cell phone cameras, thus testing the generalization of our models to an end-user application in realistic settings where adverse conditions are present.

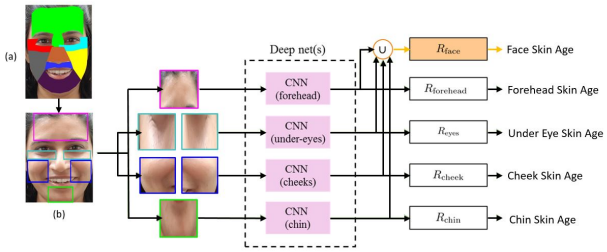

Figure 1. Algorithm flowchart of the proposed facial skin age estimation approach.

1.1 Related Work

Currently, there is no standard approach to objectively determine the perceived age of a individual that is based exclusively on facial skin. However several models and techniques have been proposed that automatically predict the perceived chronological age using holistic facial features. Recently, Escalera et al. released the ChaLearn dataset [13] which includes ground truth collected from multiple annotators that associate a perceived age with each subject based on overall image appearance. The apparent age of a subject was then estimated using the mean of the perceived age opinions among annotators. This dataset was used to train and test the DEX method[6]. Another approach that is described in [4] uses this dataset to train a network in 3 stages; the first stage trains a face identification model, the second stage fine-tunes the model to predict biological age, and the last stage fine-tunes the network on the ChaLearn dataset. While related to our work, this dataset is not suitable to develop and test our method since it is unclear which facial features in particular were the driving factors for the human-annotated age predictions, and to what extent facial skin contributed. However, we adopt a similar strategy taken by both of these approaches in that we also train our models using a multi-stage approach.

In a biological survey on the aging face process by Albert et al.[3], it is shown that horizontal and vertical craniofacial changes (shape changes) including head circumference, head length, bizygomatic breadth, and head breadth can play an important role in the aging process, as well as changes in skin complexity, and soft-tissue changes such as smoothness and elasticity. It is also noted that perceived age of skin is more heavily affected by environment and lifestyle than craniofacial changes, thus implying that one of the primary factors for the disparity between biological age and perceived age comes from facial skin characteristics. This research is what primarily motivated our use of the data augmention scheme described in 2.2.

2 Deep learning on regional facial skin patches for apparent age estimation

The main objective of this work is to estimate the regional and overall skin age given a frontal face image. Inspired by the successes of deep learning applied to face recognition [8], biological age estimation [5], and recently apparent age estimation [6], CNNs are chosen as a key component of our approach. While previously proposed prediction models implicitly use macro features (eyes, nose, etc.), or even investigate the regional importance of different facial areas for age and gender classification[12], our objective is to estimate the age based on skin alone. Furthermore, because skin age can vary across regions, we estimate the regional and overall facial skin ages using locally extracted patches as shown in Fig. 1.

2.1 Extraction of regional facial skin patches

Our method first normalizes all frontal facial image to a standard size (e.g., 716 pixels in height from the tip of forehead to bottom of the chin while keeping original image aspect ratio). The normalized facial area is then segmented into multiple regions (shown colored in Fig. 1-a) that represent the forehead, left and right cheeks, chin, and under eyes. The region shapes are derived from 68 landmarks extracted using DLIB[15]. Since CNNs require rectangular patches, we obtain the largest inscribing rectangle centered at each region's centroid. Since the rectangles can be quite small in some regions like the cheeks and under-eyes, we increase the rectangle size by 10% to obtain more skin pixels in each patch at the expense of potentially including parts of a macro-feature. Examples are shown in Fig 2. As observed in the figure, contamination by macro-features is limited to a small fraction of the pixels. Finally, the extracted skin patches are resized to 256×256. Four groups of facial regions are defined for our investigation: forehead, under-eyes, cheeks, and chin.

2.2 Training regional CNNs

Due to the limited availability of training data that it suitable to train our skin age models, we use an intermediate dataset similar to the approach described in DEX [6]. However one key difference in our approach compared to DEX is that the images used to train our intermediate and final fine-tuned models come from the same dataset. Additionally, our intermediate model uses images of full faces, which are subsequently fine-tuned on region-specific patches. This difference is explained more clearly in Sec. 2.2.2.

Figure 2. Extracted skin patches for four subjects. Each column, from left to right: forehead, left cheek, right cheek, chin, under right eye right, under left eye.

2.2.1 Datasets

There are two datasets used for training our skin age models. Note that these datasets comprise images of female subjects because they were collected as a part of a skin care application targeted for female consumers. However the methodology readily extends for male subjects and is the subject of a future study.

DS1-36K: A set of 36,000 clinically captured (even lighting, hair pulled back) frontal facial images from 1,172 female subjects with ages ranging from 18 to 67. Each subject has 2 to 87 images taken. The multiple copies for each subject can vary in facial expression such as neutral, smiling, open/close eyes, etc. Some subjects had glasses on.

DS2-Selfie: A selfie dataset collected in the wild that consists of roughly 1.5K subjets with ages ranging from 18 to 75, each with unique identities. Selfies were captured with a wide variety of smartphone cameras. Most of these images have good image quality; however some contained facial expressions and occlusions (hair on forehead, glasses, etc.)

2.2.2 Training regional facial CNNs

Although it is possible to train CNNs directly as regression models for skin age estimation, similar to DEX, we chose to train the CNNs as classifiers because a regressor often requires more training samples in order to converge [6]. Therefore, we train our age regressors separately and only use each of the region CNNs as deep feature extractors. Based on the age distribution of our DS1 dataset, we chose five age classes that led to roughly a balanced number of samples per class: $[0-32], (32-40], (40-48], (48-54], (54-100]$. Due to repeat images, we formed the training and validation sets based on subject ID (80% for training and 20% for testing).

Our progressive training scheme begins with a pre-trained AgeNet that has been designed to predict eight age-groups. Our process first fine-tunes the eight-class AgeNet model into a five-class model using full facial images. This full-face age CNN is then fine-tuned into four separate region-specific CNNs using region-specific skin patches (see Fig. 1 and Sec. 2.1). The rationale for using separate CNN's for each region is primarily based on the observation that the appearance of typical aging characteristics such as wrinkles, spots, and sagging will occur with varying severity at different time intervals across different facial areas [3]. If a single network is used, it would be required to learn multiple representations based on where each skin patch came from, which could be impractical due to macro facial cues being cropped out and micro-features across regions being visually similar.

Table 1 shows the performance of each CNN on the DS1 validation set for our five age-group classification using top-1 and 1-off accuracy. Note that we observe a clear drop in performance when estimating chronological ages from facial parts rather than the full face image (from 83% to less than 59% for any specific region).

	Top-1 accuracy	1-off accuracy
Full Face	49.4%	83.2%
Forehead	33.6%	58.7%
Under eyes	32.6%	63.2%
Cheeks	30.5%	61.1%
Chin	35.3%	65.3%

Table 1. Performance of full and regional facial skin age CNNs on DS1 validation set for our five age-group classification

3 Training the age regression models

As shown in Fig. 1, there are five age regression models to be trained: one for each facial ROI plus one additional regressor that combines features from all regions to estimate the overall skin age for the full face. They are denoted as R_{forehead}, R_{eyes}, R_{cheek}, R_{chin}, and R_{face} in this paper. The inputs to each of the the regional age regressors are the deep features of the last fully-connected layer, fc7, extracted from the corresponding fine-tuned region-specific AgeNet network[5]. The inputs to the full-face skin age regressor R_{face} are the concatenated deep features from the four parts: forehead, under-eyes, cheek, and chin. That is, we choose early-fusion as our strategy for aggregating regional skin age information to the full face.

As mentioned earlier, because we did not have the (regional) skin age ground-truth, the region-specific CNNs were trained with regional skin patches while using subject's chronological age as the surrogate ground-truth. If all of the CNNs perfectly model the training dataset, our method would estimate all regional skin ages to the same ground truth value, which is not the intended outcome. Our main goal is to learn a model that can generate meaningful/relevant deep features

to model the regionally varying characteristics of skin ages. Hence, we use a two-stage data augmentation approach that is illustrated in Fig. 3-a.

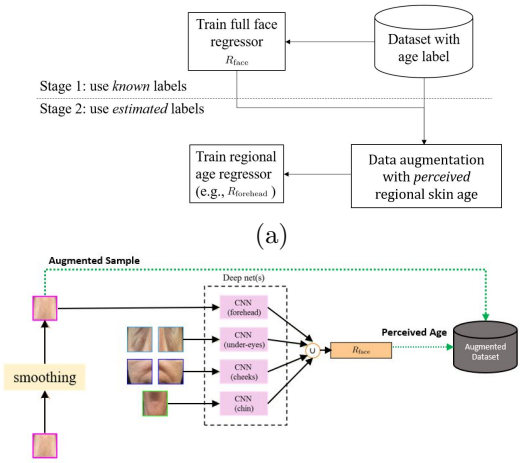

Figure 3. In (a) we show the two-stages for training each regional age regressor. In (b) we show an example of this process for forehead skin patches.

3.1 Stage 1: Full face skin age regressor

First, we train the full face skin age regressor (R_{face}) using all subjects from the DS1 dataset and the self-reported biological age as ground-truth. For subjects with multiple images, we randomly select 5 images and calculate a single deep feature vector for each face region. We do this by averaging together the deep feature vectors generated for each of the regions extracted from each of their images. The purpose of having only one deep feature vector per subject is to have a balanced representation for each subject so that the model training is not biased.

We use standard SVM regression for training R_{face}. Here, 80% of the subjects are randomly pooled for training and the remaining 20% are used as testing. The results are shown in Fig. 4. Two metrics are used as performance indicator for training and testing sets: mean absolute error (MAE) for age estimation, and R^2 for correlation between actual and estimated chronological ages. Values of these performance metrics for training and testing sets are labeled in the triplets provided in the caption of each figure, respectively. For our testing dataset, our aggregate full-face model outperforms AgeNet on entire full-face images, even though AgeNet has access to macro-facial features such as the eyes, nose, mouth, etc. We believe that one explanation for this performance gain may be due to the availability of more fine level of skin features that are not resolvable when a full face images is re-scaled to AgeNet's input 256x256 resolution.

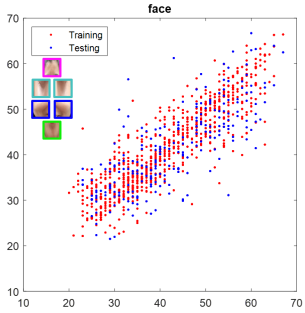

Figure 4. Performance of full face age regressor in estimating chronological ages. The performance indicators (MAE,R^2) for training and testing samples are $(3.2, 0.82)$ & $(5.6, 0.59)$, respectively.

3.2 Stage 2: Regional facial skin age regressors

To train the skin age regressors, we can no longer use the chronological age as ground-truth as was done in Stage 1 since our overall goal is to regress apparent skin age. To address this, we developed a data augmentation scheme by borrowing a technique used to develop a perceived image quality metric [10], where target images with different degrees of image quality defects come from image simulation while their labels (i.e., quality rating) come from psycho-physical experimentation with human observers. Figure 3b illustrates our data augmentation scheme. The overall idea is to augment the original training dataset by applying different amounts of smoothing (for top-hat kernels) on each region of each subject in the training set and then estimating the apparent age in two steps: first, we process the smoothed region patch through the corresponding region CNN and concatenate the resulting deep feature vector with those extracted from other regional CNNs to form a single feature vector. Next, this vector is processed through the full-face skin age regressor R_{face} to estimate the full-face skin age. We therefore treat the latter as labeled perceived age ground-truth and use this augmented dataset to train each regional age regressor R_{forehead}. Note that the concatenated deep feature vector for R_{face} is similar to the original feature vector, except the portion corresponding to the smoothed region. We expect that the estimated full-face skin age with a smoothed subregion would be lower than the original one, which is confirmed in Fig. 5. We remark that it may be beneficial to use different sets of smoothing levels for different regions rather than using a fixed set of smoothing levels for all regions while performing this data augmentation. We note also that this data augmentation procedure is very similar to how the Jacobian is computed to estimate the sensitivity of a dynamic system [11]. The data augmentation approach verifies the encoding of the desired property, smoother-means-younger, in the

full-face skin age regressor and can be used to explain how the full-face skin age regressor weighs each facial region in making its decision. This is an aspect that we wish to investigate further in our future work.

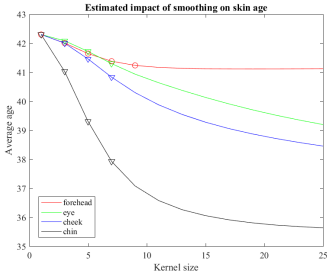

Figure 5. Trend of predicted skin age as smoothing increases.

We use an $80\% - 20\%$ split for training and testing the correlation between chronological age and our apparent age prediction in Fig. 6. Comparing these plots to Fig 4, we observe that the current full-face skin age estimation is better than the regional skin age estimation. This is expected since we have stronger modeling capacity for overall skin age model. In addition, the chronological ages (labels provided for our modeling) are likely to be a better indicator for full-face skin age compared to the skin ages of individual facial regions. However, from the MAE accuracy perspective, all models perform similarly with the exception of the forehead model. During the data augmentation for training regional regressors, we discovered that the forehead region has the smallest slope in driving full-face skin age change (see Fig. 5). That is, with the same amount of spatial smoothing in various facial regions, the smoothed forehead has the least change in the estimated full-face skin age. Another useful metric for judging the accuracy of our model is the spread of the age estimation errors. Using 90 percentile as the range, four models have the resolution of around ± 9 years while forehead skin-age model has the resolution of ± 13 years (see first column of Table 2). Although this spread may appear quite large, when comparing to the AgeNet performance in estimating age, our result is comparable. In addition, it is worth mentioning that AgeNet leverages macro-features while our method does not.

In summary, we train the age regression models in two stages. The full-face skin age regression is trained first using the chronological age as the surrogate of skin age. The regional skin age regression models are then subsequently trained using our data augmentation scheme which acts as a proxy for perceived age. Because a perceived age prediction is given for each facial region independently, we are then able to estimate the (coarse) regional variations of skin age relative to a subject's full-face skin age.

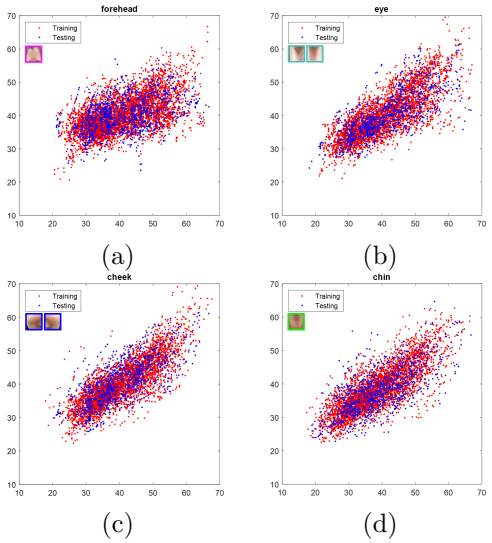

Figure 6. Performance of facial part age regression models for predicting chronological ages. The performance indicators (MAE, R^2) for training and testing samples are the following: (a) $(5.8, 0.33)$ & $(6.6, 0.19)$ for forehead, (b) $(4.3, 0.62)$ & $(4.6, 0.57)$ for eye, (c) $(3.9, 0.66)$ & $(4.6, 0.55)$ for cheek, and (d) $(4.1, 0.63)$ & $(4.9, 0.50)$ for chin.

4 Application to Selfie Images & Apparent Age Study

In the aforementioned experiments, the data used to train and evaluate our models consisted of laboratory controlled face images captured with ideal lighting conditions, fixed capture distance, hair pulled back, no-makeup, etc. In order to determine how well this model transfers to a real world application with smartphone selfie captures in the wild, we collected a selfie dataset that consists of roughly 1.5K facial images, each with unique identities, captured with a wide variety of smartphone cameras. While a large number of images captured were of good image quality, many contained one or more of facial expressions, poor lighting, facial occlusions (hair covering forehead, glasses, etc.), and graininess.

Using the same method described in the previous section (including the augmented smoothing stages), we train and test regression models using the 1.5K selfie images. Table 2 provides the accuracy of this model, as well as the original DS1 trained model discussed in previous sections. An interesting result from this experiment is that, for each of the facial regions, the overall error reduces significantly (approximately a 20% reduction) for the model trained on the 1.5K dataset. This is likely due to the increased number of unique subjects available (roughly 3 times as many)

compared to that in DS1.

	DS1-36K	DS2-Selfie 1.5K
Full Face	±8.6	±8.0
Forehead	±12.7	±9.1
Under eyes	±9.5	±7.8
Cheeks	±8.6	±7.3
Chin	±9.5	±7.8

Table 2. Range of difference (in years) betweem chronological age and predicted skin age at 90th percentile for original 35K model and selfie 1.5K model.

In our final experiment we test the hypothesis that our model has indeed learned to determine apparent age based on skin features alone. We first collected a subset of image pairs from our DS2-Selfie dataset, where each pair shows two subjects with the same chronological age, but with different predicted skin ages. A total of 2-3 pairs were collected for each age between 18-65 to create a set of 112 pairs. Next, two version of each image pair were generated; one version with macro-features present, and a masked version with macro-features hidden. A total of 15 judges were then asked to rank each image pair by selecting the subject they perceived to be older. Table 3 contains our results from this study, from which a few observations can be made: (1) across most age groups, agreement between the judges and our skin age models was higher when macro-features were masked, thereby indicating that our models in fact learned to only associate skin features with age; (2) consistency between the human judges and model predictions are closer aligned for older age groups.

Age	Pair Count	No Mask	Mask	All
≤ 25	21	.53	.50	.52
25-34	27	.59	.70	.64
35-44	25	.64	.76	.69
45-54	23	.61	.60	.61
≥ 55	16	.56	.73	.65
Overall:	**112**	**.59**	**.66**	**.62**

Table 3. Agreement between human judges and our model for picking a subject with an older apparent skin age when provided images of two subjects that have the same chronological age.

5 Conclusion

We present a novel method for estimating apparent facial skin age using a set of region-specific CNNs for feature representation, in conjunction with region-specific SVM regression models. Due to the lack of both standards for defining skin age, as well as benchmark datasets with labeled ground-truth on apparent regional skin age, we introduce a progressive fine-tuning scheme for training region-specific CNNs. We have shown positive agreement with human raters for ranking the skin age of subjects with the same chronological age. For future work, we believe that datasets with standard quantitative assessments of region-specific apparent facial skin age (e.g., through psychophysical or clinical approaches) will significantly improve progress in this relatively nascent research area.

References

[1] Jennifer R.S. Gordon and Joaquin C. Brieva, "Unilateral Dermatoheliosis," *New England Journal of Medicine*, vol.366, no.16, pp.e25, 2012 1

[2] Y. Fu, G. Guo and T. S. Huang, "Age Synthesis and Estimation via Faces: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955-1976, Nov. 2010. 1

[3] A. Midori Albert, Karl Ricanek, Eric Patterson, A review of the literature on the aging adult skull and face: Implications for forensic science research and applications, *Forensic Science International*, Volume 172, pp. 1-9, 2007 2, 3

[4] X. Liu et al., "AgeNet: Deeply Learned Regressor and Classifier for Robust Apparent Age Estimation," *2015 IEEE International Conference on Computer Vision Workshop*, Santiago, pp. 258-266, 2015 2

[5] G. Levi and T. Hassner, *Age and Gender Classification using Convolutional Neural Networks*, *IEEE Workshop on Analysis and Modeling of Faces and Gestures*, 2015. 2, 3

[6] R. Rothe, R. Timofte, and L. V. Gool, *DEX: Deep EXpectation of Apparent Age from a Single Image*, *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, 2015, pp. 252–257. 1, 2, 3

[7] Y. Chen, Z. Tan, A. P. Leung, J. Wan, and J. Zhang, Multi-Region Ensemble Convolutional Neural Networks for High-Accuracy Age Estimation, *Biometric Recognition*, British Machine Vision Conference, 2017.

[8] O. M. Parkhi, A. Vedaldi, and A. Zisserman, *Deep Face Recognition*, British Machine Vision Conference, 2015. 2

[9] P. J. Matts, B. Fink, K. Grammer, and M. Burquest, Color homogeneity and visual perception of age, health, and attractiveness of female facial skin, *Journal of the American Academy of Dermatology*, 57(6), 2007, pp. 977–984. 1

[10] W. Wu, and E. N. Dalal, *Perception-based line quality measurement*, Proceedings SPIE 5668, Image Quality and System Performance II, 111, 2005. 4

[11] D. K. Arrowsmith, and C. M. Place, *Dynamical Systems*, London: Chapman & Hall. ISBN 0-412-39080-9, 1992. 4

[12] Z. Liao, S. Petridis, M.Pantic, Local Deep Neural Networks for Age and Gender Classification, *arXiv:1703.08497* 2

[13] S. Escalera et al., "ChaLearn Looking at People 2015: Apparent Age and Cultural Event Recognition Datasets and Results" *IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, 2015, pp. 243-251. 2

[14] https://olayskinadvisor.ca/ 1

[15] Davis E. King, Dlib-ml: A Machine Learning Toolkit, *Journal of Machine Learning Research*, 10, 2009, pp. 1755–1758 2

[16] Michael Flagler et al., New biological insights into skin aging around the eye, *American Academy of Dermatology*, Volume 74, Issue 5, Supplement 1, 2016 1