Age Estimation From Facial Parts Using Compact Multi-Stream Convolutional Neural Networks

Marcus de Assis Angeloni[1,2], Rodrigo de Freitas Pereira[1], Helio Pedrini[1]
[1]Institute of Computing, University of Campinas (UNICAMP), Campinas, SP, 13083-852, Brazil
[2]AI R&D Lab, Samsung R&D Institute Brazil, Campinas, SP, 13097-160, Brazil
marcus.angeloni@ic.unicamp.br, rodrigodefreitas12@gmail.com, helio@ic.unicamp.br

Abstract

Age is a very useful property in the characterization of individuals, since it is an inherent biological attribute and plays a key role in many real-world applications such as preventing purchase of alcohol and tobacco by minors, human-computer interaction, soft biometrics, electronic customer relationship and as age synthesis in Forensic Art to find lost people. The aging process is influenced by external (health, lifestyle, smoking) and internal (genetics, gender) factors, which makes its estimation difficult for humans, and even more difficult for machines. In this work, we present and evaluate an age estimation approach in unconstrained images using facial parts (eyebrows, eyes, nose and mouth), cropped from the input images using landmarks, to feed a compact multi-stream convolutional neural network (CNN) architecture. Experimental results obtained in the challenging Adience benchmark with real-world images labeled with their respective age groups show that our method is competitive with the literature, even with a significantly smaller CNN and lower computational cost.

1. Introduction

In recent years, age estimation research has gained significant attention with many papers in journals and conferences published annually [2,20,22,26,28,30]. Age is a very useful property in the characterization of individuals, since it is an inherent biological attribute and plays a fundamental role in many real-world applications such as preventing purchase of alcohol and tobacco by minors [28], human-computer interaction, soft biometrics (improving the recognition accuracy in biometric systems and indexing large scale searches), electronic customer relationship (targeting specific customers in same age group for specific advertisements) and as age synthesis in Forensic Art to find lost people [2, 10, 22].

Human face is one of the most prominent characteristics to infer age, as it exhibits remarkable changes in its appearance related to its contour, shape of facial features (eyes, nose, mouth, etc.) and their distribution, skin pigmentation and appearance of wrinkles, among others [10, 22]. This aging process is influenced by external (health, lifestyle, smoking) and internal (genetics, gender) factors, which makes age estimation difficult for humans, and even more difficult for machines [2, 10, 22, 26].

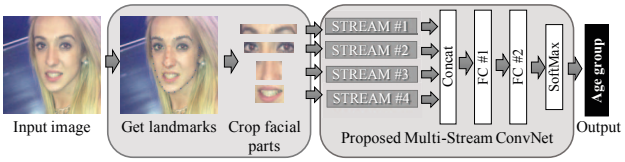

Figure 1. The proposed pipeline has a dedicated and compact stream of convolutional neural networks for each facial part and uses a multilayer perceptron to combine them.

Face age estimation is defined as the possibility of "labeling a face image automatically with the exact age (e.g., 30 years) or the age group (e.g., young, adult, 8-13 years old) of the individual face" [10, 20].

Recently, Eidinger et al. [9] provided a new and challenging data set for age and gender classification, named Adience, which is composed of face images captured in real-world conditions. They used Local Binary Patterns (LBP) descriptor variations and a dropout-SVM classifier for age estimation task. Similarly, Cirne & Pedrini [6] proposed an automatic age estimation using a combination of textural and geometric features from face images. Although our proposed method uses the same benchmark data set, it outperforms these results, using deep learning instead of handcrafted features.

In fact, with the popularity of deep learning [12, 18], the process of designing features has been automated by integrating feature extraction and classifier training in the learning process. Thus, features are learned considering the most important aspects of the data and target task, result-

ing in more robust features. Levi & Hassner [19] proposed learning representations for age estimation using a simple architecture of deep convolutional neural networks (CNN) with a full face image as input or a set of overlapping face regions. Another CNN architecture in age classification was proposed by Rattani et al. [23] focusing only on an ocular crop region. Rothe et al. [24] ensembled the age prediction of twenty VGG-16 [27] networks pre-trained on ImageNet data set followed by a refinement of the softmax expected value. In a more recent work, Rothe et al. [25] transformed the age regression into age classification problem and achieved better results.

Our proposed approach uses deep learning for age estimation, but adopts facial parts as input and a compact multi-stream CNN as architecture, which were not considered in these earlier studies.

The idea of facial parts adopted by our work is different from the patch-based approach [21], in which the facial region is split into blocks of the same size but without considering fiducial points. Our facial parts were inspired by the works of Angeloni & Pedrini [8] and Bonnen et al. [5], which use facial landmark coordinates to crop and align the images. However, we resize each part in a way that they have a similar area in pixels and use them to feed our deep network.

A similar idea was proposed by Yi et al. [32], but they extracted a set of local aligned patches, in grayscale, around some facial landmarks to fed 23 sub-networks from a multi-task CNN. These patches were extracted in four scales and, in the larger scale, the patches were composed of the entire face. On the other hand, our method has only four sub-networks (streams), we group the landmarks related to the same facial part before crop, and we directly fed into our CNN the color facial part.

About multi-stream CNN, Wang et al. [31] evaluated a very deep two-stream model for action recognition. This is slightly different from what we present in this work, since we train all streams together, our input has the same nature (raw RGB image), all streams contribute to loss function, and each stream has significantly fewer parameters than the work mentioned.

It is also worth noting that our work is different from a recent method proposed by Li et al. [20], which consisted in a continuity-aware probabilistic network for age estimation. In that work, they extracted features from aligned faces using VGG [27], and used local regressors in multiple overlapping sub-spaces to tackle heterogeneous data and gating networks learn weights for their results by employing a bridge-tree structure. Our method uses a significantly smaller CNN than VGG and combines facial parts instead of sub-spaces, through a single multi-stream network architecture.

For more comprehensive literature reviews, we refer to Osman & Yap [22], Angulu et al. [2] and Sawant & Bhurchandi [26].

This work focuses on proposing and evaluating an entire automatic pipeline of a compact multi-stream convolutional neural network architecture to explore preprocessed facial parts in order to estimate human age from a single image captured in a real-world scenario, as shown in Figure 1. The main contributions of this work include: (i) preprocessing of facial parts using only landmarks extracted by open source toolboxes; (ii) a fully differentiable and compact multi-stream CNN, allowing end-to-end learning from facial parts to age estimation; (iii) a reproducible experimental procedure for further extension of the obtained results.

The paper is organized as follows. Section 2 presents the proposed method. Section 3 reports the experimental settings and results. Finally, conclusions are drawn in Section 4.

2. Method

The main goal of this work is to propose and evaluate a compact multi-stream convolutional neural network architecture to explore preprocessed facial parts in order to estimate human age from a single image. It is important to highlight that we do not use any manual annotation of facial localization or landmarks in the entire pipeline.

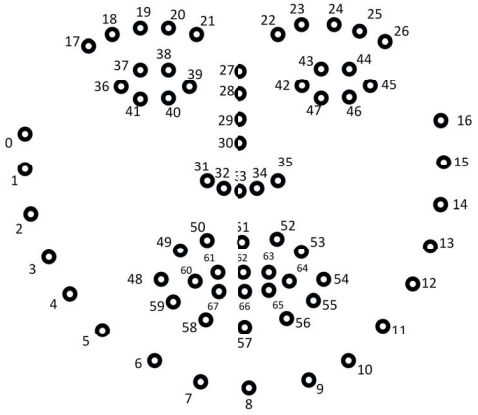

Figure 2. Locations of facial landmarks and their indices [33].

The overall pipeline of our method is illustrated in Figure 1. A single RGB image is input to the system and a face detector is applied followed by a 2D facial landmarks estimator. Thus, based on the landmark coordinates, the facial parts of interest are preprocessed and cropped. Each facial part feeds a specific stream of CNN, whose outputs are concatenated and processed by a sequence of fully connected layers. Finally, a softmax function returns the probabilities of the person belonging to each age group, such that the estimation will be the highest probability group.

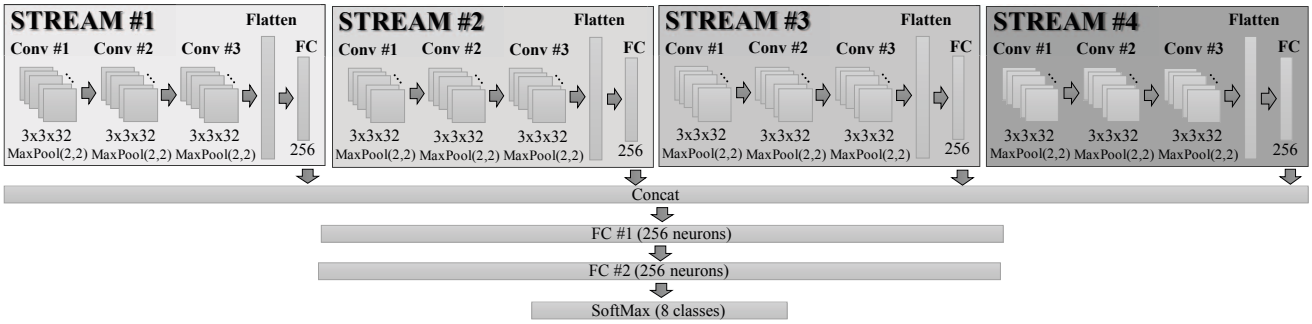

Figure 3. Architecture of the proposed Compact Multi-Stream of Convolutional Neural Network.

The following subsections present each step of the proposed method in more detail.

2.1. Pre-processing

Face detection is the first step of the pipeline, performed in the input RGB image. The bounding box returned around the face is the initialization of the facial landmark detector. These 2D landmark locations are used to align, crop and resize each facial part.

The OpenFace [4] detector, an open source facial behavior analysis toolkit, was used in this process, where among the available implementations there are a facial landmark detection and tracking [3,33]. In situations where more than one face is found, the face whose center is closest to the image center is chosen for the landmark detection. On the other hand, when no face is found, we used the DLib [16] detector, another open source toolkit that contains machine learning algorithms, such as face detection and landmark localization [15]. The order of the detectors was chosen based on experimental quantitative results using a set of face images in the wild. The coordinates returned from both toolkits are compatible, with the 68 points shown in Figure 2.

Using these landmark positions, the facial parts were extracted comprising both eyebrows, both eyes, nose and mouth. These chosen facial parts, as well as the sizes of each one, were inspired by previous works [5, 8], in order to avoid overlap among the parts. The eyebrows include the region of all coordinates between 17 and 26 expanded by a small proportion in each horizontal direction. Before cropping, we align the region using the points of the inner corner (21 and 22). In a similar way, the eyes cover the extended region of the coordinates from 36 to 47, and adopt the points 39 and 42 for alignment. The coordinates 27 to 35 are adopted to crop the nose region, which uses for aligning the points 32 and 34. Finally, the mouth includes the region of coordinates 48 to 67, and it uses to align the points of the outer corner (48 and 54).

2.2. Proposed Multi-Stream CNN

The overview of our proposed network is shown in Figure 3, in which each aforementioned facial part is processed by an independent compact CNN stream prior to concatenation with other parts. Thus, the feature learning occurs before concatenating each facial part information. Each stream is composed of three pairs of 32 convolutional filters 3×3 and maxpooling for dimensionality reduction. At the end of each stream, the output of the last convolutional layer is flattened and input into a dense layer for further concatenation. Moreover, inspired by the work of Szegedy et al. [29], the output of each dense layer is used to perform the classification of each stream individually. In other words, the classification is not only performed by the last dense layer shown in Figure 3, but also by each stream individually. This classification by each stream is performed by a second 8-unit dense layer, which is used to compute the facial part loss and adopted only in training mode.

For back-propagation purposes, the loss of each stream and the final loss were all summed up without any weighting, as shown in Equation 1.

$$loss_{global} = loss_{final} + loss_{reg} + \sum_{i=1}^{n} loss_{part_i} \qquad (1)$$

where $loss_{final}$ is the loss of the last dense layer, $loss_{reg}$ is the regularization loss and $loss_{part_i}$ is the loss of the i-th stream of facial part.

In preliminary experiments, it was concluded that this approach leads to better results than just performing classification using only the last dense layer. However, each stream estimator is not used at inference time, but only the final softmax. The concatenated streams are then forwarded through two dense layers of 256 units and finally into an 8-unit softmax layer.

Both fully connected and convolutional layers use Exponential Linear Units (elu) [7] as the activation function. Variance Scaling Function [13] is used as weight initialization since it was shown to lead to better results with non-

sigmoid activation functions compared to Xavier initialization [11]. The dropout rate is set to 0.4 and the batch size to 32. To achieve 100 epochs, 40,000 steps were performed during training. Once dealing with a classification problem, Cross-Entropy was chosen as a loss function. Adam algorithm [17] was chosen as optimizer with initial learning rate of $1e-4$ and L2 regularization was chosen for all layers. The proposed network is completely differentiable, allowing end-to-end learning, from the input facial parts to the age estimation.

3. Experiments

Our proposed multi-stream CNNs were implemented using Tensorflow 1.8.0 [1], an open source library for deep learning. The training was performed on an Intel i7-3770K 3.50GHz processor and Nvidia Geforce GTX 1080 GPU. The network training for all 5 folds required about 1.5 hours (including the time spent to save checkpoints for each 500 iterations), whereas age estimation required about 6 ms. However, inference time can be improved by running the network on image batches, since our input uses a reduced amount of memory. Note that our reported times are measured from inputting the preprocessed facial parts to outputting the results.

3.1. Data Set

We evaluated our method using the Adience benchmark [9, 19], which was designed for age and gender classification for face images captured in challenging real-world conditions. The Adience data set consists of images automatically uploaded to Flickr from smartphone devices. Since these images were uploaded without prior manual filtering, they are highly unconstrained with challenges such as extreme variations in head pose, occlusion, lightning conditions, and quality. Some images from Adience data set are shown in Figure 4 .

Figure 4. Sample images from the Adience data set.

The entire data set includes 26,580 images of 2,284 subjects, with eight unbalanced age group classes (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-above). However, we used the in-plane aligned version of the faces (19,370 images), originally used in [9, 19] in order to have a fair comparison with these previous approaches and isolate the performance gain attributed to our method.

3.2. Evaluation Protocol

As evaluation protocol, we used the standard 5-fold cross validation experiment defined for Adience benchmark [9], which is subject-exclusive and it was publicly available in [19]. For each fold, the training set is composed of around 72% of the images, validation set of 8% and test set of 20%, with samples from each age group stratified. The exact size for each set by fold is shown in Table 1, whose difference between them is mainly due to the subject-exclusive, since each subject can have different amount of images.

Table 1. Number of images in the Adience benchmark and accuracies of our proposed method for each fold.

Fold	Train	Val	Test	Exact (%)	1-off (%)
0	11,823	1,284	4,316	56.60	86.40
1	12,894	1,428	3,101	44.76	81.20
2	12,655	1,429	3,339	54.42	85.18
3	12,391	1,457	2,975	50.12	85.31
4	12,371	1,359	3,693	49.23	78.99

We performed a shallow analysis of the aligned data set, and we could observe that 96 images appear more than one time in the test set for a specific fold, 2,045 images are not used in any test set and some trials have wrong age group label, perhaps due to some failure during the alignment process in the data set preparation. Nevertheless, we use the data set and protocol as they are, so that the comparison is fair.

The two performance metrics adopted were exact accuracy, that is, when the method detects the exact age group of the input image, and 1-off accuracy, that is, when the method is off by one adjacent age group (predicts the immediately younger or older group than ground truth group). Since there are five folds, the comparison is performed based on the average and standard deviation of these metrics across all folds, following previous methods [6, 9, 19, 23].

3.3. Results

Following the pipeline of the proposed method presented in Section 2, we detect the face and its landmarks using OpenFace 2.0.0 [4], which worked for 19,160 images from data set. In the remaining 210 images, we applied Dlib 19.9 [16] and it worked for 53 of them. For the last 153 images, we provided the entire image as input to landmark detector in order to process the whole data set. A negative list was created with images in which both toolboxes could not find faces and images without the eyebrow region. Then, the idea was not to use these images as training

and validation sets, thus they only were used in the test set.

With the returned facial landmarks, we aligned and cropped each facial part while maintaining its aspect ratio. Inspired by the tuning and preprocess performed in [5, 8], we expanded by 3% the eyebrow region for each horizontal direction and resized it to 228×33 pixels. The same expansion was applied to the eye region and was resized to 202×38 pixels. The nose region was enlarged by 40% and resized to 103×73 pixels. Finally, the mouth region was enlarged by 8% and resized to 114×66 pixels. An example of these crop regions is illustrated in Figure 1.

The idea behind the chosen sizes was that each facial part has similar amount in pixels as input to the CNN. Furthermore, the sum of all facial part pixels ($7,524 + 7,676 + 7,519 + 7,424 = 30,143$) is smaller than an image of 174×174 (= $30,276$) pixels. Just as a reference, most popular CNNs have as input images larger than 200×200 pixels.

In addition to exploring a reduced input size for our network, we proposed a compact deep learning architecture designed to avoid overfitting due to limited labeled training data (we did not use data augmentation) and also composed of a reduced number of parameters. Thereby, our method does not require a sophisticated hardware to run and can be applicable even to mobile devices. The number of parameters of our proposed multi-stream CNNs is presented in Table 2. It is important to mention that the *Logits* layers are only used in training mode, which compute the facial part loss and contribute to a global loss for back-propagation purposes. This is the reason for the difference in total parameters for training and inference.

Table 2. Number of parameters of our proposed multi-stream CNNs.

Layer	**Eyebrows**	**Eyes**	**Nose**	**Mouth**
Conv #1	896	896	896	896
Conv #2	9,248	9,248	9,248	9,248
Conv #3	9,248	9,248	9,248	9,248
FC	1,188,096	1,065,216	1,065,216	1,106,176
Logits	2,056	2,056	2,056	2,056
Layer	**Concatenated**			
FC #1	262,400			
FC #2	65,792			
SoftMax	2,056			
Total	Train	**4,840,744**	Inference	**4,832,520**

As can be seen in Table 2, our proposed network has 4,832,520 parameters. It is very compact when compared to other popular CNNs such as VGG-16 (138 million), AlexNet (60 million), ResNet-50 (25 million), Inception V3 (23.2 million) and GoogleNet (6.8 million) [14]. Even when compared to Levi & Hassner network [19], which has 11,413,280 parameters and was designed to be shallow, our approach is significantly smaller.

Our experimental results in Adience benchmark are initially presented by fold in Table 1, following the evaluation protocol described in Section 3.2. The number of images are reported in each set from the second to fourth columns, whereas the last two columns present the evaluation metrics. The exact accuracy by fold has a significant standard deviation, probably due to the subjects selected by each one and the variance of the trials by age group in the test set. When we consider the 1-off accuracy, it is possible to observe an improvement in recognition and a smaller standard deviation among the folds.

The average and standard deviation of our metrics across the folds are reported in Table 3, in comparison to some previous methods. It can be seen that the proposed method achieves overall exact accuracy of 51.03% and 1-off accuracy of 83.41%.

Table 3. Age estimation results on Adience benchmark.

Method	**Exact (%)**	**1-off (%)**
Eidinger et al. [9]	45.1 ± 2.6	79.5 ± 1.4
Cirne & Pedrini [6]	46.70 ± 6.56*	81.80 ± 2.23*
Rattani et al. [23]	46.97 ± 2.90*	80.96 ± 1.10*
Levi & Hassner (single crop) [19]	49.5 ± 4.4	84.6 ± 1.7
Levi & Hassner (ov.sample) [19]	50.7 ± 5.1	84.7 ± 2.2
Proposed approach	51.03 ± 4.63	83.41 ± 3.17

We can notice that our method outperforms existing non-deep approaches [6, 9] and a deep approach that uses the ocular region [23] in terms of exact and 1-off accuracy. It is worth mentioning that Cirne & Pedrini [6] and Rattani et al. [23] evaluated their respective methods only on frontal faces of the Adience benchmark, which is a significantly less challenging scenario than the one adopted in our proposed approach. Moreover, our method achieves a slightly better exact accuracy in relation to Levi & Hassner method [19], whose network has twice the parameters of ours and an input image of 227×227 pixels. On the other hand, our achieved accuracy is not statistically higher than that obtained by the over-sample case of [19], but in their scenario its final prediction was taken by the average value of 10 input regions, covering different regions and reflections of the input image.

3.4. Ablation Study

In Table 4, we show the results varying the number of streams as an ablation study, separately removing each one of the facial parts to evaluate its contribution.

Table 4. Age estimation results on Adience benchmark by removing each facial part stream.

Method	**Exact (%)**	**1-off (%)**
Proposed approach w/o eyebrows	49.35 ± 5.05	81.64 ± 2.79
Proposed approach w/o eyes	46.02 ± 5.31	80.42 ± 3.20
Proposed approach w/o nose	47.47 ± 7.00	80.63 ± 3.93
Proposed approach w/o mouth	46.44 ± 5.71	80.97 ± 2.97

We noticed that the absences of eyes and mouth con-

tributed most to reducing the accuracy, demonstrating to be the most prominent. However, the best result is obtained when the four streams are combined, showing that they are complementary.

In another closer analysis of our results, through the confusion matrix between the age groups presented in Table 5, we can observe that, for some classes as 0-2 and 25-32, good results were achieved. In fact, 0-2 is an age group with less intra-class variations than older ones and with a clear inter-class variation. On the other hand, 25-32 is the group with more labeled images in the Adience data set. The worst results were obtained to 48-53 and 15-20 age groups, for which the adjacent age groups were incorrectly predicted.

Table 5. Age estimation confusion matrix on Adience benchmark.

age	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-
0-2	**0.731**	0.199	0.026	0.005	0.029	0.006	0.001	0.002
4-6	0.154	**0.589**	0.162	0.020	0.057	0.009	0.001	0.007
8-13	0.025	0.177	**0.488**	0.071	0.200	0.025	0.006	0.008
15-20	0.009	0.030	0.096	**0.187**	0.595	0.062	0.006	0.015
25-32	0.006	0.019	0.044	0.082	**0.721**	0.103	0.007	0.019
38-43	0.003	0.015	0.038	0.047	0.535	**0.271**	0.024	0.067
48-53	0.013	0.009	0.041	0.046	0.324	0.282	**0.076**	0.209
60-	0.007	0.012	0.027	0.029	0.221	0.237	0.070	**0.398**

4. Conclusions and Future Work

In this work, we addressed the age estimation task using a compact multi-stream convolutional neural network that employed as input a set of preprocessed facial parts. This method is different from any previous approach. We started detecting face and its landmarks in the input image, followed by an alignment, crop and resize of four chosen facial parts: eyebrows, eyes, nose and mouth. The resized parts have a similar size in pixels and each one is used to feed a specific stream, whose outputs are combined and processed by a sequence of fully connected layers. It is worth emphasizing that our network is completely differentiable, allowing end-to-end learning, from the preprocessed facial parts to the age group.

Experiments were conducted on the Adience data set, a very challenging public benchmark composed of images automatically uploaded to Flickr from smartphones, and its original protocol. This data set covers eight age groups, which are unbalanced. Our method achieved an exact accuracy of 51.03% and a 1-off accuracy of 83.41%, a competitive accuracy when compared to other available approaches. This result is similar to one of the reported methods, but our CNN has a significant reduced number of parameters and input size. Furthermore, that approach used an average value of 10 input regions as final prediction.

The reported accuracy rates show the great challenge of estimating age in the wild, and our method has the opportunity to improve the recognition mainly in some specific age groups (48-53 and 15-20). Since we provided the source code for the proposed approach[1], the entire experimental procedure can be reproduced to regenerate and extend the obtained results.

Some viable future research topics include the investigation of improvements that can be achieved by adopting data augmentation in the CNN training and a stream architecture design specific for each facial part. Furthermore, we intend to provide some quality measure for each part and consider the landmark localization confidence as inputs to the network. Finally, our method can be extended to other face-based tasks, such as gender recognition, face biometrics, anti-spoofing and emotion recognition.

Acknowledgement

The authors thank CAPES, FAPESP (grants #2014/12236-1 and #2017/12646-3), CNPq (grant #309330/2018-1) for the financial support, and NVIDIA for the donation of a GPU as part of the GPU Grant Program.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A System for Large-Scale Machine Learning. In *12th USENIX conference on Operating Systems Design and Implementation*, volume 16, pages 265–283, 2016.

[2] R. Angulu, J. R. Tapamo, and A. O. Adewumi. Age Estimation via Face Images: A Survey. *EURASIP Journal on Image and Video Processing*, 2018(1):42, June 2018.

[3] T. Baltrusaitis, P. Robinson, and L. Morency. Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild. In *IEEE International Conference on Computer Vision Workshops*, pages 354–361, Dec. 2013.

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *13th IEEE International Conference on Automatic Face Gesture Recognition*, pages 59–66, May 2018.

[5] K. Bonnen, B. Klare, and A. K. Jain. Component-Based Representation in Automated Face Recognition. *IEEE Transactions on Information Forensics and Security*, 8(1):239–253, 2013.

[6] M. V. M. Cirne and H. Pedrini. Combination of Texture and Geometric Features for Age Estimation in Face Images. In *13th International Conference on Computer Vision Theory and Applications*, pages 395–401, Jan. 2018.

[7] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[1] https://github.com/marcusangeloni/facialparts_age

[8] M. de Assis Angeloni and H. Pedrini. Part-based Representation and Classification for Face Recognition. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 002900–002905, Oct. 2016.

[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, Dec. 2014.

[10] Y. Fu, G. Guo, and T. S. Huang. Age Synthesis and Estimation via Faces: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, Nov. 2010.

[11] X. Glorot and Y. Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.

[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *CoRR*, abs/1502.01852, 2015.

[14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*, 2017.

[15] V. Kazemi and J. Sullivan. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, June 2014.

[16] D. E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[17] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7553):436, 2015.

[19] G. Levi and T. Hassner. Age and Gender Classification using Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, June 2015.

[20] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian. BridgeNet: A Continuity-Aware Probabilistic Network for Age Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[21] Z. Liao, S. Petridis, and M. Pantic. Local Deep Neural Networks for Age and Gender Classification. *arXiv preprint arXiv:1703.08497*, 2017.

[22] O. F. Osman and M. H. Yap. Computational Intelligence in Automatic Face Age Estimation: A Survey. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–15, 2018.

[23] A. Rattani, N. Reddy, and R. Derakhshani. Convolutional Neural Network for Age Classification from Smart-phone based Ocular Images. In *IEEE International Joint Conference on Biometrics*, pages 756–761, Oct. 2017.

[24] R. Rothe, R. Timofte, and L. V. Gool. DEX: Deep EXpectation of Apparent Age from a Single Image. In *IEEE International Conference on Computer Vision Workshops*, Dec. 2015.

[25] R. Rothe, R. Timofte, and L. V. Gool. Deep Expectation of Real and Apparent Age from a Single Image without Facial Landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[26] M. M. Sawant and K. M. Bhurchandi. Age Invariant Face Recognition: A Survey on Facial Aging Databases, Techniques and Effect of Aging. *Artificial Intelligence Review*, 52(2):981–1008, Aug. 2019.

[27] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] M. Singh, S. Nagpal, M. Vatsa, and R. Singh. Are You Eligible? Predicting Adulthood From Face Images via Class Specific Mean Autoencoder. *Pattern Recognition Letters*, 119:121–130, 2019. Deep Learning for Pattern Recognition.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, June 2015.

[30] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, and S. Z. Li. Efficient Group-n Encoding and Decoding for Facial Age Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2610–2623, Nov. 2018.

[31] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards Good Practices for Very Deep Two-stream Convnets. *arXiv preprint arXiv:1507.02159*, 2015.

[32] D. Yi, Z. Lei, and S. Z. Li. Age Estimation by Multi-scale Convolutional Network. In *Asian Conference on Computer Vision*, pages 144–158. Springer International Publishing, 2015.

[33] A. Zadeh, T. Baltruaitis, and L. Morency. Convolutional Experts Constrained Local Model for Facial Landmark Detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2051–2059, July 2017.