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

Abstract

- *Age is a very useful property in the characterization of individuals*
- *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(*targeting specific customers in same age group for specific advertisements*) 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*
- *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*

Introduction

- Face age estimation is defined as the possibility of "labelling 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.
- our proposed method uses the same benchmark data set, it outperforms these results, using deep learning instead of handcrafted features.
- In deep larning 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, resulting 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.
- 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. 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 multistream 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.
- we resize each part in a way that they have a similar area in pixels and use them to feed our deep network.
- our method has only four subnetworks (streams), we group the landmarks related to the same facial part before crop, and we directly fed into our CNN the color facial part.
- 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
- Our method uses a significantly smaller CNN than VGG and combines facial parts instead of sub-spaces, through a single multi-stream network architecture.
- 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

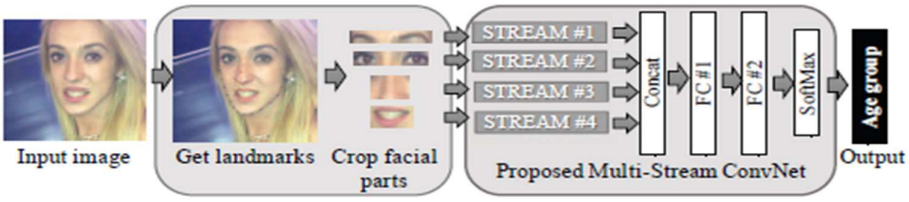

Figure 1. The proposed pipeline has a dedicated and compact stream of convolutional neural networks for each facial part and uses a multilayer perceptron to combine them.

- 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.

Method

1. 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.
2. A single RGB image is input to the system and a face detector is applied followed by a 2D facial landmarks estimator.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.
3. 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.

1. Pre-processing

- The bounding box returned around the face is the initialization of the facial landmark detector
- The OpenFace detector, an open source facial behaviour analysis toolkit, was used in this process
- 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
- The coordinates returned from both toolkits are compatible, with the 68 points shown in Figure 2.

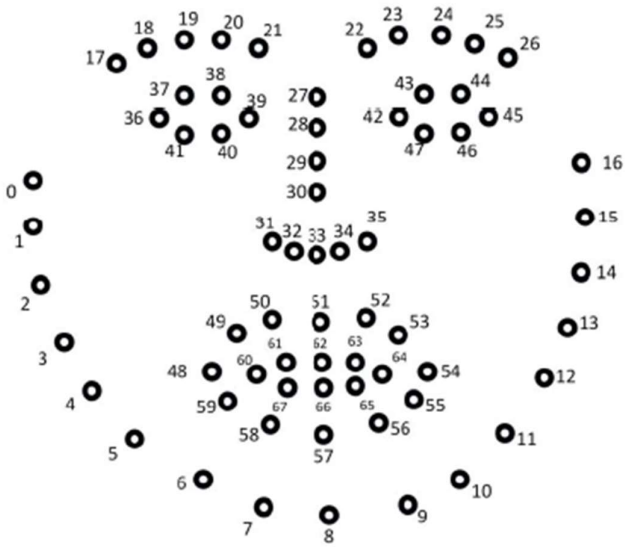

Figure 2. Locations of facial landmarks and their indices [33].

- Using these landmark positions, the facial parts were extracted comprising both eyebrows, both eyes, nose and mouth.

2. Proposed Multi-Stream CNN

The overview of our proposed network is shown in Figure 3,

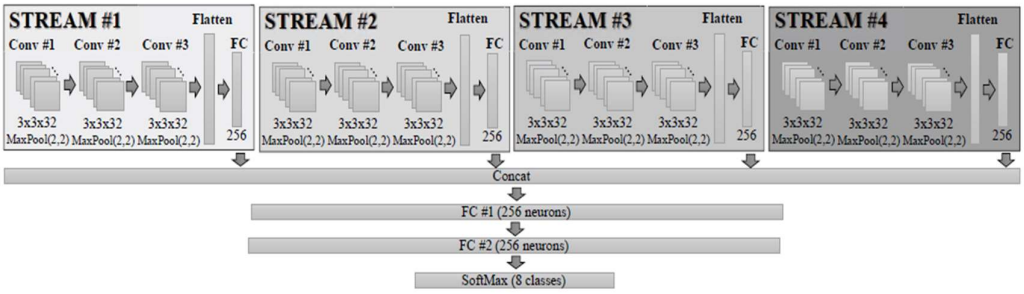

Figure 3. Architecture of the proposed Compact Multi-Stream of Convolutional Neural Network.

- each aforementioned facial part is processed by an independent compact CNN stream prior to concatenation with other parts 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.
- the output of each dense layer is used to perform the classification of 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. 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 $loss_{final}$ = loss of the last dense layer, $Loss_{reg}$ =regularization loss $loss_{part_i}$ =loss of the *i*-th stream of facial part.
- this approach leads to better results than just performing classification using only the last dense layer.
- 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
- the dropout rate is set to 0.4
- batch size to 32
- achieve 100 epochs, 40,000 steps were performed during training
- Cross-Entropy was chosen as a loss function.
- Adam algorithm 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.

Experiments

1. Our proposed multi-stream CNNs were implemented using Tensorflow 1.8.0 [1],
2. The training was performed on an Intel i7-3770K 3.50GHz processor and Nvidia Geforce GTX 1080 GPU.
3. The network training for all 5 folds required about 1.5 hours (including the time spent to save checkpoints for each 500 iterations),
4. age estimation required about 6 ms.

Dataset

- We evaluated our method using the Adience benchmark, which was designed for age and gender classification for face images captured in challenging real-world conditions.
- Adience data set consists of images automatically uploaded to Flickr from smartphone devices.
- these images were uploaded without prior manual filtering,
- 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).

- used the in-plane aligned version of the faces (19,370 images), originally used in [9, 19] in order

Evaluation protocol

- As evaluation protocol, we used the standard 5-fold cross validation experiment defined for Adience benchmark, which is subject-exclusive
- 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 two performance metrics adopted were
  1. exact accuracy i.e., when the method detects the exact age group of the input image,
  2. 1-off accuracy,I.e, 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

Results

1. we detect the face and its landmarks using OpenFace 2.0.0 which worked for 19,160 images from data set In the remaining 210 images, we applied Dlib 19.9 and it worked for 53 of them.
2. For the last 153 images, we provided the entire image as input to landmark detector in order to process the whole data set
3. A negative list was created with images in which both toolboxes could not find faces and images without the eyebrow region
4. we expanded by 3% the eyebrow region for each horizontal direction and resized it to 228×33 pixels.
5. The nose region was enlarged by 40% and resized to 103×73 pixels
6. the mouth region was enlarged by 8% and resized to 114×66 pixels
7. The idea behind the chosen sizes was that each facial part has similar amount in pixels as input to the CNN 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.
8. 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.
9. our method does not require a sophisticated hardware to run and can be applicable even to mobile devices.
10. 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.
11. our proposed network has 4,832,520 parameters. It is very compact when compared to other popular CNNs
12. It can be seen that the proposed method achieves overall exact accuracy of 51.03% and 1-off accuracy of 83.41%.

Conclusions and Future Work

1. we addressed the age estimation task using a compact multi-stream convolutional neural network that employed as input a set of preprocessed facial parts.
2. this method is different from any previous approach
3. 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.
4. This data set covers eight age groups,which are unbalanced
5. our method achieved an exact accuracy of 51.03% and a 1-off accuracy of 83.41
6. our CNN has a significant reduced number of parameters and input size
7. our method can be extended to other face-based tasks, such as gender recognition, face biometrics, anti-spoofing and emotion recognition