eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes
Master Thesis

Comparative Study of Human Age Estimation Based on Hand-crafted and Deep Face Features

---

Carlos Belver

Director:
Fadi Dornaika

Co-director:
Ignacio Arganda Carreras

informatika fakultatea
facultad de informática

2016

Abstract

In the past few years, human facial age estimation has drawn a lot of attention in the computer vision and pattern recognition communities because of its important applications in age-based image retrieval, security control and surveillance, biometrics, human-computer interaction (HCI) and social robotics. In connection with these investigations, estimating the age of a person from the numerical analysis of his/her face image is a relatively new topic. Also, in problems such as Image Classification the Deep Neural Networks have given the best results in some areas including age estimation.

In this work we use three hand-crafted features as well as five deep features that can be obtained from pre-trained deep convolutional neural networks. We do a comparative study of the obtained age estimation results with these features.

Acknowledgements

First, I would like to thank my director Dr. Fadi Dornaika for his help conducting my thesis. I also want to express my gratitude to Dr. Ignacio Arganda for supervising me and for his advice.

I would like to thank Blanca Cases for encouraging me to continue with my studies taking a Master degree.

Finally, I want to thank my family, my friends and my girlfriend for supporting me during this year.

Contents

List of Figures

Table index

1 Introduction

In the last decade, with the increasing interest in social robotics and video-based security systems, research on the numerical analysis of human faces (including face detection, face recognition, classification of gender, and recognition of facial expression) has attracted attention in the communities of computer vision and pattern recognition [1, 2, 3, 4, 5]. In connection with these investigations, estimating the age of a person from the numerical analysis of his/her face image is a relatively new topic. Age estimation by numerical analysis of the face image has many potential applications such as the development of intelligent human-machine interfaces and improving safety and protection of minors in various and diverse sectors (transport, medicine, etc.). It can be very useful for advanced video surveillance, demographic statistics collection, business intelligence and customer profiling, and search optimization in large databases. The age attribute could also be used in the verification of the face and enriching the tools used in police investigations. In general, automatic age estimation by a machine is useful in applications where the objective is to determine the age of an individual without identifying him.

The age estimator can use a machine learning approach to train a model for extracted features and make age prediction for query faces with the trained model. Generally speaking, age estimation can be viewed as a multi-class classification problem, a regression problem or a composite of these two.

The anthropometry-based approach mainly depends on measurements and distances of different facial landmarks. The anthropometry-based approaches might be useful for babies, children, and young adults, but they are impractical for adults since their facial skin appearance is the main source of information about ethnicity, gender, and age.

Estimating human age from a facial image requires a great amount of information from the input image. Extraction of these features is important since the performance of an age estimation system will heavily rely on the quality of extracted features. Lots of research on age estimation has been conducted towards aging feature extraction. Deep learning approaches claim to have the best performances in demographic classification (ethnicity, gender and age). However, this claim cannot be always true. It is known, that deep learning can provide impressive results within a single database. However when another database is used with the trained deep net, the age estimation performance can drop significantly.

1.1 Objectives

The main objective of the project is to do a comparative study of age estimation based on hand-crafted and deep features using Matlab. The original idea was to use some pre-trained neural networks and other feature extractors such as Local Binary Patterns (LBP) or Histograms of Oriented Gradients (HOG), to get the necessary information from the input images in order to estimate the age of each person.

We also work on the alignment of the images so as to get some improvement in the predicted ages.

We use feature selection such as Fisher Score to reduce the dimension of the feature vectors in order to reduce the computation time.

1.2 Related work

Human age estimation from face image has been studied for 20 years. Limited by the technology of facial analysis, early methods mainly used geometric features to estimate the age range of each image, such as baby, young adult and senior adult. Geometry features can discriminate baby and adult easily but cannot distinguish adult and old man. As the improvement of classification accuracy, researchers started to estimate the exact age instead of the coarse age range. For this purpose there have been used aging feature extractors such as the Active Appearance Model (AAM) [6], age manifold [1], AGing pattern Subspace (AGES) [7], Biologically Inspired Features (BIF) [8]. Image-based age estimation approaches view the face image as a texture pattern. Many texture features have been used like Local Binary Patterns (LBP) [9], Histograms of Oriented Gradients (HOG) [10], BIF, Binarized Statistical Image Features (BSIF) [11] and Local Phase Quantization (LPQ) in demographic estimation works. BIF and its variants are widely used in age estimation works such us [12, 13, 14]. Han et al. [14] used selected BIF features in order to estimate the age, gender and ethnicity attributes.

Due to their significant performance improvement in facial recognition domain, deep learning approaches have been recently proposed for age estimation (e.g. [3, 5]). This work shows that the full power of a pre-trained net can be exploited by simply using its deep features and only training an age regressor. This regressor training is much more efficient than retraining or tuning the whole deep net using a sheer number of images.

Experiments will show that this scheme for age estimation can be more accurate than the one obtained by the end-to-end deep net solution.

2 Neural Networks (NN)

The artificial neural network models have been studied for many years with the hope of achieving a similar performance to human behaviour in the fields of speech and image recognition. These models are composed of many nonlinear computational elements that operate in parallel and are based on patterns that resemble biological neural networks. An artificial network consists of a pool of simple processing units which communicate by sending signals to each other over a large number of weighted connections.

First Neural Network simple models appeared in 1943, and in 1958 Frank Rosenblatt created a unidirectional model composed of two layers of neurons (one input and one output, see Figure 1).

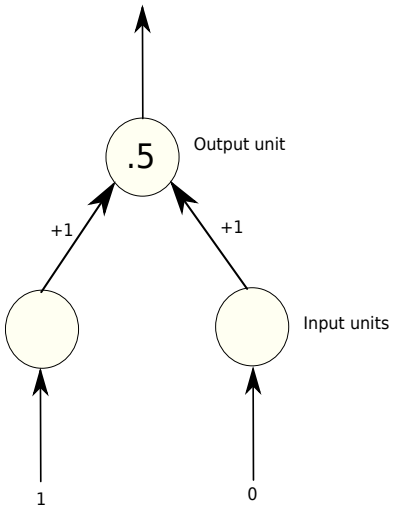


Figure 1: Perceptron: first NN model composed of two layers of neurons.

In its most basic form the perceptron learns a linear discriminant function $f(x)$. This function sets a dichotomy between two linearly separable training sets. Given two sets of points it is said to be linearly separable if there is a (straight) line in the pattern space between both sets of data.

In the late 1970's, researchers discovered that the perceptron cannot approximate many nonlinear decision functions, for example the XOR function. In the 1980's, researchers found a solution to that problem by stacking multiple layers of linear classifiers (hence the name "multilayer perceptron") to approximate nonlinear decision functions.

As we said before, a neural network is composed of some units called neurons (see Figure

2). Each neuron receives some input parameters and produces an output. This output is given by:

- A **propagation function** that generally consist of a summation of each input multiplied by the weight of its interconnection.
- An **activation function** that modifies the previous one.

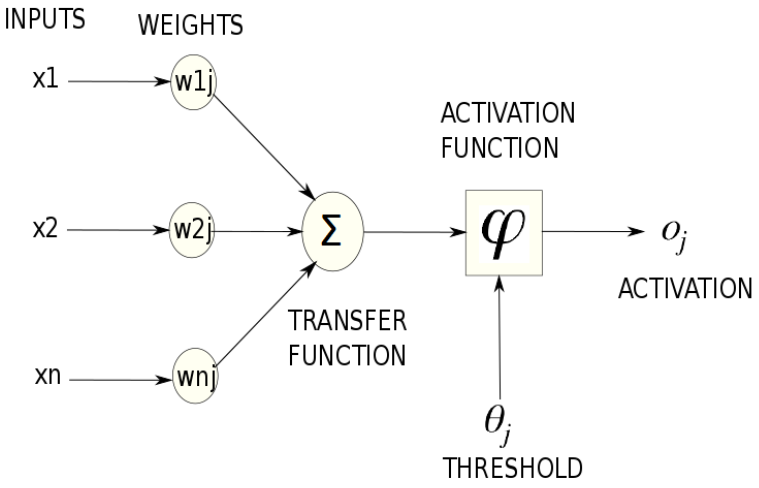


Figure 2: Simple Neural Network diagram.

2.1 Deep Learning

Since the late 2000's, neural networks have recovered importance and have become more successful thanks to the availability of inexpensive, parallel hardware (graphics processors, computer clusters) and a massive amount of labeled data. There are also new algorithms that make use of unlabeled data and achieve impressive improvements in various settings, but it can be argued that the core is almost the same with old architectures from the 1990's.

In the past few years, Deep Learning has generated much excitement in Machine Learning and industry thanks to many breakthrough results in speech recognition, computer vision and text processing. So, what is Deep Learning?

For many researchers, Deep Learning is another name for a set of algorithms that use a neural network as an architecture, but in this neural network there are more than 3 hidden

layers (see Figure 3). Due to the improvements in new activation functions and powerful Graphical Processing Units (GPUs) it is now possible to train the networks and use them for Image Classification [15].

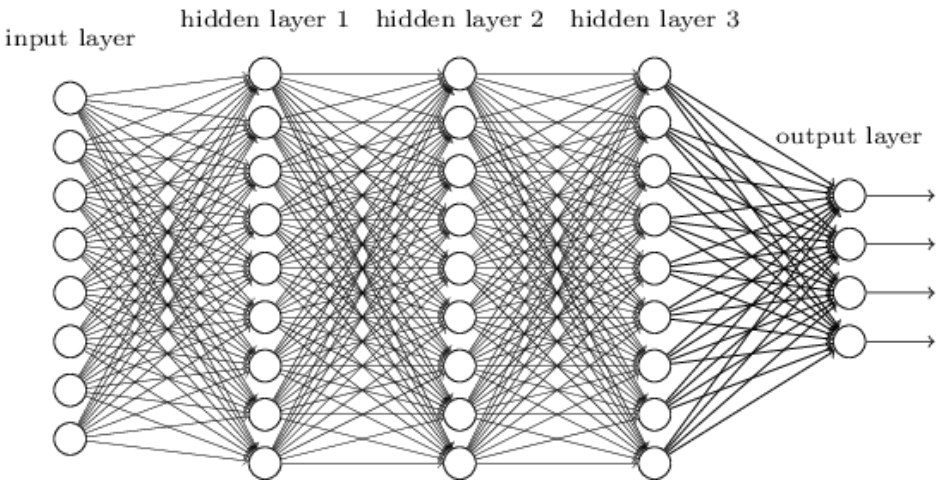


Figure 3: Deep Neural Network diagram (figure from neural networks and deep learning).

2.2 Convolutional Neural Networks (CNN)

Since 2012, one of the most important results in Deep Learning is the use of convolutional neural networks to obtain a remarkable improvement in object recognition.

In all networks that we have seen so far, every neuron in the first hidden layer connects to all the neurons in the input layer (Figure 4).

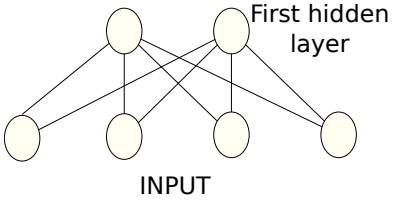


Figure 4: First hidden layer of a regular neural network (fully connected).

This does not work when the input is high-dimensional because every neuron ends up with many connections. For example, when the input is a small image of 100×100 pixels (i.e. the input vector has 10,000 dimensions), every neuron has 10,000 parameters. To make this more efficient, we can force each neuron to have a small number of connections to the input. The connection patterns can be designed to fit some structure in the inputs. For

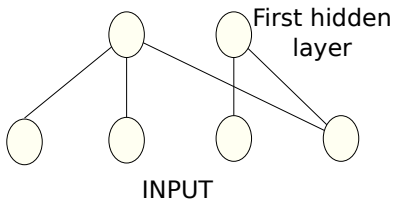


Figure 5: Non fully connected first hidden layer of a neural network.

example, in the case of images, the connection patterns involve that neurons can only look at adjacent pixels in the input image (Figure 5).

We can extend this idea to force local connectivity in many layers, to obtain a deep locally connected network. Training with gradient descent is possible because we can modify the backpropagation algorithm to deal with local connectivity: in the forward pass, we can compute the values of neurons by assuming that the empty connections have weights of zeros; whereas in the backward pass, we do not need to compute the gradients for the empty connections.

This kind of networks have many names: local networks, locally connected networks, local receptive field networks. The last name is inspired by the fact that neurons in the brain are also mostly locally connected, and the corresponding terminology in neuroscience/biology is "local receptive field".

Using locality structures significantly reduces the number of connections. Even further reduction can be achieved via another technique called "weight sharing". In weight sharing, some of the parameters in the model are constrained to be equal to each other. For example, in the following layer of a network we can see that $w1 = w6$, $w2 = w4$, $w3 = w5$ (Figure 6). With these constraints, the model can be quite compact in terms of number of actual parameters. Instead of storing all the weights we only store a few.

This idea of sharing the weights resembles an important operation in signal processing known as convolution. In convolution, we can apply a "filter" (a set of weights) to many positions in the input signals. In practice, this type of networks also come with another layer known as the "max-pooling layer". The max-pooling layer computes the maximum value of a selected set of output neurons from the convolutional layer and uses these as inputs to higher layers.

This type of networks is also known as convolutional neural networks (sometimes called *convnets*). The max-pooling layer in the CNNs is also known as the subsampling layer because it considerably reduces the size of the input data.

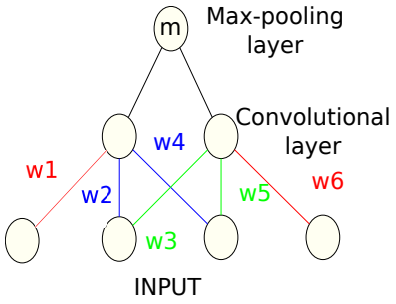


Figure 6: Weight sharing example in a CNN.

We also have to mention the rectifier, which in the context of artificial neural networks is an activation function defined as $f(x) = max(0,x)$, where x is the input to a neuron.This activation function has been argued to be more biologically plausible than the widely used logistic sigmoid (which is inspired by probability theory). The rectifier is, as of 2015, the most popular activation function for deep neural networks. A unit employing the rectifier is also called a rectified linear unit (ReLU).

3 Face alignment

Face alignment is one of the most important stages in image-based age estimation. In our experiments, the eyes of each face are detected using the Ensemble of Regression Trees (ERT) algorithm [16] which is a robust and very efficient algorithm for facial landmarks localization. Once we have the 2D positions of the two eyes, we use them to compensate for the in-plane rotation of the face. To this end, within the detected face region, the positions of right and left eyes are located as (R_x, R_y) and (L_x, L_y), respectively. Then, the angle of in-plane rotation is calculated by $\theta = artan(\frac{R_y - L_y}{R_x - L_x})$, and the input face region is rotated by the that angle.

Figure 7: Face alignment and cropping associated with one original image in MORPH II database.

After rotation correction, we use a global scale for the face image, this scale normalizes the inter-ocular distance to a fixed value l. After performing the rotation and rescaling, the face region should be cropped (aligned face). To this end, a bounding box is centered on the new eyes location (on the transformed face image) and then stretched to the left and to the right by $k_0 \cdot l$, and to top by $k_1 \cdot l$ and to bottom by $k_2 \cdot l$. Finally, in our case, k_0, k_1, k_2 and l are chosen such that the final face image has a size of 50×50 pixels for the MORPH II database (see Figure 7) and 200×200 for the PAL database.

4 Face features

This section will briefly describe some features that are very often used for extracting face features. We present three hand-crafted features as well as five deep features that can be obtained from pre-trained deep CNNs.

4.1 Local Binary Patterns (LBP)

The original LBP operator labels the pixels of an image with decimal numbers, which are called LBPs or LBP codes that encode the local structure around each pixel [17, 9]. The basic operator proceeds as follows. Each pixel is compared with its eight neighbors in a neighborhood by subtracting the central pixel value; the resulting strictly negative values are encoded with 0, and the others with 1. For each given pixel, a binary number is obtained by concatenating all these binary values in a clockwise direction, which starts from the one of its top-left neighbor. The corresponding decimal value of the generated binary number is then used for labeling the given pixel (see Figure 8). The histogram of LBP labels (the frequency of occurrence of each code) calculated over a region or an image can be used as a texture descriptor. It should be noticed the LBP descriptors can be either an LBP image or a histogram of that image.

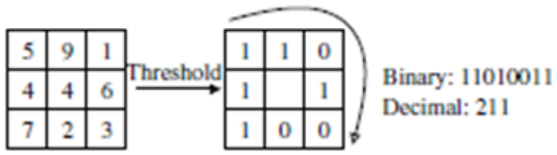


Figure 8: Basic LBP operator.

The basic LBP operator was extended to use neighborhoods of different sizes (different radio and different number of pixels). Using circular neighborhoods and interpolating the pixel values, any neighborhood of radius R and points P can be achieved.

In our work, we used the classic LBP operator that provides a histogram of 256 bins for a given face image.

4.2 Histogram of Oriented Gradients (HOG)

HOG is a feature descriptor used in computer vision and image processing for the purpose of object detection.

The essential thought behind the histogram of oriented gradients descriptor is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled. The descriptor is then the concatenation of these histograms. For improved accuracy, the local histograms can be contrast-normalized by calculating a measure of the intensity across a larger region of the image, called a block, and then using this value to normalize all cells within the block. This normalization results in better invariance to changes in illumination and shadowing. The HOG descriptor [10] has a few key advantages over other descriptors. Since it operates on local cells, it is invariant to geometric and photometric transformations, except for object orientation.

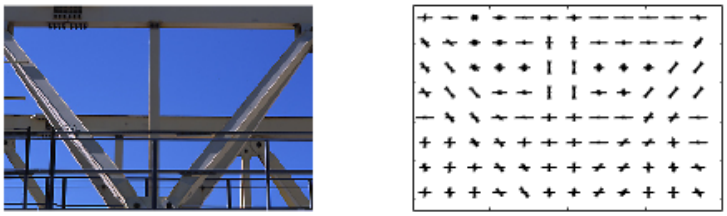


Figure 9: Original image and its HOG features (figure from mathworks).

4.3 Binarized Statistical Image Features (BSIF)

This descriptor [11] can be used in texture recognition tasks in a similar manner as LBPs. Each element (i.e. bit) in the binary code string is computed by binarizing the response of a linear filter with a threshold at zero. Each bit is associated with a different filter and the desired length of the bit string is determined by the number of filters used. The set of filters is learnt from a training set of natural image patches by maximizing the statistical independence of the filter responses. Hence, statistical properties of natural image patches determine the BSIF descriptors.

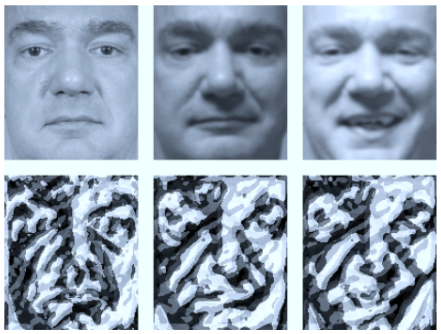


Figure 10: 3 face images and their corresponding BSIF codes (figure from [11]).

4.4 ImageNet VGG-F features

The Fast (VGG-F) architecture [18] is similar to the one used by Krizhevsky *et al.* [19]. It comprises 8 learnable layers, 5 of which are convolutional, and the last 3 are fully-connected. The input image size is 224×224. Fast processing is ensured by the 4 pixel stride in the first convolutional layer. The main differences between this architecture and that of Krizhevsky are the reduced number of convolutional layers and the dense connectivity between convolutional layers (Krizhevsky used sparse connections to enable training on two GPUs). The network was trained on ILSVRC-2012 using gradient descent with momentum. The hyper-parameters are the same as used by Krizhevsky. The authors applied data augmentation in the form of random crops, horizontal flips, and RGB color jittering. We extracted the deep features from the 4K dimensional feature vector after removing the last classification layer. The resulting vector is L2 normalized. The only image pre-processing consists on resizing the input images to the network input size and subtracting the average image (provided by the authors in the network metadata).

Arch.	conv1	conv2	conv3	conv4	conv5	full6	full7	full8
CNN-F	64x11x11 st. 4, pad 0 LRN, x2 pool	256x5x5 st. 1, pad 2 LRN, x2 pool	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 x2 pool	4096 drop-out	4096 drop-out	1000 soft-max

Figure 11: ImageNet VGG-F architecture (figure from [18]).

4.5 ImageNet VGG-verydeep-16 features

This network is part of the evaluation of networks of increasing depth carried out by Simonyan and Zisserman [20] that proved to be very performant at the ImageNet 2014 challenge. The configuration is quite different from the ones used in the top-performing

entries of the 2012 and 2013 competitions. Rather than using relatively large receptive fields in the first convolutional layers, they used very small 3×3 receptive fields throughout the whole net, which are convolved with the input at every pixel. More specifically, the convolution stride is fixed to 1 pixel; the spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution, i.e. the padding is 1 pixel for 3×3 convolutional layers. Spatial pooling is carried out by 5 max-pooling layers, which follow some of the convolutional layers (not all the convolutional layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2. In total, the network we used has 13 convolutional layers and 3 FC. The only preprocessing we do is subtracting the mean RGB value of the input image. The 4K features are collected from the last FC layer and L2 normalized.

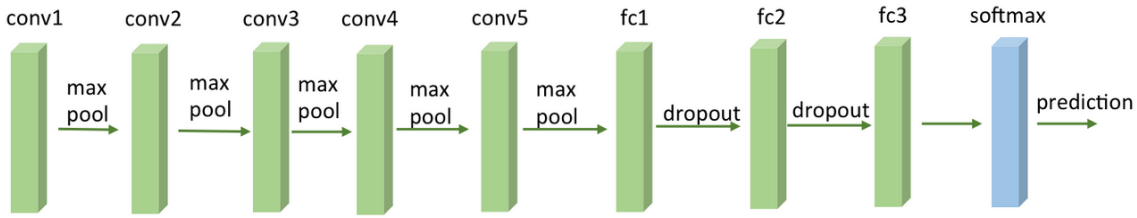


Figure 12: ImageNet-VGG-16-layer network structure. Each conv includes three convolutional layers (figure from Homepage of Zhicheng Yan).

4.6 VGG Face features

The CNN architecture comprises 11 blocks (see Figure 13), each containing a linear operator followed by one or more non-linearities such as ReLU and max pooling [21]. The first eight such blocks are said to be convolutional as the linear operator is a bank of linear filters (linear convolution). The last three blocks are instead called Fully Connected (FC); they are the same as a convolutional layer, but the size of the filters matches the size of the input data, such that each filter "senses" data from the entire image. All the convolution layers are followed by a rectification layer (ReLU). The first two FC layers output are 4,096 dimensional vectors. This multi-way CNN is trained to discriminate between the 2,622 identities using about 2.6M images. The deep features of this network are extracted by taking the 4K dimensional features and removing the last classification layer. The resulting vector is L2 normalized.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

Figure 13: Network configuration. Details of the VGG-Face network architecture. The FC layers are listed as "convolution" as they are a special case of convolution. For each convolution layer, the filter size, number of filters, stride and padding are indicated (figure from [21]).

4.7 DEX-IMDB-WIKI and DEX-ChaLearn-ICCV2015 features

The Deep EXpectation (DEX) on apparent age method [22, 23] uses the VGG-16 architecture for its networks (see Figure 14), which are pre-trained on ImageNet for image classification. In addition, the authors explored the benefit of fine-tuning over crawled Internet face images with available age. In total, they collected more than 500,000 images of celebrities from IMDb and Wikipedia. The networks of DEX were fine-tuned on the crawled images and then on the provided images with apparent age annotations from the ChaLearn LAP 2015 challenge on apparent age estimation. We extracted the features provided by two networks: DEX-IMDB-WIKI and DEX-ChaLearn-ICCV2015[1]. The first one was trained on real age estimation using the cropped and aligned faces of the IMDB-WIKI dataset, while the second one is a fine-tuned version of the previous model, trained on apparent age using the challenge images. An ensemble of these models led to 1st place at the challenge (115 teams). The 4K features are collected from the previous to the last FC layer.

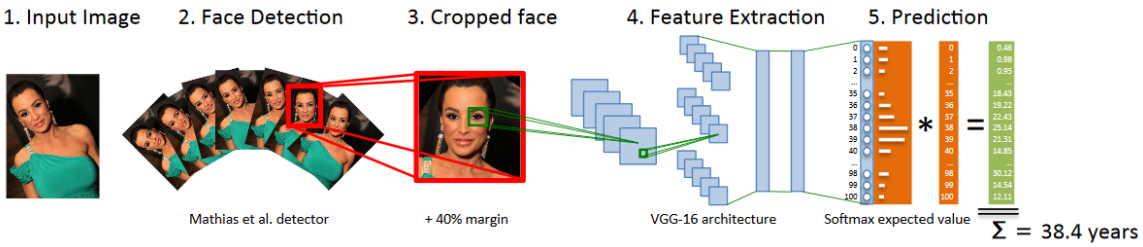


Figure 14: Pipeline of DEX method (with one CNN) for apparent age estimation (figure from [22]).

[1]For simplicity, we will refer to DEX-ChaLearn-ICCV2015 as DEX-CHALEARN in many parts of this thesis.

5 Experimental setup

Our study concerns three hand-crafted features and five deep features. The deep features were obtained by pre-trained CNNs. Two CNNs were trained on images of objects for the purpose of image categorization (ImageNet VGG-F and ImageNet VGG-verydeep-16). One net was trained on face images for the purpose of face identification (VGG-Face). The last two nets (DEX-IMDB-WIKI and DEX-ChaLearn-ICCV2015) were trained on face images for the purpose of age estimation. One can also notice that the first one was trained on real ages and the second one was trained on apparent age. With regards to feature sizes, the LBP, HOG and BSIF descriptors have 256, 832/1872/4212[2], and 256, respectively. All deep features are given by 4096 elements.

We also use Fisher Score [24] in order to reduce the dimension of each descriptor with the objective of reducing the computational time and getting almost the same prediction results.

5.1 Datasets

In our study, two public datasets are used:

5.1.1 MORPH (Album 2)

The MORPH (Album 2), or simply MORPH II, database from the University of North Carolina Wilmington [25] contains $\sim$55,000 unique images of 13,618 individuals (11,459 male and 2,159 female) in the age range of 16 to 77 years old. The distribution of ages in the dataset is shown in Figure 16.

The average number of images per individual is 4. The MORPH (Album 2) database can be divided into three main ethnicities: African (42,589 images), European (10,559 images) and other ethnicities (1,986 images). Some samples are illustrated in Figure 15a.

MORPH II is a database of longitudinal sides developed for researchers investigating all aspects of the progression of age, for example, face modeling, photorealistic animation, facial recognition, etc. This database contributes to several areas of active research, particularly the recognition by providing: the largest set of longitudinal images available to the

[2] for original and aligned MORPH II images and PAL images respectively.

(a)

(b)

Figure 15: Database images. **(a)** Sample images from MORPH II database. **(b)** Sample images from PAL database.

public, longitudinal sections of a few months to more than twenty years, and the inclusion of key physical parameters affecting appearance with aging.

As a first approximation to the problem, we evaluate the sample of white women from MORPH II using a 5-fold cross-validation. We select this subset of images because, as we said before, MORPH II contains $\sim 55,000$ images and selecting only the white women images we get $\sim 2,490$ and we can decrease the computation time in order to do the first experiments.

Later, we also use a 5-fold cross-validation evaluation for the whole MORPH II dataset.

The folds are selected in such a way to prevent algorithms from learning the identity of the persons in the training set by making sure that all images of individual subjects are only in one fold at a time.

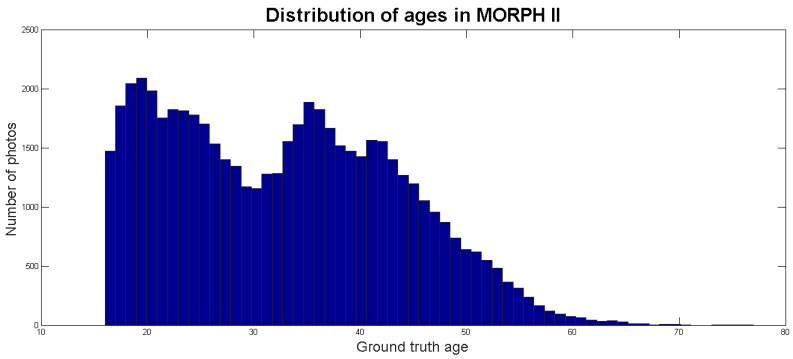


Figure 16: Distribution of ages in MORPH II.

5.1.2 PAL

The Productive Aging Lab Face (PAL) database from the University of Texas at Dallas [26] contains 1,046 frontal face images from different subjects (430 males and 616 females) in the age range of 18 to 93 years old. The PAL database can be divided into three main ethnicities: African-American subjects (208 images), Caucasian subjects (732 images) and other subjects (106 images). The database images contain faces having different expressions. Some samples are illustrated in Figure 15b.

For the evaluation of the approach, we conduct again a 5-fold cross-validation. In the experiments, we consider three cases: (i) original images, (ii) aligned images with loose crop (face plus some background), and (iii) aligned/cropped images. These cases are illustrated in Figure 17. The corresponding sizes are 230×350 pixels, 200×200 pixels, and 200×200 pixels, respectively.

Figure 17: Three types of PAL images. The left one is the original face image. The middle and right images correspond to the aligned and cropped face. The middle image correspond to a loose face cropping and the left one to a tight face cropping.

5.2 Evaluation protocol

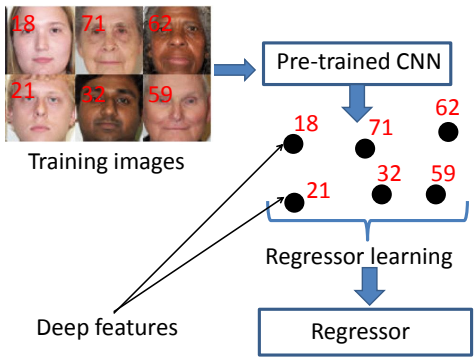
Figure 18 illustrates the training and testing processes used for evaluating the performances of the eight face features. The procedure is the same whether the features are hand crafted or provided by the pre-trained CNNs. We used five-fold cross-validation that allows to test every test image in the considered database. In our experiments, we used the Partial Least Square (PLS) regressor [27]. This is a statistical method that retrieves relations between groups of observed variables X and Y through the use of latent variables. It is a powerful statistical tool which can simultaneously perform dimensionality reduction and classification/regression. It estimates new predictor variables, known as components, as linear combinations of the original variables, with consideration of the observed output values.

It is worthy to notice that, for deep features, the training phase concerns only the regressor. We use two measures that are very common in the literature for evaluating the performance of automatic age estimators. The first measure is the Mean Age Error (MAE) (expressed in years) which is given by the average of absolute age error between the ground-truth ages and the predicted ones. The second measure is given by the Cumulative Score (CS). The Cumulative Score reflects the percentage of tested images for which the age estimation error is less than a threshold.

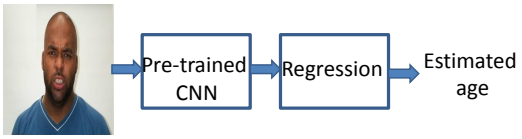
5.3 Results

5.3.1 Results on white women subset

As explained before, we started our experiments on a small subset of samples of MORPH II. Table 1 shows the MAE obtained on the white women subset from the MORPH II database using different percentages of the descriptors based on Fisher Score. We can see that using the 50% or the 80% of the descriptors the MAE does not increase very much, so in case we want to reduce the computational time it is not a bad decision to use the 50% of the descriptors since using the 100% of the descriptors we only get some decimals of improvement.



(a) Training.



(b) Testing.

Figure 18: **(a)** In the training phase, the training images are fed to a given pre-trained CNN in order to obtain the deep features. The deep features are used for learning an age regressor. **(b)** For a test face, the age is estimated by using the learned regressor using the corresponding deep features.

5.3.2 Results on whole MORPH II

Table 2 illustrates the MAE obtained on the MORPH II database using the eight face features. In this table, we considered two cases: the original images and the aligned/cropped images. We can observe that with face alignment and cropping the performances obtained with the hand-crafted features have increased. This is very intuitive since the hand-crafted features need to focus on the face region only. On the other hand, for the last two deep features, the use of the original images provided better performance. This can be explained by the fact that these ones were trained on face images having significant background.

Whether the original images or the aligned and cropped images were used, the deep features supplied by DEX-IMDB-WIKI and DEX-ChaLearn-ICCV2015 nets provided the best performances. Moreover, we can observe that among deep features the best perfor-

Face features	FS 20%	FS 50%	FS 80%	FS 100%
LBP	7.08	6.93	6.92	6.80
HOG	9.00	6.76	6.04	5.86
BSIF	6.83	6.45	6.20	5.98
IMAGENET-VGG-F	**5.64**	**5.82**	**5.83**	5.85
VGG-FACE	5.88	6.17	**5.83**	**5.72**

Table 1: Mean Age Error (years) obtained with different face features on the white women sample (2490 images) from MORPH II database and using different percentages of the descriptors based on Fisher Score (FS).

mances were obtained with nets that were trained on face images, i.e. VGG-Face, DEX-IMDB-WIKI and DEX-CHALEARN.

Face features	Original images	Aligned+cropped
LBP	7.20	6.53
HOG	6.26	4.84
BSIF	7.34	6.69
IMAGENET-VGG-F	5.11	5.04
IMAGENET-VERY-DEEP-16	5.53	5.47
VGG-FACE	4.72	4.79
DEX-CHALEARN	**3.67**	4.77
DEX-IMDB-WIKI	3.77	**4.76**

Table 2: Mean Age Error (years) obtained with different face features on MORPH II database.

The performances of some state-of-the-art approaches are shown in Table 3. As can be seen, our deep feature results are comparable to the performance obtained by the work of Han *et al.* (2015) [14]. The latter uses coarse-to-fine and hierarchical age estimation via binary decision trees for classifying non-overlapping age groups and within-group age regressors. In our case, only one single regressor is used.

Figure 19 represents the cumulative score associated with the eight face features using the aligned and cropped versions of the images. As can be seen, for some face features the cumulative scores are similar and even some of the handcrafted features present results close to the results of the deep features.

Having a look at the distribution of ages in MORPH II in Figure 16, we can appreciate that there are many more images of young people than those of elder people. That is the reason that can explain why as the age increases the MAE increases too. We can see it in Figure 20.

Publication	Approach	MAE (years)
Guo and Mu (2011) [28]	BIF* +KPLS†	4.2
Chang *et al.* (2011) [29]	BIF*	6.1
Geng *et al.* (2013) [30]	BIF*	4.8
Guo and Mu (2013) [31]	BIF*	4.0
Huerta *et al.* [3]	CNN‡	3.9
Han *et al.* (2015) [14]	DIF††	**3.6**
Our result	DEX-CHALEARN	3.67

* Biologically Inspired Features † Kernel Partial Least Squares
‡ Convolutional Neural Networks
†† Demographic Informative Features

Table 3: MAE (years) obtained with different state-of-the-art approaches on MORPH II database.

Another interesting way of analyzing the age prediction results consists on displaying the predicted age vs the real age of all samples. In this manner, we can observe if the error of our regressor is homogeneous among age ranges. In Figure 21, we show such a graph for our best predictions (obtained using the DEX-CHALEARN features). We can observe how, in general, the predicted ages for young people are lower than their real ages while the predicted ones for elder people are higher than their real ones.

5.3.3 Results on PAL

Tables 4, 5 and 6 show the MAE obtained on the PAL database with original images, with alignment and loose cropping (face plus some background) and with alignment and cropping respectively; using different percentages of the descriptors based on Fisher Score. As we can see, we are in the same situation as we were with MORPH II database. We can see that using the 50% or the 80% of the descriptors the MAE does not increase very much, so in case we want to reduce the computational time it is not a bad decision to use the 50% of the descriptors.

Table 7 illustrates the MAE obtained on the PAL database using the eight face features. In this table, we considered three cases: (i) original images, aligned images with loose crop

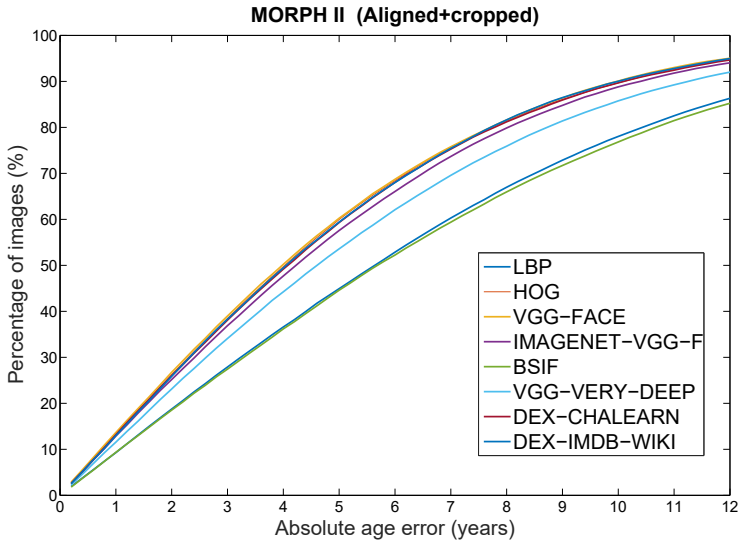


Figure 19: Cumulative scores obtained with eight face features for MORPH II database (aligned and cropped images).

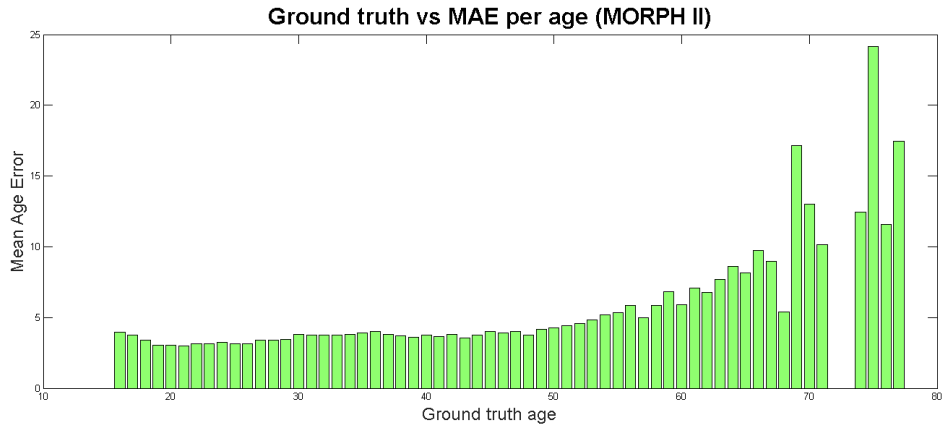


Figure 20: MAE for each age in MORPH II (using DEX-CHALEARN features).

Face features	FS 20%	FS 50%	FS 80%	FS 100%
LBP	13.30	11.87	11.72	11.39
HOG	14.26	10.78	8.68	8.68
BSIF	13.71	12.07	11.86	10.70
IMAGENET-VGG-F	7.87	6.98	6.91	6.88
IMAGENET-VERY-DEEP-16	8.98	8.22	8.06	8.04
VGG-FACE	9.10	6.78	6.19	5.90
DEX-CHALEARN	4.30	**4.02**	**3.97**	**3.97**
DEX-IMDB-WIKI	**4.25**	4.12	4.04	4.03

Table 4: Mean Age Error (years) obtained with different face features on PAL database and using different percentages of the descriptors based on Fisher Score (FS).

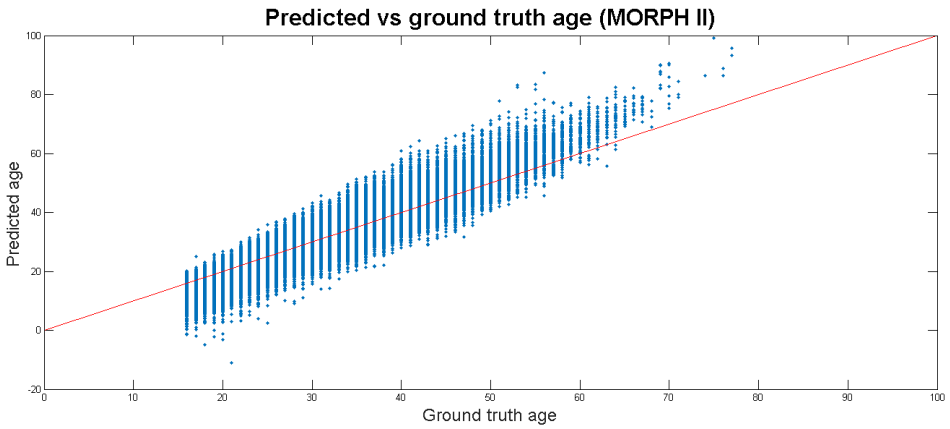


Figure 21: Predicted age vs real age in MORPH II (using DEX-CHALEARN features).

Face features	**FS 20%**	**FS 50%**	**FS 80%**	**FS 100%**
LBP	12.26	11.67	11.12	11.16
HOG	11.53	8.80	7.89	7.61
BSIF	13.02	12.04	11.42	11.25
IMAGENET-VGG-F	7.70	6.93	6.88	6.80
IMAGENET-VERY-DEEP-16	13.02	12.04	11.42	11.25
VGG-FACE	8.01	5.79	5.34	5.13
DEX-CHALEARN	**4.16**	**3.91**	**3.79**	**3.78**
DEX-IMDB-WIKI	4.25	3.92	3.80	3.79

Table 5: Mean Age Error (years) obtained with different face features on PAL database with alignment and loose crop and using different percentages of the descriptors based on Fisher Score (FS).

Face features	**FS 20%**	**FS 50%**	**FS 80%**	**FS 100%**
LBP	11.42	11.18	11.02	10.99
HOG	11.12	9.22	7.41	7.00
BSIF	10.76	10.23	10.18	10.08
IMAGENET-VGG-F	8.02	7.19	7.18	7.13
IMAGENET-VERY-DEEP-16	8.51	8.53	8.43	8.41
VGG-FACE	8.05	6.11	5.41	5.22
DEX-CHALEARN	6.01	5.32	5.15	5.12
DEX-IMDB-WIKI	**5.68**	**5.07**	**4.90**	**4.90**

Table 6: Mean Age Error (years) obtained with different face features on PAL database with alignment and crop and using different percentages of the descriptors based on Fisher Score (FS).

(face plus some background), and (iii) aligned/cropped images. We can observe that with face alignment and cropping the performances obtained with the hand-crafted features have increased. The use of original images and a loose crop have improved the performances obtained by the last two deep features. Whether the original images or the aligned and cropped images were used, those two deep features provided the best performances.

Face features	**Original images**	**Aligned+Loose crop**	**Aligned+crop**
LBP	11.40	11.16	10.99
HOG	8.68	7.61	7.00
BSIF	10.71	11.26	10.09
IMAGENET-VGG-F	6.89	6.81	7.14
IMAGENET-VERY-DEEP-16	8.04	8.64	8.41
VGG-FACE	5.91	5.13	5.23
DEX-CHALEARN	**3.97**	**3.79**	5.12
DEX-IMDB-WIKI	4.04	**3.79**	**4.90**

Table 7: Mean Age Error (years) obtained with different face features on PAL database.

The predictions are in general quite accurate for the DEX-CHALEARN features. In Figure 22 we show some examples of good and bad predictions calculated with our regressor using those features.

The performances of some state-of-the-art approaches are shown in Table 8. As can be seen, by adopting the proposed scheme, we got a significant improvement in performance. The best state-of-the art MAE was 5.4 years, whereas the best MAE obtained by our adopted scheme was 3.79 years.

Figure 23 represents the cumulative score associated with the eight face features using the aligned and cropped versions of the PAL images. As can be seen, the cumulative scores of the three deep features trained on facial images are very similar and outperform the rest of features.

Table 9 illustrates a comparison between the MAE of the end-to-end CNNs and that obtained by the use of deep features. The table corresponds to the MORPH II database with two different types of images. For each CNN, the upper row illustrates the MAE obtained by applying the net in order to estimate the age. The lower row depicts the MAE where the net is used to provide only the deep features. As can be seen, by adopting the deep features the obtained MAE was better than that of the end-to-end CNN. Table 10 illustrates a comparison between the MAE of the end-to-end CNNs and that obtained by the use of

Figure 22: Examples of good and bad predictions on PAL database (aligned + loose crop) using DEX-CHALEARN features.

Publication	Approach	MAE
Gunay and Nabiyev (2016) [32]	AAM+GABOR+LBP	5.4
Nguyen *et al.* (2014) [33]	MLBP+GABOR+SVR	6.5
Bekhouche *et al.* (2014) [34]	LBP+BSIF+SVR	6.2
Choi *et al.* (2010) [35]	GHPF*+SVR	8.4
Luu *et al.* (2011) [2]	CAM†+SVR	6.0
Our result	DEX-CHALEARN	**3.79**

* Gaussian High Pass Filter † Contourlet Appearance Model
‡ Multi-Quantized Local Binary Patterns

Table 8: Mean Age Error (years) obtained with different state-of-the art approaches on PAL database.

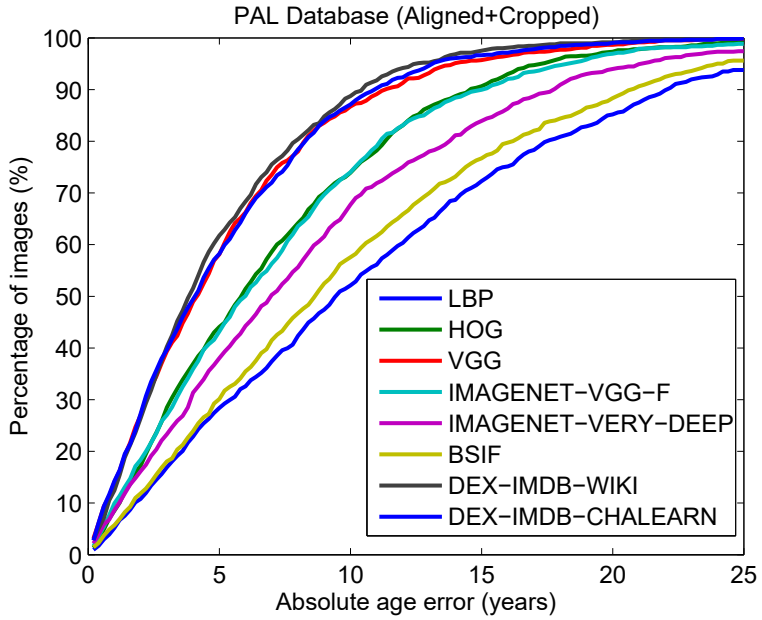


Figure 23: Cumulative scores obtained with eight face features for PAL database (aligned and cropped images).

deep features for PAL database. The table corresponds to the database with three different types of images. We can observe a similar behavior to that obtained with the MORPH II database. This tends to confirm that by only retraining the regressor, we are able to transfer the power of the pre-trained CNN without having to retrain the whole network.

CNN	Scheme	Original	Aligned+crop
DEX-CHALEARN	End-to-end	5.34	11.1
	Deep features	**3.67**	**4.77**
DEX-IMDB-WIKI	End-to-end	5.77	11.6
	Deep features	**3.77**	**4.76**

Table 9: Mean Age Error (in years) obtained with two deep CNNs on MORPH II database. For each CNN, the upper row illustrates the MAE obtained by applying the CNN as an end-to-end solution. The lower row depicts the MAE where the net is used to provide only the deep features.

CNN	Scheme	Original	Aligned+Loose crop	Aligned+crop
DEX-CHALEARN	End-to-end	7.12	5.43	8.53
	Deep features	**3.97**	**3.79**	**5.12**
DEX-IMDB-WIKI	End-to-end	6.99	4.72	7.98
	Deep features	**4.04**	**3.79**	**4.90**

Table 10: Mean Age Error (in years) obtained with two deep CNNs on PAL database. For each CNN, the upper row illustrates the MAE obtained by applying the CNN as an end-to-end solution. The lower row depicts the MAE where the net is used to provide only the deep features.

As we can see in Figure 24, it happens the same as with MORPH II database. The predicted ages in PAL for young people are lower than their real ages while the predicted ones for elder people are higher than their real ones.

5.3.4 Number of components for the PLS regressor

Table 11 illustrates the MAE as function of the number of latent component associated with the PLS regressor. We can observe that the use of 20 latent variables for almost all face features provided the best results.

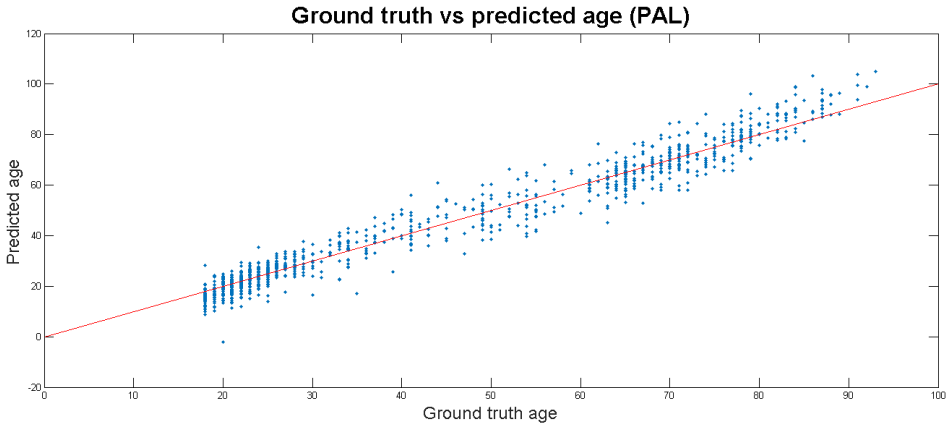


Figure 24: Predicted age vs real age in PAL (using DEX-CHALEARN features).

Features \# of PLS components	10	20	30	40
LBP	11.10	**10.84**	11.16	11.16
HOG	**7.02**	7.41	7.61	7.75
BSIF	11.33	**10.91**	11.26	11.34
IMAGENET-VGG-F	6.65	**6.45**	6.81	7.10
IMAGENET-VERY-DEEP-16	8.76	**8.54**	8.63	8.74
VGG-FACE	**5.04**	5.07	5.13	5.14
DEX-CHALEARN	3.79	**3.7**4	3.79	3.87
DEX-IMDB-WIKI	3.84	**3.79**	3.79	3.94

Table 11: MAE as a function of the latent variables used by the Partial Least Square regressor. The results correspond to PAL database.

6 Conclusions and Future Lines

This work has addressed the issue of comparing several face features for the task of age estimation from facial images. In the study, we have considered three hand-crafted image features as well as five deep features provides by pre-trained CNNs.

The comparison shown in this work yields several conclusions. First, the solution adopted in the work shows that efficient and stable age estimation can be obtained from deep features on the premise that the age regressor is retrained. The last process is by far more efficient than re-training the whole deep CNN on the new set of images. Second, the use of deep features gave better results than using hand-crafted features. Furthermore, we obtained the best results with nets that were trained on face images, i.e. VGG-Face, DEX-IMDB-WIKI and DEX-ChaLearn.

We also feel that we have made a significant improvement on PAL age estimation results. As we said before, the best state-of-the art MAE obtained for PAL was 5.4 years, while the best MAE obtained by our adopted scheme was 3.79 years.

In future work we will investigate feature fusion provided by deep CNNs and we will also use graph-based semi-supervised learning to see if we can get some improvements in the age prediction.

References

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

[2] Luu, K., Seshadri, K., Savvides, M., Bui, T.D., Suen, C.Y.: Contourlet appearance model for facial age estimation. In: Biometrics (IJCB), 2011 International Joint Conference on. (2011) 1–8

[3] Huerta, I., Fernández, C., Segura, C., Hernando, J., Prati, A.: A deep analysis on age estimation. Pattern Recognition Letters **68, Part 2** (2015) 239 – 249 Special Issue on "Soft Biometrics".

[4] Levi, G., Hassncer, T.: Age and gender classification using convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (2015) 34–42

[5] Ranjan, R., Zhou, S., Chen, J.C., Kumar, A., Alavi, A., Patel, V.M., Chellappa, R.: Unconstrained age estimation with deep convolutional neural networks. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). (2015) 351–359

[6] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 681—685

[7] Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 2234–2240

[8] Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: Computer Vision Pattern Recognition. (2009)

[9] Bereta, M., Karczmarek, P., Pedrycz, W., Reformat, M.: Local descriptors in application to the aging problem in face recognition. Pattern Recognition **46** (2013) 2634—-2646

[10] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2005)

[11] Kannala, J., Rahtu, E.: BSIF: Binarized statistical image features. In: Pattern Recognition (ICPR), 2012 21st International Conference on. (2012) 1363–1366

[12] Han, H., Jain, A.K.: Age, gender and race estimation from unconstrained face images. Technical Report MSU-CSE-14-5, Department of Computer Science, Michigan State University, East Lansing, Michigan (2014)

[13] Guo, G., Mu, G.: A framework for joint estimation of age, gender and ethnicity on a large database. Image and Vision Computing **32** (2014) 761 – 770 Best of Automatic Face and Gesture Recognition 2013.

[14] Han, H., Otto, C., Liu, X., Jain, A.K.: Demographic estimation from face images: Human vs. machine performance. IEEE Transactions on Pattern Analysis and Machine Intelligence **37** (2015) 1148–1161

[15] Le, Q.V., et al.: A tutorial on deep learning part 1: Nonlinear classifiers and the backpropagation algorithm (2015)

[16] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1867–1874

[17] Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28** (2006) 2037–2041

[18] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: British Machine Vision Conference. (2014)

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

[20] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[21] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. Volume 1. (2015) 6

[22] Rothe, R., Timofte, R., Gool, L.V.: DEX: Deep expectation of apparent age from a single image. In: IEEE International Conference on Computer Vision Workshops (ICCVW). (2015)

[23] Rothe, R., Timofte, R., Gool, L.V.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision (IJCV) (2016)

[24] Yang, M., Song, J.: A novel hypothesis-margin based approach for feature selection with side pairwise constraints. Neurocomputing **73** (2010) 2859–2872

[25] Ricanek, K., Tesafaye, T.: MORPH: a longitudinal image database of normal adult age-progression. In: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on. (2006) 341–345

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

[27] Rosipal, R., Kramer, N.: Overview and recent advances in partial least squares. In: Subspace, Latent Structure and Feature Selection Techniques. Springer (2006) 34–51

[28] Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. (2011) 657–664

[29] Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. (2011) 585–592

[30] Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 2401–2412

[31] Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: CCA vs. PLS. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. (2013) 1–6

[32] Günay, A., Nabiyev, V.V.: Age Estimation Based on Hybrid Features of Facial Images. In: Information Sciences and Systems 2015: 30th International Symposium on

Computer and Information Sciences (ISCIS 2015). Springer International Publishing, Cham (2016) 295–304

[33] Nguyen, D.T., Cho, S.R., Shin, K.Y., Bang, J.W., Park, K.R.: Comparative study of human age estimation with or without pre-classification of gender and facial expression. The Scientific World Journal **2014** (2014) 15

[34] Bekhouche, S., Ouafi, A., Taleb-Ahmed, A., Hadid, A., Benlamoudi, A.: Facial age estimation using BSIF and LBP. In: Proceeding of the first International Conference on Electrical Engineering ICEEB'14. (2014)

[35] Sung Eun, C., Youn Joo, L., Sung Joo, L., Kang Ryoung, P., Jaihie, K.: A comparative study of local feature extraction for age estimation. In: Control Automation Robotics *Vision (ICARCV), 2010 11th International Conference on. (2010) 1280–1284*