



A Deep Analysis on Age Estimation

Ivan Huerta^{†a,**}, Carles Fernández^{†b}, Carlos Segura^b, Javier Hernando^c, Andrea Prati^a

^aDPDCE, University IUAV, Santa Croce 1957, 30135 Venice, Italy

^bHerta Security, Pau Claris 165 4-B, 08037 Barcelona, Spain

^cUniversitat Politècnica de Catalunya, Jordi Girona 1, 08034 Barcelona, Spain

ABSTRACT

The automatic estimation of age from face images is increasingly gaining attention, as it facilitates applications including advanced video surveillance, demographic statistics collection, customer profiling, or search optimization in large databases. Nevertheless, it becomes challenging to estimate age from uncontrollable environments, with insufficient and incomplete training data, dealing with strong person-specificity and high within-range variance. These difficulties have been recently addressed with complex and strongly hand-crafted descriptors, difficult to replicate and compare. This paper presents two novel approaches: first, a simple yet effective fusion of descriptors based on texture and local appearance; and second, a deep learning scheme for accurate age estimation. These methods have been evaluated under a diversity of settings, and the extensive experiments carried out on two large databases (MORPH and FRGC) demonstrate state-of-the-art results over previous work.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Estimating age from images has been historically one of the most challenging problems within the field of facial analysis. Some of the reasons are the uncontrollable nature of the aging process, the strong specificity to individual traits (Weng et al. (2013)), high variance of observations within the same age range, camouflage due to beards, moustache, glasses and makeup (this latter specifically used to alter the perceived age), and the difficulty to gather complete and sufficient training data (Geng et al. (2013)).

As in most image recognition tasks, a large and representative amount of data/images is required to successfully train the classifier. Moreover, in the case of supervised classifiers, data/images need to be annotated, with real age in our case. In the past, however, available databases were limited and strongly skewed. This is especially disadvantageous for video surveillance and forensics, where unknown subjects are common and

often not collaborative. Fortunately, the availability of large databases like MORPH (Ricanek and Tesafaye (2006)) and FRGC (Phillips et al. (2005)) offers opportunities to progress in the field. However, training data sets can never represent the whole population fully, and methods with substantial robustness need to be developed in order to exploit large databases.

The inherent difficulties in the facial age estimation problem, such as limited imagery, challenging subject variability, or subtle visual age patterns, have derived research in the field into building particularly complex feature extraction schemes. The most typical ones consist of either hand-tuned multi-level filter banks (Guo et al. (2009); Geng et al. (2013); Han et al. (2013)), that intend to emulate the behavior of primary visual cortex cells, or fine-grained facial meshes to accomplish precise alignment through dozens of facial landmarks (Chang et al. (2011); Geng et al. (2007); Lanitis et al. (2004)). In any case, the resulting extraction schemes are difficult to replicate, and the high-dimensional visual descriptors in many cases take considerable time to be computed.

This paper addresses these issues from a very practical perspective: given the above-mentioned limitations of the existing approaches, none of which can fully handle all the issues, we aim at proposing two possible orthogonal ways. The first one aims at simplifying the estimation process by avoiding hand-crafted features, while proposing a simple yet effective fusion

**Corresponding author: Tel.: +39-041-257-2169; fax: +39-041-257-2424;
e-mail: huertacasado@iuav.it (Ivan Huerta)

† I. Huerta and C. Fernández contributed equally.

of well-known descriptors. By carefully selecting the features to fuse we can ideally borrow the best from all of them. On the other hand, previously hand-crafted and complex schemes for extracting visual features are progressively being replaced by deep learning procedures, which automatically train layered network architectures to tackle a defined problem. To the best of our knowledge, this paper conducts the first thorough evaluation of a deep learning framework for estimating age from face images.

With these premises, the title of this paper contains a pun: the word "deep" has a twofold meaning, referring both to the thorough (deep) analysis of commonly used local visual descriptors and to the proposal of deep learning approaches, in order to investigate their utility towards the automatic facial age estimation problem. Based on the limitations of existing proposals to face age estimation, the main contributions are stated next:

1. We extensively review effective descriptors based on texture and appearance, and show that their fusion improves over complex, state-of-the-art feature extraction schemes. Even though no new descriptors are proposed, their comprehensive evaluation and the demonstration of the superior performance achievable by fusing some of these (orthogonal) features, represent interesting results for the scientific community.
2. We investigate learning schemes to automatically train deep neural networks for age estimation. As mentioned above, we first conduct thorough evaluation of deep learning for age estimation. Deep learning has been proposed in the past and proved to be a viable and effective classifier for several applications. However, its performance for age estimation was still questionable, given the high variability and limited data available.
3. The proposed methods are exhaustively evaluated over two large databases, regarding optimal parameters and regularization. Both methods showed state-of-the-art results, despite the use of a simple eye alignment as preprocessing.

The paper is structured as follows. Next section gathers previous work regarding facial age estimation. Section 3 reviews the proposed candidate descriptors, along with the chosen classification scheme, and comments on the investigated deep learning scheme. Evaluation for both methods is presented out in Section 4, first reviewing available age-annotated large databases, and then describing the experiments carried out over fused local descriptors and deep neural networks. Finally, Section 5 summarizes the results and draws some conclusions.

2. Related work

Initial attention on automatic age estimation from images dated back to the early 2000s (Lanitis et al. (2004, 2002); Minear and Park (2004)). However, research in the field has been experiencing a renewed interest from 2006 on, since the availability of large databases like MORPH-Album 2 (Ricanek and Tesafaye (2006)), which increased by 55× the amount of real age-annotated data compared to databases at that time.

This database has been consistently evaluated in recent works through different feature extraction and classification schemes.

Feature extraction scheme. In age estimation from images, typically the first phase after pre-processing is to extract visual features which need to be (1) discriminative among different classes, (2) robust within the same class, and (3) with a minimal dimensionality. One class of methods relies on flexible shape and appearance models such as ASM (Active Shape Model) and AAM (Active Appearance Model) to model aging patterns (Chang et al. (2011); Geng et al. (2013, 2007); Lanitis et al. (2004)). Such statistical models capture the main modes of variation in shape and intensity observed in a set of faces, and allow to encode face signatures based on such characterizations.

Other methods extract a set of visual features which are then fed into the classifier to estimate the age. For instance, Bio-Inspired Features (BIF) (Riesenhuber and Poggio (1999)) and its derivations have consistently been used for age estimation in the last years (Geng et al. (2013); Han et al. (2013)). These feed-forward models intertwine a number of convolutional and pooling layers. First, an input image is mapped to a higher-dimensional space by convolving it with a bank of multi-scale and multi-orientation Gabor filters. Later, a pooling step down-scales the results with a non-linear reduction, typically a MAX or STD operation, progressively encoding the results into a vector signature. In Guo et al. (2009), the authors carefully design a two-layer simplification of this model for age estimation by manually setting the number of bands and orientations for convolution and pooling. Such features are also used in their posterior works, e.g. Guo and Mu (2011, 2013, 2014).

Features extracted from local neighborhoods have been used for the purpose of age estimation, for example in Yang and Ai (2007), Gunay and Nabiyevev (2008) and Choi et al. (2011). In Weng et al. (2013), LBP histogram features are combined with principal components of BIF, shape and textural features of AAM, and PCA projection of the original image pixels. Independent HOG features have been used for age estimation in Fernández et al. (2014) and Huerta et al. (2014).

Classification scheme. With regards to the learning algorithm, several approaches have been proposed, including, among others, Support Vector Machines / Regressors (Guo et al. (2009); Han et al. (2013); Chang et al. (2011); Weng et al. (2013)), neural networks (Lanitis et al. (2004)) and their variant of Conditional Probability Neural Network (Geng et al. (2013)), Random Forests (Montillo and Ling (2009)), and projection techniques such as Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA), along with their regularized and kernelized versions (Guo and Mu (2011, 2013, 2014)). An extensive comparison of these classification schemes for age estimation has been reported (Fernández et al. (2014); Huerta et al. (2014)), and the advantageousness of CCA was demonstrated over others, both regarding accuracy and efficiency.

For this reason, specific attention must be given to the CCA technique. The PLS and CCA subspace learning algorithms were originally conceived to model the compatibility between two multidimensional variables. PLS uses latent variables to learn a new space in which such variables have maximum cor-

relation, whereas CCA finds basis vectors such that the projections of the two variables using these vectors are maximally correlated to each other. Both techniques have been adapted for label regression. To the best of our knowledge, the best current result over MORPH is achieved by combining BIF features with kernel CCA (Guo and Mu (2013)), although in that case the size of training folds is limited to 10K samples due to computational limitations.

Deep learning. Recently, convolutional networks and deep learning schemes have been successfully employed for many tasks related to facial analysis, including face detection, face alignment (Sun et al. (2013)), face verification (Taigman et al. (2014)), and demographic estimation (Yang et al. (2011)). This last work actually exploits age and gender cues in order to address face recognition, whereas we specifically focus on analyzing and evaluating convolutional network architectures for age estimation. The basic methodology is generally common to all, i.e., combining a number of convolutional, pooling and fully or partially connected neuron layers, with variations in the order, repetition and connectivity of the layers. Nonetheless, the particular choice of parameters, which are typically shared across layers, is the key to their success.

One of the main contributions of this paper is the proposal of a novel combination of well-known local descriptors capturing texture and contour cues for the purpose of facial age estimation. The different nature of these features allows the exploitation of the benefits of each of them, bringing to performance which are superior than in the case of them applied separately. Another contribution is the evaluation of deep learning frameworks to the problem of age estimation. In this field, to the best of our knowledge, approaches based on local features and deep learning have never been compared to each other under the same experimental settings, and across several databases. Our experiments demonstrate a comparable performance of both proposals with respect to state-of-the-art results provided by complex and fine-tuned feature extraction schemes such as BIF (Guo and Mu (2014)). Moreover, for the sake of simplicity and efficiency, a simple eye alignment operation is carried out through similarity transformation, as opposed to precise alignment approaches typically fitting active shape and appearance models with tens of facial landmarks.

3. Methodology

We present two approaches, one based on local features and their combination, and the other exploiting deep learning. Both methodologies employ the same basic preprocessing, described next. A global view of the methodology is presented in Figure 1.

Preprocessing. The facial region of each image has been detected with the face detector described in Oro et al. (2011). Differently from other methods which rely on tens of facial landmarks for accurate alignment (e.g., ASM and AAM), we exploit the relative alignment invariance of local descriptors based on concatenated cell histograms to work with simple

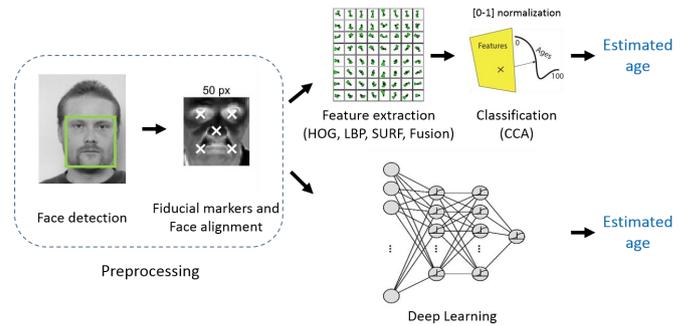


Fig. 1. General view of the two methodologies presented in this paper.

eye-aligned images. The fiducial markers corresponding to the eye centers have been obtained using the convolutional neural network for face alignment presented in Sun et al. (2013). The aligned version of each detected face is obtained by a non-reflective similarity image transformation that yields an optimal least-square correspondence between the eye centers and the target locations, that have been symmetrically placed at 25% and 75% of the alignment template. Unlike previous works like Guo and Mu (2013), which use input images of 60×60 pixels, our aligned images are resized to only 50×50.

Descriptors. The choice of visual features to be extracted from aligned images plays a fundamental role on the resulting estimation accuracy. In this paper, we have selected a number of significant local invariant descriptors that have been useful for image matching and object recognition in the past due to their expressiveness, fast computation, compactness, and invariance to misalignment and monotonic illumination changes. They include local appearance descriptors as HOG and texture descriptors as LBP and SURF.

Histograms of Oriented Gradients (HOG) (Dalal and Triggs (2005)) have largely been used as robust visual descriptors in many computer vision applications related to object detection and recognition. The image region is divided into $C_x \times C_y$ grid cells. A histogram of orientations is assigned to each cell, in which every bin accounts for an evenly split sector of either the $[0, \pi]$ or $[-\pi, \pi]$ domain (for unsigned and signed versions, respectively). At each pixel location, the gradient magnitude and orientation is computed, and that pixel increments the assigned orientation bin of its correspondent cell by its gradient magnitude. Cell histograms are concatenated to provide the final descriptor. We use $HOG_{C_x, C_y}^{x, S}$ to denote $C_x \times C_y$ square grids (where $C = C_x = C_y$) and B orientation bins, at S different scales.

Local Binary Patterns (LBP) (Ojala et al. (2002)) have been long used as a textural descriptor for image classification, and more recently, variations of the original proposal have provided state-of-the-art results in fields like face and object recognition. The original operator describes every pixel in the image by thresholding its surrounding 3×3-neighborhood with its intensity value, and concatenating the 8 boolean tests as a binary number. To build an LBP compact descriptor, a histogram is computed over the filtered result, in which each bin corresponds to a LBP code. Another typical extension reduces the dimensionality of the descriptor by assigning all *non-uniform* codes to

a single bin, whereas uniform codes are defined as those having not more than 2 bitwise transitions from 0 to 1 or vice versa (e.g., 00111000, versus non-uniform 01001101). An LBP descriptor of generic neighborhood size P and radius R using uniform patterns at S scales is referred as $LBP_{P,R}^{u2 \times S}$, e.g. $LBP_{8,2}^{u2 \times 1}$.

Speeded-Up Robust Features (SURF) (Bay et al. (2006)) is an interest point detector and descriptor that is particularly invariant to scale and rotation. It has commonly been used in image matching and object recognition as a faster and comparable alternative to SIFT. In our case, we concentrate on the upright version of the technique (U-SURF). The square image region to describe is partitioned into 4×4 subregions. Horizontal and vertical wavelet responses d_x and d_y are computed and weighted with a Gaussian. The sum of these responses and their absolute values are stored, generating a 4-dimensional vector $(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$ for each subregion, and these are concatenated to form the final 64-dimensional descriptor of the image region, SURF₆₄. A common extension consists of doubling the number of features, by separately computing the sums of d_x and $|d_x|$ for $d_y < 0$ and $d_y \geq 0$, and equally for d_y given the sign of d_x , thus yielding SURF₁₂₈. We will use the notation SURF_D^{F_S} to refer to the concatenation of D -dimensional SURF descriptors at S different scales.

As gradient information is used to describe image content by most descriptors, we have included raw magnitude gradient images $(\delta I := \sqrt{(\frac{\delta I}{\delta x})^2 + (\frac{\delta I}{\delta y})^2})$ as a baseline in our experiments for the evaluation of the proposed descriptors.

Classification. From the wide variety of learning schemes presented in the literature on facial age estimation, *Canonical Correlation Analysis (CCA)* and its derivations have recently obtained state-of-the-art results in challenging large databases such as MORPH (Guo and Mu (2014)). This projection technique involves low computational effort and unprecedented accuracy in the field, for which we use it as our chosen regression learning algorithm. CCA is posed as the problem of relating data \mathbf{X} to labels \mathbf{Y} by finding basis vectors w_x and w_y , such that the projections of the two variables on their respective basis vectors maximize the correlation coefficient

$$\rho = \frac{w_x^T \mathbf{X} \mathbf{Y}^T w_y}{\sqrt{(w_x^T \mathbf{X} \mathbf{X}^T w_x)(w_y^T \mathbf{Y} \mathbf{Y}^T w_y)}}, \quad (1)$$

or, equivalently, finding $\max_{w_x, w_y} w_x^T \mathbf{X} \mathbf{Y}^T w_y$ subject to the scaling $w_x^T \mathbf{X} \mathbf{X}^T w_x = 1$ and $w_y^T \mathbf{Y} \mathbf{Y}^T w_y = 1$. For age estimation, the data matrix \mathbf{X} is $M \times N$ and the label matrix \mathbf{Y} is $M \times 1$, being M the number of examples and N the dimension of the descriptor. Hence, since \mathbf{Y} becomes a vector, the vector w_y turns to be a simple scaling factor, so a least squares fitting suffices to relate labels \mathbf{Y} to the projected data features $w_x^T \mathbf{X}$. Thus, only w_x (of size $M \times 1$) needs to be computed, by solving the following generalized eigenvalue problem:

$$\mathbf{X} \mathbf{Y}^T (\mathbf{Y} \mathbf{Y}^T + \gamma_y I)^{-1} \mathbf{Y} \mathbf{X}^T w_x = \lambda (\mathbf{X} \mathbf{X}^T + I) w_x \quad (2)$$

When projecting through the solution w_x , the dimensionality of data features is reduced to one dimension per output (a single

numerical value in our case), so the aforementioned label fitting simply consists on finding the scalar value that optimally adapts the projected values to the ground truth age, in the least-squares sense. The described procedure can be stabilized through regularization, by modifying the eigenvalue problem as follows:

$$\mathbf{X} \mathbf{Y}^T ((1 - \gamma_y) \mathbf{Y} \mathbf{Y}^T + \gamma_y I)^{-1} \mathbf{Y} \mathbf{X}^T w_x = \lambda ((1 - \gamma_x) \mathbf{X} \mathbf{X}^T + \gamma_x I) w_x \quad (3)$$

Regularization terms $\gamma_x, \gamma_y \in [0, 1]$ have been included in Eq. 3 to prevent overfitting. Although CCA also admits extension to a kernelized version, kCCA, in that case covariance matrices become computationally intractable over 10K samples. In practice, regularized CCA (rCCA) works comparably to kCCA (Guo and Mu (2013)), it is much less computationally demanding, and will allow us to reproduce the same exact validation schemes than other algorithms over large databases.

Deep Learning. Neural network formulations have regained remarkable popularity in the computer vision and machine learning communities, in the form of deep learning schemes. This is explained by a number of reasons, namely the availability of larger datasets to be exploited automatically by these schemes, and the recent availability of more efficient hardware devoted to scalable computation.

Large datasets are crucial for generalizing computer vision solutions to non-constrained settings, due to the multiple sources of variability, e.g. view, illumination, or occlusion. Previously popular machine learning techniques such as support vector machines or subspace learning methods (PCA, LDA, ICA, CCA) become seriously limited when dealing with large training sets. For instance, we have mentioned that kCCA can work in practice up to 10K training samples (Guo and Mu (2013)), whereas large volumes of data are actually recommended or even required for conducting deep learning. Moreover, deep learning frameworks are especially useful when the problem involves the exploitation of non-trivial features, due to the fact that the feature extraction and classification steps are jointly optimized during the learning process. The resulting network internally extracts suitable features for minimizing an objective cost function, hence crafting adequate features for better tackling the problem.

Following the success of recent works on deep learning for facial analysis, we incorporate types of layers that are devoted to learn the appropriate features for the problem, followed by layers that serve for interrelating the information globally and conducting the regression or classification process. The first type typically includes convolutional and pooling layers, whereas the second type is represented by locally or fully connected neurons. For many problems, it is best to repeat the first group of layers a number of times, in order to extract features of progressively higher order, from edges and contours to blobs and textures. The particular choice of layer-specific parameters (e.g. filter sizes and number), as well as those related to the learning process itself (e.g. learning rates, weight regularization) is described in the following section.

Table 1. Description of popular databases for age estimation. Our evaluation considers those in bold.

Database	Reference	Samples	Subjects	Comments
PAL	Minear and Park (2004)	580	580	Limited number of samples
FG-NET	Lanitis et al. (2002)	1,002	82	Limited number of samples and subjects
GROUPS	Gallagher and Chen (2009)	28,231	28,231	Ages discretized into seven age intervals
FRGC v2.0	Phillips et al. (2005)	44,278	568	Large database; many samples per subjects
MORPH-II	Ricanek and Tesafaye (2006)	55,134	13,618	Large database; high diversity

4. Experimental Results

Age databases. Due to the nature of the age estimation problem, there is a restricted number of publicly available databases providing a substantial number of face images labelled with accurate age information. Table 1 shows the summary of the existing databases with main reference, number of samples, number of subjects, and comments.

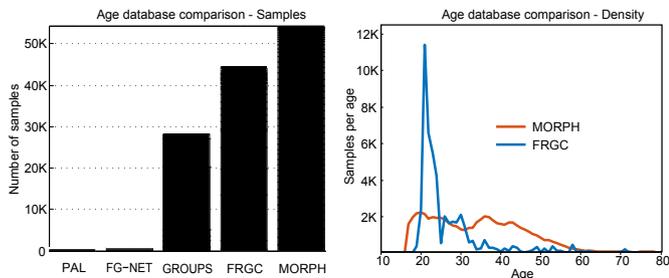


Fig. 2. Left: Number of face samples per database. Right: global density per age. PAL and FG-NET are relatively negligible, and GROUPS annotates only intervals. We focus on MORPH-II and FRGC. Samples are skewed towards 20–30 years old.

From Table 1, it is quite clear that older datasets like PAL and FG-NET are composed by a negligible number of samples when compared to the other newer datasets. GROUPS, instead, contains a good number of samples. However, age annotations are discretized into seven age intervals, which makes it unsuitable for training accurate age estimation models. Moreover, FG-NET contains only 82 subjects, so a *leave-one-person-out* validation scheme is employed by convention, to avoid optimistic biasing by identity replication. Given such limitations, and the recent tendency to use MORPH as a standard for age estimation, we concentrate on this database and on FRGC to provide experimental evaluations. Although the FRGC database is comparable to MORPH regarding number of samples, image quality and age range coverage, we have only found one previous publication on age estimation including FRGC as part of their experiments (Fernández et al. (2014)). Figure 2 offers a graphical visualization and comparison of the analyzed databases, by number of samples and age density. Figure 2 also shows the age distribution of the different datasets: it is evident that both MORPH and FRGC have samples with age mostly concentrated on the range 20-55.

Metrics. To evaluate the accuracy of the age estimators, the conventional metrics are the Mean Average Error (MAE) and the Cumulative Score (CS). MAE computes the average age

deviation error in absolute terms, $MAE = \sum_{i=1}^M |\hat{a}_i - a_i|/M$, with \hat{a}_i the estimated age of the i -th sample, a_i its real age and M the total of samples. CS is defined as the percentage of images for which the error e is no higher than a given number of years l , as $CS(l) = M_{e \leq l}/M$ (Chang et al. (2011); Weng et al. (2013); Han et al. (2013)). Related publications typically supply either an eleven-point curve for age deviations $[0 - 10]$, or simply the value $CS(5)$.

Through the rest of this paper, the optimal parameters are searched so as to minimize the MAE score over MORPH, using 5-fold cross-validation in all cases ¹. In particular, the division into training and validation sets is made so that all the instances of the same subject are contained in one single fold at a time; this applies to all the presented experiments. Descriptors are always extracted from the aligned version of detected faces.

Cx	Cy	B																		
		7	8	9	10	11	12	13	14	15	16	17	18	19	20					
3	3	7.11	6.97	6.88	6.86	6.73	6.77	6.66	6.58	6.60	6.56	6.55	6.48	6.48	6.49					
4	4	6.62	6.56	6.35	6.42	6.28	6.28	6.17	6.18	6.16	6.16	6.08	6.06	6.04	6.06					
5	5	5.76	5.75	5.55	5.53	5.47	5.44	5.39	5.37	5.38	5.35	5.33	5.31	5.31	5.29					
6	6	5.53	5.43	5.30	5.32	5.26	5.23	5.17	5.18	5.16	5.14	5.13	5.11	5.12	5.10					
7	7	5.13	5.10	4.98	4.99	4.93	4.93	4.89	4.89	4.88	4.85	4.85	4.85	4.85	4.84					
8	8	4.97	4.94	4.84	4.86	4.80	4.80	4.76	4.77	4.75	4.75	4.74	4.73	4.74	4.73					
9	9	4.85	4.81	4.73	4.75	4.71	4.69	4.66	4.67	4.64	4.64	4.65	4.64	4.63	4.64					
10	10	4.77	4.73	4.68	4.69	4.64	4.61	4.61	4.60	4.59	4.58	4.59	4.59	4.59	4.59					
11	11	4.66	4.62	4.55	4.57	4.54	4.50	4.50	4.50	4.49	4.47	4.50	4.49	4.49	4.50					
12	12	4.84	4.82	4.77	4.78	4.72	4.71	4.70	4.72	4.70	4.69	4.70	4.70	4.71	4.71					
13	13	4.66	4.65	4.60	4.61	4.57	4.56	4.54	4.56	4.55	4.54	4.55	4.56	4.56	4.57					
14	14	4.57	4.55	4.50	4.51	4.47	4.46	4.45	4.48	4.46	4.46	4.46	4.47	4.47	4.48	4.49				
15	15	4.47	4.45	4.41	4.42	4.39	4.38	4.38	4.40	4.39	4.39	4.40	4.41	4.41	4.43					
16	16	5.14	5.09	5.08	5.10	5.04	5.06	5.05	5.06	5.07	5.04	5.09	5.10	5.11	5.11					
17	17	5.00	4.95	4.96	4.97	4.92	4.92	4.92	4.94	4.95	4.92	4.96	4.98	4.99	5.01					
18	18	4.84	4.81	4.81	4.82	4.77	4.77	4.79	4.80	4.81	4.80	4.83	4.85	4.88	4.88					
19	19	4.65	4.64	4.62	4.63	4.61	4.60	4.62	4.63	4.63	4.63	4.67	4.69	4.71	4.72					
20	20	4.55	4.55	4.54	4.54	4.54	4.53	4.54	4.56	4.57	4.57	4.60	4.63	4.65	4.67					

Fig. 3. Results for HOG_{C,B} feature over a single scale image at size 50×50 px with grid size $C = C_x = C_y$ (rows) and B bins (columns). The bordered cell shows the best value.

Parameter analysis for local features. In order to evaluate in depth the performance of the analyzed features for age estimation, we have conducted a deep analysis of the different parameters for the compared feature detectors. Table 2 lists the parametric choices that we have considered, and gives names to successful configurations for HOG, LBP and SURF descriptors that will be used for fusion experiments. The multiscale

¹ Cross-Validation folder structure of the images used for each database are available for comparison purposes in <https://sites.google.com/site/ivanhuertacasado/deepanalysisage>

Cx	Cy	B															
		5	6	7	8	9	10	11	12	13	14	15	16	17	18		
8	8				4.62	4.63	4.58	4.59	4.58	4.58							
9	9				4.50	4.51	4.48	4.48	4.47	4.48							
10	10	4.72	4.67	4.56	4.55	4.51	4.52	4.49	4.49	4.48	4.50	4.50	4.49	4.64	4.52		
11	11	4.61	4.56	4.48	4.47	4.43	4.44	4.42	4.43	4.43	4.44	4.45	4.44	4.47	4.48		
12	12	4.72	4.68	4.60	4.61	4.57	4.59	4.57	4.58	4.58	4.61	4.60	4.63	4.64	4.66		
13	13	4.74	4.73	4.61	4.62	4.58	4.60	4.57	4.57	4.57	4.59	4.58	4.59	4.59	4.62		
14	14	4.63	4.62	4.53	4.53	4.48	4.52	4.49	4.49	4.49	4.53	4.52	4.54	4.55	4.57		
15	15	4.52	4.51	4.45	4.45	4.41	4.45	4.42	4.43	4.44	4.47	4.46	4.49	4.51	4.54		
16	16				5.04	5.04	5.08	5.04	5.07	5.07	5.09	5.12					

Fig. 4. Results for the concatenation of $HOG_{C,B}^{\times 3}$ features over 3-scale images at 50×50 , 25×25 , and 13×13 px, with grid size $C = C_x = C_y$ (rows) and B bins (columns). The bordered cell shows the best value.

Cx	Cy	B															
		6	7	8	9	10	11	12	13	14	15	16	17	18			
7	7	5.39	5.13	5.09	4.97	4.95	4.87	4.88	4.85	4.82	4.82	4.81	4.80				
8	8	5.15	4.93	4.91	4.80	4.79	4.73	4.72	4.70	4.67	4.66	4.65	4.66				
9	9	4.85	4.70	4.65	4.59	4.59	4.53	4.51	4.49	4.48	4.48	4.47	4.48				
10	10	4.87	4.67	4.62	4.54	4.55	4.49	4.49	4.46	4.44	4.44	4.44	4.43				
11	11	4.64	4.50	4.48	4.41	4.42	4.37	4.37	4.36	4.34	4.35	4.34	4.34				
12	12	4.63	4.51	4.47	4.41	4.42	4.38	4.38	4.37	4.36	4.36	4.35	4.36				
13	13	4.52	4.41	4.38	4.33	4.33	4.30	4.29	4.28	4.28	4.28	4.28	4.28				
14	14	4.47	4.36	4.33	4.31	4.30	4.28	4.29	4.27	4.26	4.28	4.27	4.29				
15	15	4.37	4.28	4.26	4.23	4.23	4.21	4.22	4.20	4.20	4.21	4.22	4.24				
16	16	4.44	4.35	4.33	4.30	4.31	4.30	4.28	4.29	4.27	4.29	4.29	4.30				
17	17	4.36	4.28	4.26	4.24	4.25	4.23	4.23	4.23	4.22	4.24	4.24	4.25				
18	18	4.30	4.23	4.21	4.20	4.20	4.19	4.18	4.19	4.19	4.20	4.21	4.22				
19	19	4.26	4.20	4.18	4.17	4.18	4.17	4.16	4.17	4.17	4.19	4.19	4.22				
20	20	4.41	4.34	4.24	4.33	4.33	4.32	4.34	4.34	4.35	4.38	4.37	4.40				

Fig. 5. Results for $HOG_{C,B}$ feature for a single scale image at size 100×100 px, with grid size $C = C_x = C_y$ (rows) and B bins (columns). The bordered cell shows the best value.

Table 2. Reference tables summarizing the parametric choices we took to conduct the experiments, and the naming for recurrent configurations.

Scheme	Parameter	Description	Values	Image size (px \times px)
$HOG_{C,B}^{\times 3}$	$C=C_x=C_y$	#cells	{3, 4, ..., 20}	–
	B	#bins	{7, 8, ..., 20}	–
	S	#scales	{1, 3}	50 or 100 all 50, 25 and 13
$LBP_{P,R}^{u,2 \times 3}$	P	#neighbors	{8, 16}	–
	R	radius	{2, 3, ..., 10}	–
	S	#scales	{1, 3}	50 all 50, 25 and 13
$SURF_D^{\times 3}$	D	dimension (for base descriptor)	{64, 128}	–
	S	#scales	{1, 2, 3}	–
	V	scale values	{1.6, 1.8, 2, 2.4, 3, 4, 5}	50
DNN	architecture	$NC_kRP_k - NC_kRP_k - UFRD_k - F$		50
	N	#filters	{16, ..., 128}	–
	C_k	convolutional	{3, ..., 11}	–
	P_k	pooling	2	–
	U	#units	{256, ..., 1000}	–
	F	fully connected	–	–
	R	rectifier	–	–
	D_k	dropout	0.5	–
Name	Parameters	Image sizes (px \times px)		
HOG_A	$C_x=C_y=8, B=9, S=1$	50		
HOG_B	$C_x=C_y=15, B=13, S=1$	50		
LBP_A	$P=16, R=3, S=3$	50, 25 and 13		
$SURF_A$	$D=64, S=3, V=\{1.6, 2, 2.4\}$	50		
$SURF_B$	$D=128, S=3, V=\{1.6, 2, 2.4\}$	50		

versions result from concatenating base descriptors at different scales.

HOG parameters. When referring to $HOG_{C,B}$, we are con-

sidering a grid size $C_x \times C_y$ and number of bins B , whose optimal values have been obtained through exhaustive logarithmic grid search and 5-fold cross-validation, for single and multiple scales. Best results were obtained when $C_x=C_y=C$. As an implementation detail, a 50% cell overlapping for smoothness and global L2 normalization, instead of per-cell, have been used in our experiments. Other more sophisticated and systematic approaches could be used to reduce the parameters' combinations, but this is not the main focus of this paper. Multiscale variations are achieved by concatenating the feature vectors obtained by the descriptor at different scales. In order to have a fair comparison with the results reported in Guo and Mu (2014), images have been processed at 50×50 (similar to the 60×60 size used in that paper). However, we also evaluate the effect of different image sizes on the final performance in Figure 5, where images of size 100×100 were used. In summary, Figures 3, 4 and 5 report the individual analysis of HOG descriptors for a single scale at 50×50 pixels; for multiple scales (3 scales at 50×50 , 25×25 , and 13×13 ; and for a single scale at 100×100 , respectively. Figure 5 shows that 100×100 images provide even better scores than the traditional sizes in the literature, but we conduct the rest of experiments for 50×50 pixels for fair comparison. A single HOG scale performed better.

LBP parameters. For $LBP_{P,R}^{u,2}$ the analysis has been carried out by searching the optimal number of sampled neighbors P and radius R , for one and three scales, constraining the neighbors to be either 8 or 16, see Table 3. In the multiscale case, the smallest image size restricts the maximum radius to 6 pixels.

Table 3. MAE for the single-scale descriptor $LBP_{P,R}^{u,2}$ at 50×50 pixels, and for the 3-scale $LBP_{P,R}^{u,2 \times 3}$ concatenating 50×50 , 25×25 , and 13×13 . Neighborhoods of 8 and 16 are shown.

	(Size)	Radius R									
		2	3	4	5	6	7	8	9	10	
$LBP_{8,R}^{u,2}$	(59)	7.17	7.12	7.15	7.30	7.55	7.82	8.04	8.11	8.08	
$LBP_{16,R}^{u,2}$	(243)	6.88	6.70	6.66	6.76	7.06	7.25	7.40	7.51	7.81	
$LBP_{8,R}^{u,2 \times 3}$	(177)	6.48	6.49	6.66	6.82	10.75	-	-	-	-	
$LBP_{16,R}^{u,2 \times 3}$	(729)	6.18	6.13	12.41	11.32	12.26	-	-	-	-	

SURF parameters. In the case of SURF, 5 descriptors are extracted at a certain scale from fiducial markers at the eyes, nose tip and mouth corners, as provided during alignment, and concatenated into a single descriptor. Multiple scales have been tested for both the original and the extended descriptor ($SURF_{64}$ and $SURF_{128}$), as shown in Table 4.

Table 4. MAE results for SURF at one and multiple scale combinations (size in brackets).

Scale	$SURF_{64}$	$SURF_{128}$	Multiscale	$SURF_{64}^{\times 5}$	$SURF_{128}^{\times 5}$
1.6	6.09 (320)	5.72 (640)	{1.6, 2}	5.73 (640)	5.39 (1280)
1.8	6.21 (320)	5.77 (640)	{1.6, 2, 2.4}	5.71 (640)	5.41 (1280)
2.0	6.24 (320)	5.81 (640)	{2, 3}	5.95 (640)	5.60 (1280)
2.4	6.65 (320)	6.24 (640)	{1.6, 1.8, 2}	5.67 (960)	5.34 (1920)
3.0	6.93 (320)	6.59 (640)	{1.6, 2, 2.4}	5.59 (960)	5.30 (1920)
4.0	7.46 (320)	7.12 (640)	{1.6, 2.4, 3}	5.60 (960)	5.33 (1920)
5.0	7.52 (320)	7.26 (640)	{2, 2.4, 3}	5.84 (960)	5.53 (1920)

CCA optimal regularization. The optimal regularization cost γ^* , as defined in Section 3, differs for each computed feature and parameter. For this reason, initially the above-mentioned grid search has been performed without regularization ($\gamma = 0$). Once the best parameters for the feature detectors have been identified, the optimal regularization cost has been searched by looking for the minimum MAE. Additionally, we impose $\gamma_x = \gamma_y$. However, our experiments suggest that no significant changes are noticed when incorporating regularization due to the relative size of the database to the descriptor, as shown in Fig. 6. Each curve represents a subset of examples of different size. As the number of database examples M increases well over the feature dimensionality N , i.e. $M \gg N$, the optimal regularization cost γ^* (minimum of each curve) tends to zero.

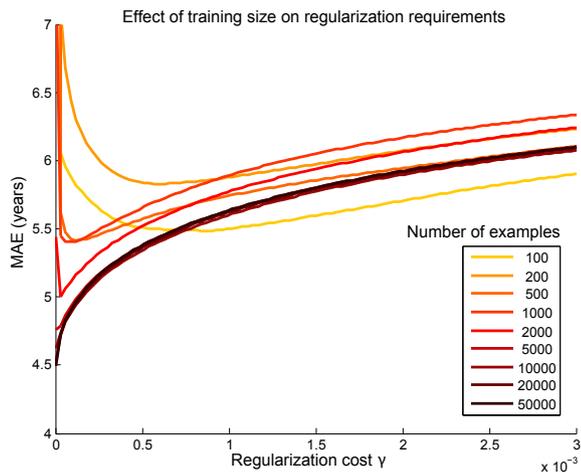


Fig. 6. The need for regularization depends strongly on the ratio between training examples M and feature dimensionality N . This figure shows 5CV results using 576-dimensional HOG_A over a single scale at 50×50 px and CCA, through different values of γ and increasing examples from 100 to 50K. As M increases the optimal γ^* decays, dropping to zero for $M \gg N$.

Feature combination. In order to improve the accuracy of the estimation, and taking advantage of the orthogonal nature of different descriptors, a thorough analysis of fusion combinations among feature candidates has been carried out. Different features are combined by simply concatenating them, as proposed, for instance, for pedestrian detection in Liang et al. (2012). Feature pooling and/or dimensionality reduction techniques (Huang et al. (2014)) might be used as well, but we prefer to stick with a simple approach and the obtained results reported in the following are promising. Similarly, we have employed an early-fusion strategy, combining the features from the very beginning, before the classification and decision take place. Other strategies could have been used, such as a late-fusion strategy, where each feature is coupled with its own classification, and the fusion is performed at decision level, as in Tan and Triggs (2010).

Although more combinations have been tested, Table 5 shows the most significant ones: single-scale $HOG_{8,9}$ and $HOG_{15,13}$, over 50×50 px images (HOG_A and HOG_B); 3-scale $LBP_{16,3}^{u2 \times 3}$, computed over 50×50 , 25×25 and 13×513 px images, and concatenated (LBP_A); the raw gradient magnitude δI over

50×50 px images; and the 3-scale $SURF_{64}^{\times 3}$ and $SURF_{128}^{\times 3}$ computed over 5 fiducial points at scales 1.6, 2, and 2.4, and concatenated ($SURF_A$ and $SURF_B$). Feature combinations concatenate the descriptors using the best parameters individually found, as described above.

Table 5. MAE for the fusion of the best descriptors. HOG_A , HOG_B and δI are computed over a single scale (50 px). LBP_A , $SURF_A$ and $SURF_B$ aggregate 3 scales. The best result is achieved by fusing $HOG_B + LBP_A + SURF_A$.

#	HOG_A	HOG_B	LBP_A	δI	$SURF_A$	$SURF_B$	(Size)	MAE
1	•						(576)	4.84
2		•					(2925)	4.38
3			•				(729)	6.13
4				•			(2500)	5.58
5					•		(960)	5.59
6						•	(1920)	5.30
7	•		•				(1305)	4.66
8	•		•	•			(3805)	4.53
9	•		•		•		(2265)	4.42
10	•		•			•	(3225)	4.61
11	•		•	•	•		(4765)	4.51
12	•		•	•		•	(5725)	4.72
13		•	•				(3654)	4.33
14		•		•			(5420)	4.33
15		•	•	•			(6154)	4.30
16		•			•		(3885)	4.30
17		•				•	(4845)	4.33
18		•			•		(4614)	4.27
19		•				•	(5574)	4.33
20		•	•	•	•		(7114)	4.31
21		•	•	•		•	(8074)	4.34
22			•	•			(3229)	5.07
23			•		•		(1689)	5.31
24			•			•	(2649)	6.45

The columns in Table 5 report a reference row number to ease the description in the following text, the feature names, the size of the combined descriptor, and its MAE. The table is vertically divided in four parts. The uppermost part (rows 1 to 6) shows, with a bullet in the corresponding column, the results with a single feature. The second part (rows 7 to 12) shows the combinations with HOG_A and LBP_A in common, while the third (rows 13 to 21) and fourth (rows 22 to 24) show the results with different combinations, by keeping HOG_B and LBP_A only fixed, respectively.

As observed from the summary of results in Table 5, $SURF_B$ reduces its MAE when fused with other features (from 5.30 years – row 6 – down to 4.33 when combined with HOG_B and LBP_A – row 17 and 19), and performs worse than $SURF_A$ under the same combination (see rows 9-10, 11-12, 16-17, 18-19, 20-21 and 23-24). However, when considered in isolation, $SURF_B$ performs better than $SURF_A$ (row 6 compared with row 5).

The best result is obtained when combining HOG_B , LBP_A and $SURF_A$. This combination has the advantage of fusing texture and local appearance-based descriptors. Another noticeable remark is the so-called curse of dimensionality: the addition of further descriptors into higher dimensional features not always enhances the result (compare, for instance, row 15 with 20 or 21, or row 8 with 12).

The specific size of the most accurate descriptors does not

seem to be correlated to their accuracy either, at least not after proper regularization has been applied. The HOG family of descriptors behaves particularly well for the different granularities that were tested, HOG_A and HOG_B , of 576 and 2925 dimensions respectively. This suggests that local appearance information is particularly useful and quite sufficient for capturing age patterns. The size of the descriptor deserves important consideration in the case of CCA, as it strongly affects the computational efficiency of the training process, and plays an important role in the stability of the solution: higher $\frac{M}{N}$ ratios result in more stable pseudo-inverse matrices when searching for the CCA projection matrix.

Table 6. Results for non-regularized CCA ($\gamma = 0$) and for CCA with the regularization cost γ^* yielding the best MAE, for each descriptor (BIF from Guo and Mu (2014)).

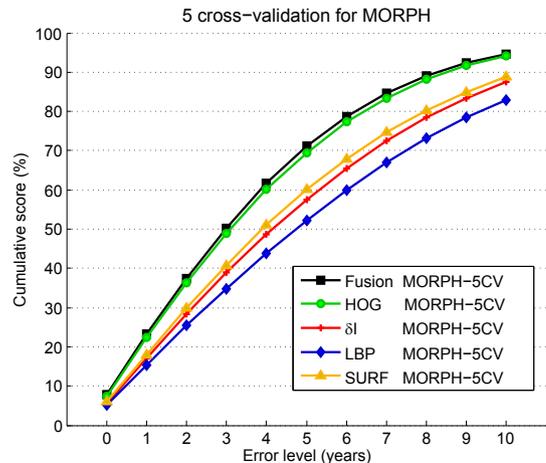
	HOG_B	δI	LBP_A	$SURF_B$	BIF	Fusion
(Size)	(2925)	(2500)	(729)	(1920)	(4376)	(4614)
MAE ($\gamma = 0$)	4.38	5.58	6.13	5.30	5.37	4.27
MAE (best γ^*)	4.34	5.49	6.13	5.29	4.42	4.25
	$\gamma^*=0.001$	$\gamma^*=0.002$	$\gamma^*\rightarrow 0$	$\gamma^*\rightarrow 0$	$\gamma^*=0.05$	$\gamma^*\rightarrow 0$

Table 6 shows the effect of regularization on the features that yielded best MAE scores in our experiments, over the MORPH database and using the regularized CCA regression technique. The optimal regularization costs are provided. We have also included the best results (to the best of our knowledge) achieved using the BIF descriptor, which is very commonly used in age estimation and provides the lowest MAE for MORPH in the literature (Guo and Mu (2014)). The size of BIF after dimensionality reduction (4376) is very similar to the proposed fusion without any further processing (4614). Nonetheless, our proposed fusion of local descriptors improves over the best registered result in this database, reducing it from 4.42 down to 4.25. It is noteworthy to see how differently regularization contributes to each descriptor. For instance, it does not affect LBP, but it improves BIF by 18%.

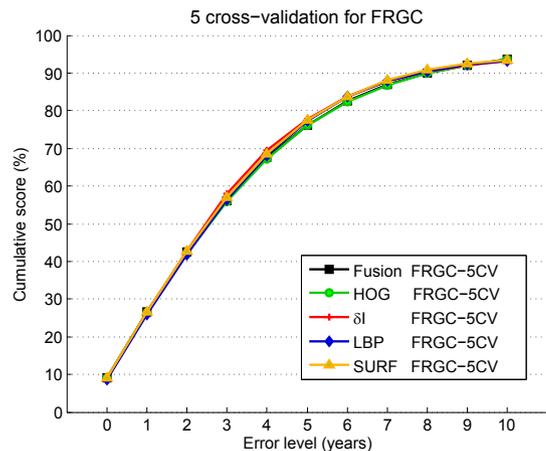
Table 7. MAE and CS(5) scores for MORPH and FRGC. Best possible descriptors are used.

	HOG_B		δI		MAE		$SURF_B$		Fusion		CS(5)	
	HOG_B	δI	LBP_A	$SURF_B$	LBP_A	$SURF_B$	LBP_A	$SURF_B$	LBP_A	$SURF_B$	LBP_A	$SURF_B$
MORPH-5CV	4.34	5.49	6.13	5.29	4.25							
FRGC-5CV	4.19	4.38	4.45	4.44	4.17	69.5%	57.6%	52.1%	60.2%	71.2%	76.0%	77.9%

Finally, these results have been obtained for FRGC as well. Table 7 contains global MAE errors and CS(5) values for MORPH and FRGC, whereas Figure 7 shows the complete cumulative score curves for error levels between 0 and 10. From Figure 7(a) it can be seen that for the MORPH database, the fusion of descriptors consistently improves over individual features, even for their optimal configuration of parameters and regularization. On the other hand, the FRGC curves are practically identical. As stated at the beginning of this section, this may be due to the lack of variability in the images of this database, in which every individual averages 80 images, and all very alike. In terms of MAE, the fusion of descriptors always obtains the best score.



(a)



(b)

Fig. 7. 5-fold cross-validation (5CV) Cumulative Score curves of the techniques evaluated in (a) MORPH and (b) FRGC.

Parameter analysis for deep learning. The experimental evaluation for deep neural networks has been conducted through variations of some of the most relevant parameters in the network architecture: the number and nature of the layers, the rate of the learning process, and specific internal parameters such as the number of filters or the size of the convolution kernels. The validation scheme continues to be 5CV, i.e. we divide the dataset into five folds, train the network parameters from scratch uniquely using the 50×50 images in four folds, test the network on the remaining one, and average the testing results from the five possible assignments. Axial symmetry of the faces has been exploited as a form of data augmentation.

The choice and adaptation of the learning rate is crucial for the success of a model, see Figure 8. Currently, common good practices include either using (i) an automatic, progressive rate update as the iterations progress, or (ii) a fixed rate with significant manual decrements after the learning curve stabilizes. Both intend to learn finer characteristics once an optimization minimum has been coarsely approached. During our evaluation, we decided to use a fixed learning rate and manually readjust it after a reasonable number of iterations, as this enabled us

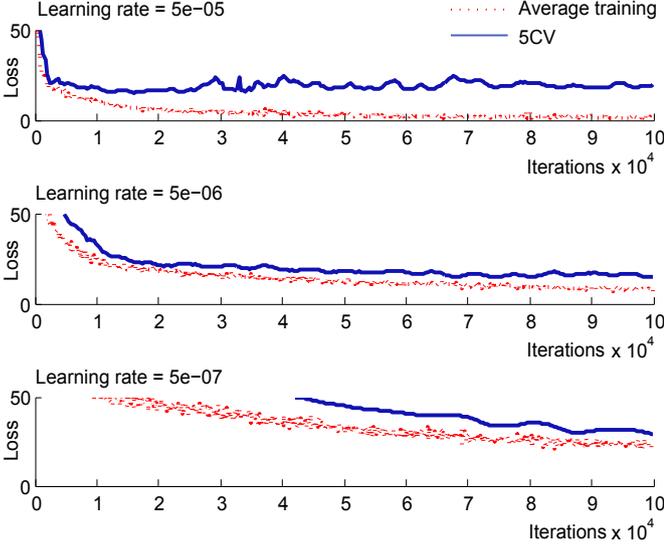


Fig. 8. The learning rate parameter is chosen first due to criticality. We show how validation loss evolves for 3 fixed values of learning rate, for the same architecture. These values yield, from top to bottom, to instability, convergence, and slow learning.

to better assess the effect of a single parameter in the network, when modifying it across different experiments.

Table 8 shows a significant subset of the experiments that were carried out while training the network. As it is common practice in deep learning approaches, our initial setup replicated a previously successful CNN (LeNet), and further adjustments were applied from that initial state. This architecture consists of a number of convolution and pooling layers, followed by a number of fully connected layers. In our case, the final stage is always a single regression unit.

We evaluated architectures with different number of layers (ranging from 1 to 3), layer types, number of units, activation functions, and regularization techniques. Concretely, we use the notation C_k for convolutional units of kernel size k ; P_k for max-pooling units of kernel size k ; and F for layers featuring full connection among the units. The strides are always set to 1 pixel for convolutions, and for pooling we always use the same value as their kernel size. R is included for those layers employing rectified linear units (ReLU), in the form $f(x) = \max(0, x)$. Layers with D_n incorporate dropout, i.e. random subsampling of n of the total units of the layer, which has been proved useful to prevent overfitting. Here, half of the neurons are randomly disconnected. For instance, $32C_{11}P_2 - 500FR - F$ represents a layer of 32 convolutional filters with 11-pixel kernels, followed by a max-pooling operation that reduces the output to half, a layer of 500 full-connected units with ReLU activation, and an output regressor. The learning rate and the weight decay of the network is explicitly stated for all the experiments in the table.

The computational requirements for training the deep neural network differ substantially from the previous approach. Regarding training time for one fold of 5CV-MORPH, 2-layer architectures take usually 6–7 hours on an i7-3770K with NVIDIA GTX770 graphics card using the Caffe framework (Jia et al. (2014)), whereas 3-layer ones need 8–9 hours.

Table 8. Selection of network configurations and their 5CV validation results on MORPH after 10⁵ iterations.

Architecture	Learning rate	Weight decay	MAE
$32C_{11}P_2 - 500FR - F$	10^{-5}	10^{-6}	4.09
$20C_{11}P_2 - 500FR - F$	10^{-5}	10^{-6}	4.15
$32C_7P_2 - 64C_7P_2 - 500FR - F$	10^{-5}	10^{-6}	4.39
$20C_7P_2 - 50C_9P_2 - 500FR - F$	10^{-5}	10^{-6}	4.48
$20C_7P_2 - 50C_9P_2 - 500FR - F$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	3.96
$20C_5P_2 - 50C_5P_2 - 500FR - F$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-6}$	3.97
$20C_5P_2 - 50C_5P_2 - 500FR - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-6}$	4.17
$20C_5P_2 - 50C_5P_2 - 500FR - F$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-8}$	5.75
$32C_{11}P_2 - 64C_9P_2 - 500FR - F$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-6}$	4.31
$32C_{11}P_2 - 64C_9P_2 - 500FR - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-6}$	4.25
$32C_{11}P_2 - 64C_9P_2 - 500FR - F$	$5 \cdot 10^{-7}$	$5 \cdot 10^{-8}$	6.14
$16C_3RP_2 - 32C_7R - 512FR - F$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-4}$	3.98
$16C_3RP_2 - 32C_7R - 512FR - F$	10^{-4}	10^{-3}	4.01
$16C_3RP_2 - 64C_7R - 256FR - F$	$5 \cdot 10^{-5}$	10^{-4}	3.96
$16C_3RP_2 - 64C_7R - 256FR - F$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-4}$	4.00
$32C_3RP_2 - 16C_7R - 512FR - 256FR - F$	$5 \cdot 10^{-5}$	10^{-3}	4.12
$20C_5P_2 - 50C_5P_2 - 512FR - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-7}$	4.14
$20C_5P_2 - 50C_5P_2 - 512FR - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-8}$	4.16
$20C_5P_2 - 50C_5P_2 - 512FR - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-9}$	4.14
$20C_5P_2 - 50C_5P_2 - 512FR - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-10}$	4.10
$20C_5P_2 - 50C_5P_2 - 512FR - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-11}$	4.12
$20C_5P_2 - 50C_5P_2 - 500FRD_{0.5} - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-11}$	3.90
$20C_5P_2 - 50C_5P_2 - 1000FRD_{0.5} - F$	$5 \cdot 10^{-6}$	$5 \cdot 10^{-11}$	3.88
$32C_3P_2 - 64C_5P_2 - 128C_3P_2 - 500FR - F$	10^{-5}	10^{-6}	4.07

Compared to them, extracting the chosen descriptors takes approximately 1 minute for HOG, 2 minutes for SURF and 12 seconds for LBP (less than 4 minutes for their fusion), and learning the CCA model of the fused descriptor takes 15 seconds. Prediction times also differ: for deep learning models it takes about 6 seconds, whereas CCA over fused descriptors is in the order of milliseconds. In general, we observe that many of the deep learning architectures produce similar results, and fine parametric adjustment progressively decreases the error. The inclusion of more layers and units increases the learning capacity of the model, but also contributes to its instability. By leveraging regularization techniques such as weight decay, rectifiers for sparsity, and unit dropout, we manage to achieve a more stable and accurate network, yielding a 5CV-MAE of 3.88 for MORPH.

Discussion. Table 9 summarizes some of the most relevant contributions to facial age estimation to date which supply cross-validation MAE over either MORPH or FRGC, including the methods proposed in this paper. Unlike ours, most of these contributions rely on flexible models with tens of fiducials (ASM or AAM), or hand-crafted BIF features. Moreover, our proposals exploit the whole available sets of 55K samples for MORPH and 44K samples for FRGC, by training from 4 folds, testing over the remaining one and averaging all five combinations.

The 5CV-MAE given by the early fusion of local descriptors improves over the best 5CV approach. On the other hand, the model obtained by the deep learning technique has produced a 5CV-MAE of 3.88 for MORPH. This value reduces the previ-

Table 9. Age estimation results in MORPH II and FRGC for the compared algorithms and visual descriptors, in a variety of settings.

MORPH-5CV	Technique	Proposed by	Feature	Train / test	MAE	CS(5)
	WAS	Lanitis et al. (2002)	AAM+BIF	55K	9.21	–
	AAS	Geng et al. (2013)	AAS+BIF	55K	10.10	–
	AGES	Geng et al. (2007, 2013)	AAM+BIF	55K	6.61	–
	RED (SVM)	Chang et al. (2011)	AAM	6K	6.49	48.9%
	OHRank	Chang et al. (2011)	AAM	6K	6.07	56.4%
	OHRank	Chang et al. (2011)	AAM+BIF	55K	6.28	–
	PLS	Guo and Mu (2011, 2013)	BIF	10K/55K	4.56	–
	kPLS	Guo and Mu (2011, 2013)	BIF	10K/55K	4.04	–
	IIS-LLD	Geng et al. (2013)	AAM+BIF	55K	5.67	–
	CPNN	Geng et al. (2013)	AAM+BIF	55K	4.87	–
	CCA	Guo and Mu (2013)	BIF	10K/55K	5.37	–
	rCCA	Guo and Mu (2013)	BIF	10K/55K	4.42	–
	kCCA	Guo and Mu (2013)	BIF	10K/55K	3.98	–
	MFOR	Weng et al. (2013)	PCA+LBP+BIF	4K	4.20	72.0%
	SVM+SVR	Han et al. (2013)	BIF+ASM	78K	4.20	72.4%
	SVR	Fernández et al. (2014)	HOG	55K	4.83	63.4%
	rCCA	<i>This paper</i>	Fusion	55K	4.25	71.17%
	CNN	<i>This paper</i>	CNN	55K	3.88	–
FRGC-5CV	Technique	Proposed by	Feature	Train / test	MAE	CS(5)
	rCCA	<i>This paper</i>	Fusion	44K	4.17	76.24%
	CNN	<i>This paper</i>	CNN	44K	3.31	–

ous best result for this database (Guo and Mu (2013)), which additionally was not employing a standard 5CV scheme due to computational limitations caused by the kCCA approach. The resulting CNN architecture has been also validated under FRGC, resulting in a 5CV-MAE of 3.31, thus also decreasing the previous result by fusing local descriptors.

5. Conclusions

Two very different techniques have been proposed for age estimation from facial images. The first method is based on the early fusion of local invariant descriptors coupled with a standard subspace learning technique, which requires few feature tuning, and demonstrates that local appearance and texture are sufficient for capturing age information. On the other hand, we also provided a powerful deep learning framework that couples the extraction and regression of meaningful cues by jointly optimizing both stages. Both approaches apply over eye-aligned 50×50 images, and do not require complex statistical facial models for precise alignment nor additional cues, unlike many traditional techniques for age estimation.

We have provided a thorough evaluation on the stability and effectiveness of these two approaches. Regarding local descriptors, our experiments show that the early fusion of HOG, LBP and SURF improves over the best MAE score reported using the non-kernelized CCA technique, resulting in 4.25 years compared to the 4.42 of hand-crafted BIF at 60×60 pixels. The experiments also show that this distance can be further increased when using larger images as it has been demonstrated using a single HOG descriptor (MAE 4.16). On the other hand, our deep learning architecture, although requiring more specific parameter tuning, decreased the minimum error to date from 3.98

to 3.88, without imposing the restriction on the number of training samples caused by the kernel matrix size in kCCA. We explored the robustness of these techniques in terms of parameter settings and in the presence and lack of regularization.

Overall, we have conducted a quite comprehensive set of experiments on the two largest and most used datasets to date (MORPH and FRGC). These experiments aim at not only demonstrating the superior accuracy of the proposed approaches (as described above), but also to draw some considerations about the dimensionality of the feature used. In fact, as a lesson learned, even though combining multiple orthogonal features may result in lower MAE, it also increases the complexity, and for some classifiers such as CCA and kCCA this may bring instability.

We can imagine future directions in our research. First of all, alternative feature fusion strategies, such as feature pooling or sophisticated dimensionality reduction techniques, as well as late-fusion strategies, should be developed and tested to either confirm that our simple combination of feature suffices or to show better performance. Next, other features should be tested both in isolation and combined with other features. Other possible future directions include the addition of a frontalization stage during preprocessing, particularly important when dealing with real images, which are rarely frontal. Additionally, the proposed deep learning approach can be further refined by new forms of data augmentation, the exploitation of multiscale versions of the input image, and carefully designed deeper network architectures. Finally, the evaluation can be extended by further investigating the distribution of errors across age ranges, gender and ethnicity; and the generalization capabilities can be tested through cross-database validation schemes.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation (MICINN) through the Torres-Quevedo funding program (PTQ-11-04401 and PTQ-11-04400). We thank the reviewers for their useful comments and suggestions.

References

- Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features, in: *Computer Vision—ECCV 2006*. Springer, pp. 404–417.
- Chang, K.Y., Chen, C.S., Hung, Y.P., 2011. Ordinal hyperplanes ranker with cost sensitivities for age estimation, in: *CVPR, IEEE*. pp. 585–592.
- Choi, S.E., Lee, Y.J., Lee, S.J., Park, K.R., Kim, J., 2011. Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition* 44, 1262–1281.
- Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: *CVPR, IEEE*. pp. 886–893.
- Fernández, C., Huerta, I., Prati, A., 2014. A comparative evaluation of regression learning algorithms for facial age estimation, in: *FFER in conjunction with ICPR*, in press, IEEE.
- Gallagher, A.C., Chen, T., 2009. Understanding images of groups of people, in: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE*. pp. 256–263.
- Geng, X., Yin, C., Zhou, Z.H., 2013. Facial age estimation by learning from label distributions, in: *TPAMI, IEEE*. pp. 2401–2412.
- Geng, X., Zhou, Z.H., Smith-Miles, K., 2007. Automatic age estimation based on facial aging patterns. *TPAMI* 29, 2234–2240.
- Gunay, A., Nabyev, V.V., 2008. Automatic age classification with lbp, in: *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on, IEEE*. pp. 1–4.
- Guo, G., Mu, G., 2011. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression, in: *CVPR, IEEE*. pp. 657–664.
- Guo, G., Mu, G., 2013. Joint estimation of age, gender and ethnicity: CCA vs. PLS, in: *10th Int. Conf. on Automatic Face and Gesture Recognition, IEEE*.
- Guo, G., Mu, G., 2014. A framework for joint estimation of age, gender and ethnicity on a large database. *Image and Vision Computing*.
- Guo, G., Mu, G., Fu, Y., Huang, T.S., 2009. Human age estimation using bio-inspired features, in: *CVPR, IEEE*. pp. 112–119.
- Han, H., Otto, C., Jain, A.K., 2013. Age estimation from face images: Human vs. machine performance, in: *International Conference on Biometrics (ICB), IEEE*.
- Huang, R., Ye, M., Xu, P., Li, T., Dou, Y., 2014. Learning to pool high-level features for face representation. *The Visual Computer*, 1–13.
- Huerta, I., Fernández, C., Prati, A., 2014. Facial age estimation through the fusion of texture and local appearance descriptor, in: *Soft Biometrics in conjunction with ECCV*, in press, IEEE.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.
- Lanitis, A., Draganova, C., Christodoulou, C., 2004. Comparing different classifiers for automatic age estimation. *TSMC-B* 34, 621–628.
- Lanitis, A., Taylor, C.J., Cootes, T.F., 2002. Toward automatic simulation of aging effects on face images. *TPAMI* 24, 442–455.
- Liang, J., Ye, Q., Chen, J., Jiao, J., 2012. Evaluation of local feature descriptors and their combination for pedestrian representation, in: *Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE*. pp. 2496–2499.
- Minear, M., Park, D.C., 2004. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers* 36, 630–633.
- Montillo, A., Ling, H., 2009. Age regression from faces using random forests, in: *ICIP, IEEE*. pp. 2465–2468.
- Ojala, T., Pietikainen, M., Maenpaa, T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24, 971–987.
- Oro, D., Fernández, C., Saeta, J.R., Martorell, X., Hernando, J., 2011. Real-time GPU-based face detection in HD video sequences, in: *ICCV Workshops*, pp. 530–537.
- Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W., 2005. Overview of the Face Recognition Grand Challenge, in: *CVPR, IEEE*. pp. 947–954.
- Ricanek, K., Tesafaye, T., 2006. MORPH: a longitudinal image database of normal adult age-progression, in: *Automatic Face and Gesture Recognition*, pp. 341–345. doi:10.1109/FGR.2006.78.
- Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. *Nature neuroscience* 2, 1019–1025.
- Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection, in: *CVPR, IEEE*. pp. 3476–3483.
- Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: Closing the gap to human-level performance in face verification, in: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE*. pp. 1701–1708.
- Tan, X., Triggs, B., 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on* 19, 1635–1650.
- Weng, R., Lu, J., Yang, G., Tan, Y.P., 2013. Multi-feature ordinal ranking for facial age estimation, in: *AFGR, IEEE*.
- Yang, M., Zhu, S., Lv, F., Yu, K., 2011. Correspondence driven adaptation for human profile recognition, in: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE*. pp. 505–512.
- Yang, Z., Ai, H., 2007. Demographic classification with local binary patterns, in: *Advances in Biometrics*. Springer, pp. 464–473.