Facial Age Estimation Through the Fusion of Texture and local appearance Descriptors

Ivan Huerta[1], Carles Fernández[2], and Andrea Prati[1]

[1] DPDCE, University IUAV, Santa Croce 1957, 30135 Venice, Italy
huertacasado@iuav.it, aprati@iuav.it
[2] Herta Security, Pau Claris 165 4-B, 08037 Barcelona, Spain
carles.fernandez@hertasecurity.com

Abstract. Automatic extraction of soft biometric characteristics from face images is a very prolific field of research. Among these soft biometrics, age estimation can be very useful for several applications, such as advanced video surveillance [5, 12], demographic statistics collection, business intelligence and customer profiling, and search optimization in large databases. However, estimating age from uncontrollable environments, with insufficient and incomplete training data, dealing with strong person-specificity, and high within-range variance, can be very challenging. These difficulties have been addressed in the past with complex and strongly hand-crafted descriptors, which make it difficult to replicate and compare the validity of posterior classification schemes. This paper presents a simple yet effective approach which fuses and exploits texture- and local appearance-based descriptors to achieve faster and more accurate results. A series of local descriptors and their combinations 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.

Keywords: Age estimation, CCA, HOG, LBP, SURF

1 Introduction

The problem of age estimation from images has historically been one of the most challenging within the field of facial analysis. Some of the reasons are the uncontrollable nature of the aging process, the strong specificity to the personal traits of each individual [24], high variance of observations within the same age range, and the fact that it is very hard to gather complete and sufficient data to train accurate models [7].

This process can be made easier by having available large and representative collections of age-annotated images. However, in the past the available databases were often very limited and strongly skewed. This is especially disadvantageous for applications like video surveillance and forensics, which need to work correctly when facing unknown subjects and a lack of any additional cues. Fortunately, the recent availability of large databases like MORPH [21] and FRGC [20] offers a great opportunity to make advances in the field. Keeping in mind that any training data set which is representative of the whole population cannot exist, the only viable option is to develop methods that are able to exploit large databases in order to gain substantial generalization capabilities.

The inherent difficulties in the facial age estimation problem, such as limited imagery, challenging subject variability, and 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, 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. 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 extracted and processed.

On the other hand, during the last decade, several fields within image classification and object recognition have proposed different families of very fast and descriptive feature extraction schemes, which have become well-known for being especially invariant to rotation, scale, illumination, and alignment. Such histogram-based descriptors, which typically capture local intensity variations or local neighborhood patterns from spatial grids, are nowadays a fundamental tool to deal with highly adverse and unconstrained environments for a variety of applications.

In this paper we conduct a thorough evaluation of a series of common local visual descriptors, in order to investigate their utility towards the automatic facial age estimation problem. The contributions are as follows:

- We review some of the most efficient and effective local visual descriptors from image classification, and explore their suitability to extract age-related discriminative patterns.
- We demonstrate that the fusion of textural and local appearance-based descriptors achieves state-of-the-art results, improving over complex feature extraction schemes that were previously proposed.
- Candidate descriptors are exhaustively evaluated regarding optimal parameters and regularization, in terms of mean average errors and cumulative score curves over two large databases.

The paper is structured as follows: next section gathers and comments on previous related work on facial age estimation. The candidate descriptors to be evaluated are reviewed in Section 3, along with the selected classification scheme. Evaluation is presented out in Section 4, by first describing available large databases with age annotations, and subsequently analyzing the extensive experiments carried out over the combinations of local descriptors. Finally, Section 5 summarizes the results and draws some conclusions.

2 Related work

After an initial interest on automatic age estimation from images dated back in the early 2000s [13–15], research in the field has experienced a renewed interest from 2006 on, since the availability of large databases like MORPH-Album 2 [21], which increased by 55× the amount of real age-annotated data with respect to traditional age databases. Therefore, this database has deeply been employed in recent works by applying over it different descriptors and classification schemes.

Feature extraction scheme. Regarding visual features, flexible shape and appearance models such as ASM (Active Shape Model) and AAM (Active Appearance Model) have been some of the primary cues used to model aging patterns [2, 7, 8, 13]. Such statistical models capture the main modes of variation in shape and intensity observed in a set of faces, and allow face signatures based on such characterizations to be encoded.

Bio-Inspired Features (BIF) [22] and its derivations have consistently been used for age estimation in the last years [7, 12]. These feed-forward models consist of a number of layers intertwining convolutionally and pooling processes. 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 downscales the results with a non-linear reduction, typically a MAX or STD operation, progressively encoding the results into a vector signature. In [17], 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 [9–11].

Features extracted from local neighborhoods have very rarely been used for the purpose of age estimation. In [24], LBP histogram features are combined with principal components of BIF, shape and textural features of AAM, and PCA projection of the original image pixels. HOG features have independently been used for age estimation in [4].

Classification scheme. With regards to the learning algorithm, several approaches have been proposed, including, among others, Support Vector Machines (SVM) / Support Vector Regressors (SVR) [17, 12, 2, 24], neural networks [13] and their variant of Conditional Probability Neural Network (CPNN) [7], Random Forests (RF) [16], and projection techniques such as Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA), along with their regularized and kernelized versions [9–11]. An extensive comparison of these classification schemes for age estimation has been reported in our previous paper [4], and in particular the advantageousness of CCA was demonstrated over the others, both regarding accuracy and efficiency.

Specific attention must be given to the CCA technique, which is the main focus of this paper from the classification perspective. The PLS and CCA subspace learning algorithms were originally conceived to model the compatibility between two multi-dimensional variables. PLS uses latent variables to learn a new space in which such variables have maximum correlation, 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 [10], although in that case the size of training folds is limited to 10K samples due to computational limitations.

The main contribution 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 orthogonal 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. To the best of our knowledge, this approach has never been employed before for age estimation, and our experiments demonstrate comparable performance with respect to state-of-the-art results provided by complex and fine-tuned feature extraction schemes such as BIF [11]. 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

Preprocessing. In general, existing works tackle the problem of age estimation with visual features that are either complex and fine-tuned (e.g., BIF), or require precise statistical models involving tens of facial landmarks for accurate alignment (e.g., ASM and AAM). As opposed to them, we do not rely on precisely aligned appearance models; instead, our experiments will be evaluated using a simple alignment through the fiducial landmarks of the detected eye regions.

The facial region of each image has been detected with the face detector described in [19]. The relative alignment invariance of local descriptors based on concatenated cell histograms allows us to work with simple eye-aligned images. The fiducial markers corresponding to the eye centers have been obtained using the convolutional neural network for face alignment presented in [23]. 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 [10], which use input images of 60×60 pixels, our aligned image are resized to only 50×50 pixels.

Descriptors. The choice of visual features to be extracted from aligned images and sent to the classification scheme 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) [3] have largely been used as robust visual descriptors in many computer vision applications related to object detection and recognition. The horizontal and vertical gradients of the input image are computed, and 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 $\text{HOG}_{C,B}$ to denote $C\times C$ square grids and B orientation bins.

Local Binary Patterns (LBP) [18] 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. A common extension considers generic pixel neighborhoods formed by P sampled pixel values at radius R from the central pixel. 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 and radius using uniform patterns is referred as $LBP^{u2}_{P,R}$, e.g. $LBP^{u2}_{8,2}$.

Speeded-Up Robust Features (SURF) [1] 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 descriptor component of 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, SURF$_{64}$. 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 SURF$_{128}$.

As gradient information is typically a very relevant cue to describe image content for many image descriptors, we have included raw magnitude gradient images (GRAD) 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 [11]. 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

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

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, labels in $\mathbf{Y}$ are unidimensional, so a least squares fitting suffices to relate these labels to the projected data features. Thus, only w_x is computed, by solving the following generalized eigenvalue problem:

$$\mathbf{X}\mathbf{Y}^T \left(\mathbf{Y}\mathbf{Y}^T +_y I\right)^{-1} \mathbf{Y}\mathbf{X}^T w_x = \lambda \left(\mathbf{X}\mathbf{X}^T + I\right) w_x \tag{2}$$

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 in the following manner:

$$\mathbf{XY}^T \left((1-\gamma_y)\mathbf{YY}^T + \gamma_y I\right)^{-1} \mathbf{YX}^T w_x = \lambda \left((1-\gamma_x)\mathbf{XX}^T + \gamma_x I\right) w_x \tag{3}$$

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, in that case covariance matrices become computationally intractable with over 10K samples. In practice, regularized CCA works comparably to KCCA [10], it is much less computationally demanding, and will allow us to reproduce the same exact validation schemes over large databases.

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 labeled with accurate age information. Table 1 shows the summary of the existing databases with main reference, number of samples, number of subjects, and comments.

Database	Samples	Subjects	Comments
PAL [15]	580	580	Limited number of samples
FG-NET [14]	1,002	82	Limited number of samples and subjects
GROUPS [6]	28,231	28,231	Ages discretized into seven age intervals
FRGC v2.0 [20]	44,278	568	Large database; many samples per subjects
MORPH II [21]	55,134	13,618	Large database; high diversity

Table 1. Description of the existing databases for age estimation.

From the information in Table 1, we see that PAL and FG-NET are comparatively negligible to the rest in terms of number of samples. Additionally, age annotations in GROUPS 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 [4]. Figure 1 offers a graphical visualization and comparison of the

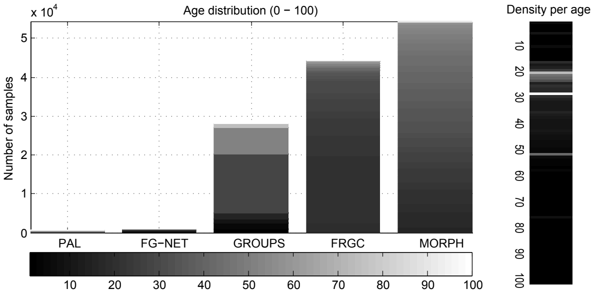


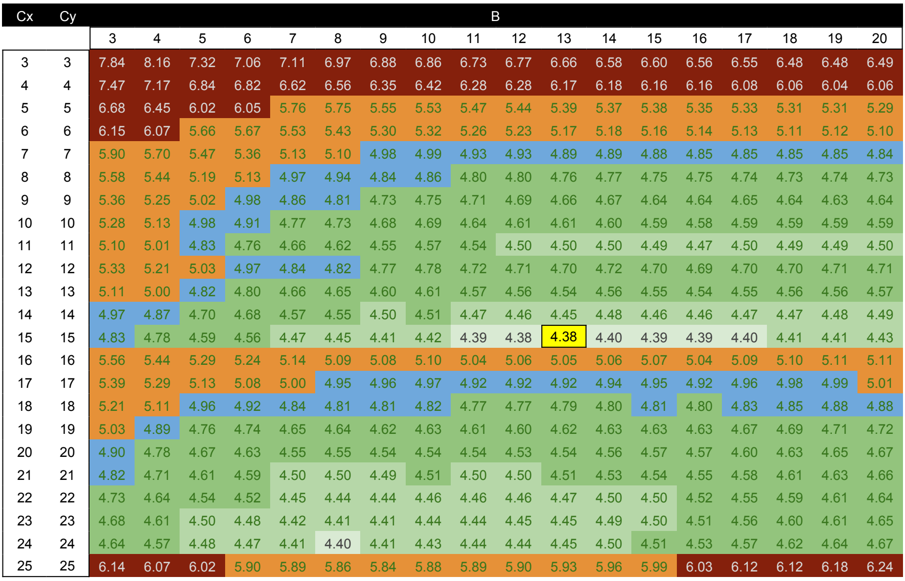
Fig. 1. Age distribution and density per database. In the left graphic (Age distribution) different ages are represented by the intensity. In the right graphic (Desity per age) the intensity represent the density (white color more density). PAL and FG-NET are relatively negligible compared to others, and GROUPS only provides age intervals, so we focus on MORPH II and FGRC. Age samples are mainly skewed towards 20–30 and 50 year old.

analyzed databases, by number of samples and density of age ranges.

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 $CS(l) = M_{e \leq l}/M$ [2, 24, 12] . Related publications typically supply either an eleven-point curve for age deviations $[0 - 10]$, or simply the value $CS(5)$.

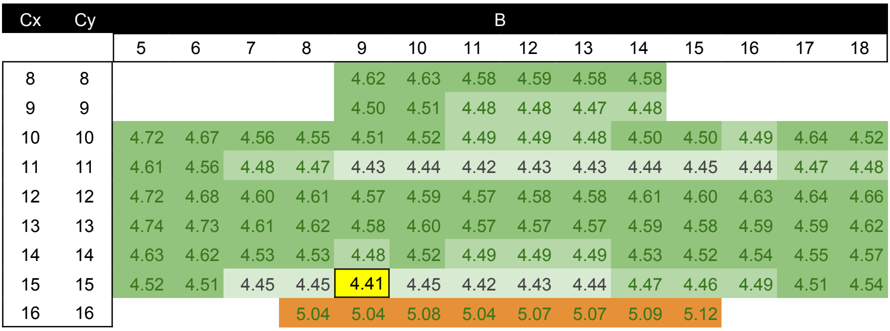
All through the rest of this paper, the optimal parameters are searched so as to minimize the MAE score over MORPH, using 5-fold cross-validation in all cases. In particular, the division into training and validation sets is made so that all the instances of the same subject are contained in one single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

Parameter analysis. In order to evaluate in depth the performance of the analyzed features for age estimation, we have conducted an analysis of the different parameters for the compared feature detectors. In the case of HOG$_{C,B}$, the optimal parameters for grid size $C{\times}C$ and number of bins B have been obtained through exhaustive logarithmic grid search and 5-fold cross-validation, for single and multiple scales. 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 [11], images have been processed at $50{\times}50$ (similar to the $60{\times}60$ size used in that paper). However, we also evaluate the effect of different image sizes on the final performance in Fig. 4, where images of size $100{\times}100$ were used. In summary, Figs. 2, 3 and 4 report the individual analysis of HOG descriptors for a single scale at $50{\times}50$ pixels; for 3-scales at $\{50{\times}50,\ 25{\times}25,\ 13{\times}13\}$; and for a single scale at $100{\times}100$, respectively.



Cx	Cy	B																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3	7.84	8.16	7.32	7.06	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	7.47	7.17	6.84	6.82	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	6.68	6.45	6.02	6.05	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	6.15	6.07	5.66	5.67	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	5.90	5.70	5.47	5.36	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	5.58	5.44	5.19	5.13	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	5.36	5.25	5.02	4.98	4.86	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	5.28	5.13	4.98	4.91	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	5.10	5.01	4.83	4.76	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	5.33	5.21	5.03	4.97	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	5.11	5.00	4.82	4.80	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	4.97	4.87	4.70	4.68	4.57	4.55	4.50	4.51	4.47	4.46	4.45	4.48	4.46	4.46	4.47	4.47	4.48	4.49
15	15	4.83	4.78	4.59	4.56	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	5.56	5.44	5.29	5.24	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	5.39	5.29	5.13	5.08	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	5.21	5.11	4.96	4.92	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	5.03	4.89	4.76	4.74	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	4.90	4.78	4.67	4.63	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
21	21	4.82	4.71	4.61	4.59	4.50	4.50	4.49	4.51	4.50	4.50	4.51	4.53	4.54	4.55	4.58	4.61	4.63	4.66
22	22	4.73	4.64	4.54	4.52	4.45	4.44	4.44	4.46	4.46	4.46	4.47	4.50	4.50	4.52	4.55	4.59	4.61	4.64
23	23	4.68	4.61	4.50	4.48	4.42	4.41	4.41	4.44	4.44	4.45	4.45	4.49	4.50	4.51	4.56	4.60	4.61	4.65
24	24	4.64	4.57	4.48	4.47	4.41	4.40	4.41	4.43	4.44	4.44	4.45	4.50	4.51	4.53	4.57	4.62	4.64	4.67
25	25	6.14	6.07	6.02	5.90	5.89	5.86	5.84	5.88	5.89	5.90	5.93	5.96	5.99	6.03	6.12	6.12	6.18	6.24

Fig. 2. Results for HOG$_{C,B}$ feature for a single scale with image size 50×50 at varying grid size C (rows) and number of bins B (columns). The bordered cell shows the best value.

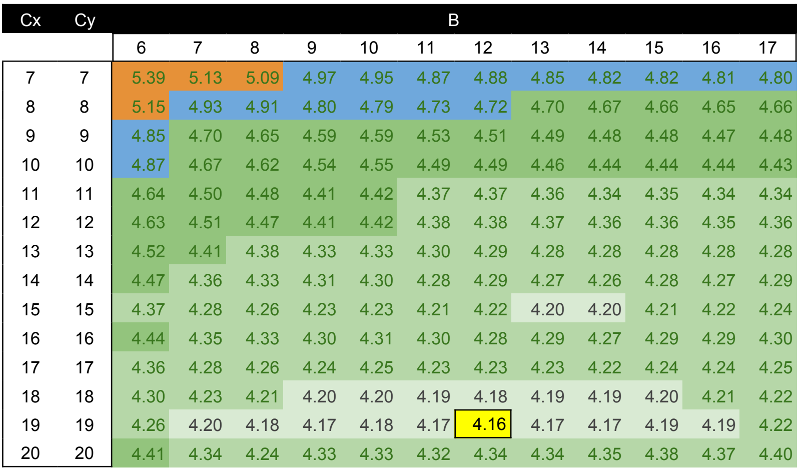


Cx	Cy	B													
		5	6	7	8	9	10	11	12	13	14	15	16	17	18
8	8					4.62	4.63	4.58	4.59	4.58	4.58				
9	9					4.50	4.51	4.48	4.48	4.47	4.48				
10	10	4.72	4.67	4.56	4.55	4.51	4.52	4.49	4.49	4.48	4.50	4.50	4.49	4.64	4.52
11	11	4.61	4.56	4.48	4.47	4.43	4.44	4.42	4.43	4.43	4.44	4.45	4.44	4.47	4.48
12	12	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. 3. Results for HOG$^{\times 3}_{C,B}$ feature for 3 scales concatenating descriptors over 50×50, 25×25, and 13×13 images, at varying grid size C (rows) and number of bins B (columns). The bordered cell shows the best value.

Fig. 4 shows that 100×100 images provide even better scores than the traditional sizes in the literature, although we conduct the rest of experiments for 50×50 pixels for fair comparison. Single scale HOG performed better than multiscale.

A similar grid search procedure has been chosen to optimize the parameters of LBP and SURF descriptors. In the case of LBP$^{u2}_{P,R}$ 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 number of neighbors to either 8 or 16, see Table 2. In the case



Cx	Cy	B											
		6	7	8	9	10	11	12	13	14	15	16	17
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. 4. Results for HOG$_{C,B}$ feature for a single scale with size image 100×100 at varying grid size C (rows) and number of bins B (columns). The bordered cell shows the best value.

	(Size)	Radius R								
		2	3	4	5	6	7	8	9	10
LBP$^{u2}_{8,R}$	(59)	7.17	7.12	7.15	7.30	7.55	7.82	8.04	8.11	8.08
LBP$^{u2}_{16,R}$	(243)	6.88	6.70	**6.66**	6.76	7.06	7.25	7.40	7.51	7.81
LBP$^{u2\times3}_{8,R}$	(177)	6.48	6.49	6.66	6.82	10.75	-	-	-	-
LBP$^{u2\times3}_{16,R}$	(729)	6.18	**6.13**	12.41	11.32	12.26	-	-	-	-

Table 2. MAE for the single-scale descriptor LBP$^{u2}_{P,R}$ at 50×50 pixels, and for the 3-scale LBP$^{u2\times3}_{P,R}$ concatenating 50×50, 25×25, and 13×13. Neighborhoods of 8 and 16 are shown.

Scale	SURF$_{64}$	SURF$_{128}$	Multiscale	SURF$^{\times S}_{64}$	SURF$^{\times S}_{128}$
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, 2.4}	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)

Table 3. MAE results for SURF at one and multiple scale combinations. Size in brackets.

of SURF, multiple scales have been tested for both the original and extended descriptor (SURF$_{64}$ and SURF$_{128}$), as shown in Table 3.

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 optimal (minimum) MAE. Additionally, we impose $\gamma_x = \gamma_y$. However, our experiments suggest that no significant changes can be noticed when incorporating regularization because of the relative size of the database to the descriptor, as shown in Fig. 5. As the number of database examples M increases well over the dimensionality of the feature N, i.e. $M \gg N$, the optimal regularization cost γ^* tends to zero.

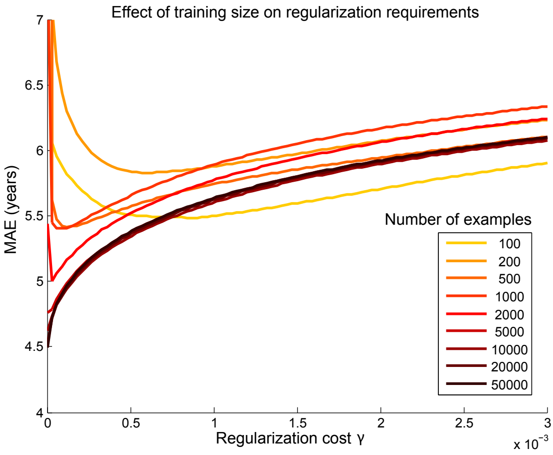


Fig. 5. The need for regularization depends strongly on the ratio between training examples M and feature dimensionality N. This figure shows 5-fold cross-validation results using 576-dimensional $\text{HOG}_{8,9}$ and CCA, through different values of γ and increasing examples from 100 to 50K. As M increases the optimal γ^* decays, dropping to zero for $M \gg N$.

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. Although more combinations have been tested, Table 4 shows the most significant ones: single-scale $\text{HOG}_{8,9}$ and $\text{HOG}_{15,13}$; 3-scale $\text{LBP}^{u2\times3}_{16,3}$; the raw gradient magnitude GRAD; and the 3-scale $\text{SURF}^{\times3}_{64}$ and $\text{SURF}^{\times3}_{128}$ with scales 1.6, 2, and 2.4. Feature combinations have been obtained by concatenating the descriptors and exploiting the best parameters obtained previously.

As observed from the results summarized in Table 4, $\text{SURF}^{\times3}_{128}$ reduces its MAE when fused with other features (from 5.30 years down to 4.33 when combined with $\text{HOG}_{15,13}$ and $\text{LBP}^{u2\times3}_{16,3}$), and performs worse than SURF_{64} under the same combination. The best result is obtained when combining $\text{HOG}_{15,13}$, $\text{LBP}^{u2\times3}_{16,3}$ and $\text{SURF}^{\times3}_{64}$. This combination has the advantage of fusing texture and local appearance-based descriptors. Another noticeable remark is the so-called curse of dimensionality: the ad-

$\text{HOG}_{8,9}$	$\text{HOG}_{15,13}$	$\text{LBP}^{u2\times3}_{16,3}$	GRAD	$\text{SURF}^{\times3}_{64}$	$\text{SURF}^{\times3}_{128}$	(Size)	MAE
●						(576)	4.84
	●					(2925)	4.38
		●				(729)	6.13
			●			(2500)	5.58
				●		(960)	5.59
					●	(1920)	5.30
$\text{HOG}_{8,9}$	$\text{HOG}_{15,13}$	$\text{LBP}^{u2\times3}_{16,3}$	GRAD	$\text{SURF}^{\times3}_{64}$	$\text{SURF}^{\times3}_{128}$	(Size)	MAE
●		●				(1305)	4.66
●		●	●			(3805)	4.53
●		●		●		(2265)	4.42
●		●			●	(3225)	4.61
●		●	●	●		(4765)	4.51
●		●	●		●	(5725)	4.72
	●	●				(3654)	4.33
	●		●			(5420)	4.33
	●	●	●			(6154)	4.30
	●			●		(3885)	4.30
	●				●	(4845)	4.33
	●	●		●		(4614)	**4.27**
	●	●			●	(5574)	4.33
	●	●	●	●		(7114)	4.31
	●	●	●		●	(8074)	4.34
		●	●			(3229)	5.07
		●		●		(1689)	5.31
		●			●	(2649)	6.45

Table 4. MAE results for the fusion of different descriptors that yielded best results. $\text{HOG}_{8,9}$ and $\text{HOG}_{15,13}$ have a single scale. $\text{LBP}^{u2\times3}_{16,3}$ is computed at the original, half and quarter image size. GRAD is formed concatenating all gradient magnitude values. $\text{SURF}^{\times3}_{64}$ and $\text{SURF}^{\times3}_{128}$ are aggregated SURF descriptors with scales $\{1.6, 2, 2.4\}$. The best result is achieved by combining $\text{HOG}_{15,13}$, $\text{LBP}^{u2\times3}_{16,3}$, and $\text{SURF}^{\times3}_{64}$.

dition of further descriptors into higher dimensional features not always enhances the result.

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, $\text{HOG}_{8,9}$ and $\text{HOG}_{15,13}$, 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.

	$\mathrm{HOG}_{15,13}$	GRAD	$\mathrm{LBP}_{16,3}^{u2\times3}$	$\mathrm{SURF}_{128}^{\times3}$	BIF [11]	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**
	(γ^*=0.001)	(γ^*=0.002)	($\gamma^*\to$0)	($\gamma^*\to$0)	(γ^*=0.05)	($\gamma^*\to$0)

Table 5. Results for non-regularized CCA ($\gamma = 0$) and for CCA with the regularization cost γ^* yielding the best MAE, for each descriptor.

Table 5 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 [11]. 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%.

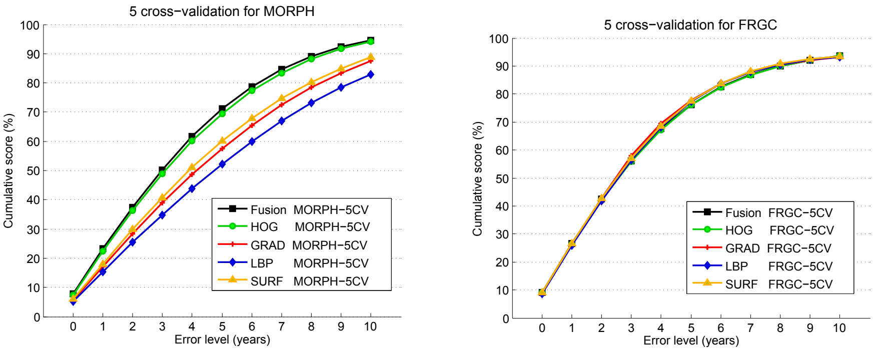
Finally, these results have been obtained for FRGC as well. Table 6 contains global MAE errors and CS(5) values for MORPH and FRGC, whereas Figure 6 shows the complete cumulative score curves for error levels between 0 and 10. From Figure 6(a) it can be seen that for the MORPH database, the fusion of 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.

	MAE					CS(5)				
	HOG	GRAD	LBP	SURF	Fusion	HOG	GRAD	LBP	SURF	Fusion
MORPH–5CV	4.34	5.49	6.13	5.29	**4.25**	69.5%	57.6%	52.1%	60.2%	**71.2%**
FRGC–5CV	4.19	4.38	4.45	4.44	**4.17**	76.0%	**77.9%**	77.4%	77.5%	76.2%

Table 6. MAE and CS(5) scores for MORPH and FRGC. Each descriptor has optimal parameters.

5 Conclusions

We have provided a thorough evaluation on the effectiveness of local invariant descriptors, both individually and combined, towards the automatic estimation of apparent age



(a) (b)

Fig. 6. 5-fold cross-validation (5CV) Cumulative Score curves of the Feature descriptor techniques evaluated in: (a) MORPH and (b) FRGC databases.

from facial images, using a standard classification technique. In our experiments, the early fusion of HOG, LBP and SURF descriptors over eye-aligned images provides state-of-the-art results over two large databases, MORPH and FGRC. Concretely, the proposed fusion of descriptors at 50×50 pixel images improves over the best MAE score reported using the CCA technique, resulting in 4.25 years compared to the 4.38 of BIF at 60×60 pixels. Our experiments also show that this distance can be further increased when using larger images and a single HOG descriptor (MAE 4.16).

Our approach requires few feature tuning; it does not involve statistical face models requiring precise annotation of tens of facial landmarks; and it does not require additional cues. We have explored the robustness of the descriptors in terms of parameter settings and in the presence and lack of regularization. Finally, we have demonstrated that local appearance information is sufficient for capturing age information from faces, although it is further improved with textural cues. Canonical Correlation Analysis has proved to be a very effective and efficient technique for age estimation, working consistently for an ample variety of descriptors.

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).

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer Vision–ECCV 2006, pp. 404–417. Springer (2006)
2. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: CVPR. pp. 585–592. IEEE (2011)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. vol. 1, pp. 886–893. IEEE (2005)

4. Fernández, C., Huerta, I., Prati, A.: A comparative evaluation of regression learning algorithms for facial age estimation. In: FFER in conjunction with ICPR, in press. IEEE (2014)
5. Fu, Y., Guo, G., Huang, T.: Age synthesis and estimation via faces: A survey. TPAMI 32(11), 1955–1976 (2010)
6. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 256–263. IEEE (2009)
7. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. In: TPAMI. vol. 35, pp. 2401–2412. IEEE (2013)
8. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. TPAMI 29(12), 2234–2240 (2007)
9. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: CVPR. pp. 657–664. IEEE (2011)
10. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: CCA vs. PLS. In: 10th Int. Conf. on Automatic Face and Gesture Recognition. IEEE (2013)
11. Guo, G., Mu, G.: A framework for joint estimation of age, gender and ethnicity on a large database. Image and Vision Computing (2014)
12. Han, H., Otto, C., Jain, A.K.: Age estimation from face images: Human vs. machine performance. In: International Conference on Biometrics (ICB). IEEE (2013)
13. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. TSMC-B 34(1), 621–628 (2004)
14. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. TPAMI 24(4), 442–455 (2002)
15. Minear, M., Park, D.C.: A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, & Computers 36(4), 630–633 (2004)
16. Montillo, A., Ling, H.: Age regression from faces using random forests. In: ICIP. pp. 2465–2468. IEEE (2009)
17. Mu, G., Guo, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: CVPR. pp. 112–119. IEEE (2009)
18. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24(7), 971–987 (2002)
19. Oro, D., Fernández, C., Saeta, J.R., Martorell, X., Hernando, J.: Real-time GPU-based face detection in HD video sequences. In: ICCV Workshops. pp. 530–537 (2011)
20. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the Face Recognition Grand Challenge. In: CVPR. pp. 947–954. IEEE (2005)
21. Ricanek, K., Tesafaye, T.: MORPH: a longitudinal image database of normal adult age-progression. In: Automatic Face and Gesture Recognition. pp. 341–345 (2006)
22. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature neuroscience 2(11), 1019–1025 (1999)
23. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR. pp. 3476–3483. IEEE (2013)
24. Weng, R., Lu, J., Yang, G., Tan, Y.P.: Multi-feature ordinal ranking for facial age estimation. In: AFGR. IEEE (2013)