Real-Time Age Estimation from Face Imagery Using Fisher Vectors

Lorenzo Seidenari[1](✉), Alessandro Rozza[2], and Alberto Del Bimbo[1]

[1] University of Florence, Firenze, Italy
{lorenzo.seidenari,alberto.delbimbo}@unifi.it
[2] Hyera Software, Coccaglio, Italy
alessandro.rozza@hyera.com

Abstract. In the last decade facial age estimation has grown its importance in computer vision. In this paper we propose an efficient and effective age estimation system from face imagery. To assess the quality of the proposed approach we compare the results obtained by our system with those achieved by other recently published methods on a very large dataset of more than 55K images of people with different gender and ethnicity. These results show how a carefully engineered pipeline of efficient image analysis and pattern recognition techniques leads to state-of-the-art results at 20FPS using a single thread on a 1.6GHZ i5-2467M processor.

Keywords: Age estimation · Face analysis · Biometrics

1 Introduction

In the last decade age estimation from facial imagery has grown its importance in the computer vision field. The process of age determination has many potential application areas, such as: age-based access control and verification, where a person's age is verified prior to physical access to a place or product being sold or virtual access to a website is granted; age-adaptive human-computer interaction, where as example, a digital sign can display advertisements based on the age of the audience walking past; age-based indexing of face images, that is the use of age as criterion for indexing into huge-scale biometric databases for faster retrieval.

To guarantee the success of all the aforementioned applications it is required to obtain fast (or real-time) estimation of the attribute of interest (the age). This requirement is particularly severe when it exists a limited window of time for a decision based on the outcome, such as when a person walks past a digital sign. Nevertheless, even in the case of the usage of age as criterion for indexing into huge-scale biometric databases, high speed of the age estimation algorithms are required to make it operationally viable.

Age estimation is usually performed as a multi-class classification problem or as a regression task. In the first case, given an image feature $\phi(\boldsymbol{I})$ computed

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 172–182, 2015.
DOI: 10.1007/978-3-319-23234-8_17

from a face image $\boldsymbol{I}$, the task is to predict the class associated to the interval (*age group*) containing the actual age. Precisely, the age labels are quantized in a set of age groups, e.g. $\{[16, 25], [26, 35] \ldots [56, 65]\}$. This approach is intuitive but has a few drawbacks. First of all, if the aim is a precise estimation the age groups must be kept small, but this comes at the cost of reducing the amount of positive samples per class and increasing the dataset imbalance. Second, if the relationship among labels is discarded, a classical classification loss function would equally penalize errors among close and distant age groups.

In the regression case the age is treated as a real number and a function $\text{age}\,(\phi(\boldsymbol{I}))$ is estimated to minimize the age estimation error. This approach has several advantages over the multi-class classification task. First, all data can be used to fit a single model. This avoids the quantization problem and reduces the amount of models needed to estimate the age leading to higher efficiency at evaluation time. Second, the loss function can be formulated more naturally penalizing models proportionally to the error they commit.

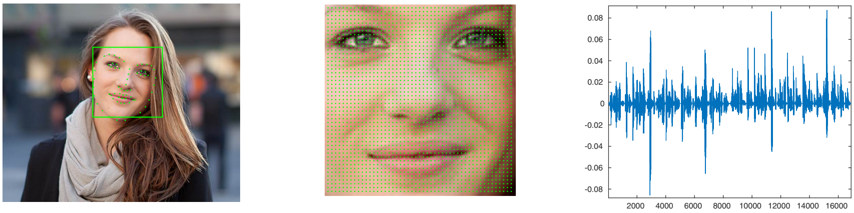
Many related works exploit shape features based on active appearance models [1] and Biological Inspired Features (BIF). BIF are firstly proposed for age estimation by Guo *et al.* [2] combined with a linear SVM. In this work the authors employ a pyramid of Gabor filters with small sizes and they suggest to determine the number of orientations and bands with a problem-specific approach, rather than using a predefined number. In [3] Guo *et al.* investigate the variations of age estimation performance under variations across race and gender. They observe that crossing race and gender can result in significant error increases for age estimation. To leverage the aging pattern of different gender and ethnicity they employ the feature presented in their previous work [2] and they propose a 3-step method learning separate classifiers for different combinations of age and genders and applying the age estimator only after predicting the gender and ethnicity of the subject.

Guo *et al.* also propose to use the kernel partial least squares regression (KPLS) for age estimation [4]. The strength of this approach is twofold. First, the KPLS simultaneously performs the feature dimensionality reduction and learns the aging function; furthermore, since KPLS can find a small number of latent variables to reduce the dimensionality of the original space, this can improve the efficiency of the proposed approach.

In [5] a hierarchical part based representation for face age estimation has been proposed. This method identifies different facial components and extracts BIF feature vectors describing these parts; subsequently, each facial component is classified into one of four disjoint age groups using a binary decision tree based on SVM; finally, a separate SVM age regressor is trained to predict the actual age.

Chang *et al.* in [6] proposed an ordinal hyperplane ranker on Active Appearance Models (AAM [1]) exploiting the distribution of training labels. The key idea is try to obtain multiple decisions on who is the older of two people to finally determine the person's actual age. To perform this task the authors present an approach that is able to efficiently compute the input face age as the result of a series of comparisons between the target face and the training ones, and then to estimate the person's age by integrating the result. Precisely, all the facial images are separated by each ordinal hyperplane into two groups according to the relative order, and a cost-sensitive property is exploited to find better hyperplanes based on the classification costs. The actual age is inferred by aggregating a set of preferences from the ordinal hyperplanes with their cost sensitivities.

Geng *et al.* propose two algorithms exploiting the label distributions [7] of the face images. Instead of considering each face image as an instance with a single label (the age), the author consider each face image as an instance associated with a label distribution. The label distribution covers a certain number of class labels, representing the degree that each label describes the instance. This approach guarantees that one face image can contribute also to the learning of its adjacent ages. One of the main assumptions of the first proposed algorithm is that the distribution of each face image can be derivated by the maximum entropy model. Nevertheless, there is no particular evidence supporting it in the problem of age estimation. To relax this assumption the authors propose to use a three layer neural network to approximate the distributions. A comprehensive list of recent age estimation approaches can be found in [5].



(a) Landmark estimation (b) Feature Sampling (c) Fisher Vector

Fig. 1. Our image representation pipeline. Face detection and landmark estimation (a) followed by dense multi-scale SIFT extraction on the aligned face (b) and Fisher Vector computation (c).

In this paper we describe our age estimation system (see Fig. 1) designed with efficiency in mind. Differently from previous works we use a high-dimensional modern feature [8] that proves to be accurate yet efficient. We use regularized linear regression that is efficient to evaluate requiring a single dot product per face and allows to directly minimize the error in years.

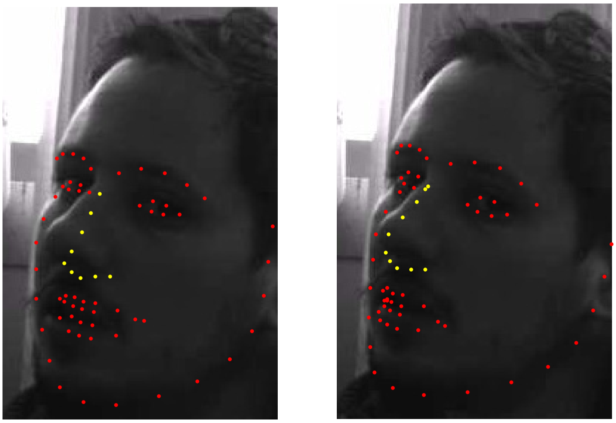
This paper is organized as follows: in Section 2 the employed face detection approach and the alignment technique are described; in Section 3 our face representation is summarized; in Section 4 the regression approach used for age estimation is presented; in Section 5 the achieved results on a very large dataset are shown; in Section 6 our conclusions are highlighted.

2 Fast Face Detection and Alignment

The first block of our processing chain is an image pre-processing one followed by face detection. Subsequently, face alignment is performed in order to obtain a consistent geometric reference for image features. These steps will be exploited in the face representation step as explained in Sect. 3.

2.1 Features and Image Pre-processing

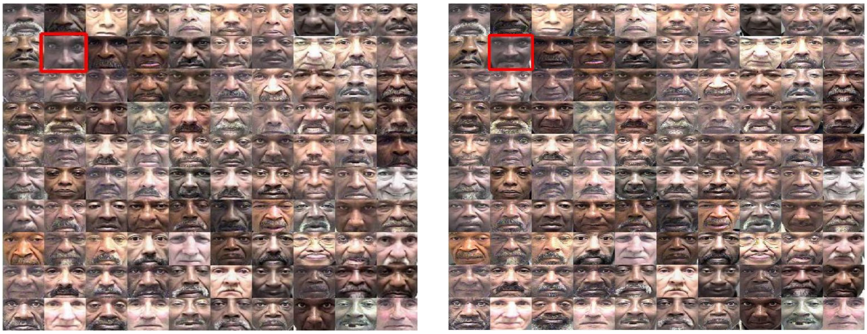
To avoid missing faces, especially in highly saturated images, we apply an histogram equalization to the image. Several approaches have been developed to normalize images in order to gain invariance to illumination. Usually these techniques aim at normalizing a face crop in a way that recognition does not suffer from illumination variations. In our case we are interested in reducing the effect of sensor saturation in presence of strong lighting. Our concern is to detect as many faces as possible and have a reliable landmark estimation without sacrificing real-time performance. Among many available algorithms we evaluated rank normalization and wavelet based normalization [9]. In our experiments we found that, for detection and landmark estimation purposes, i.e. to keep discriminative features from faces a basic histogram equalization is enough to guarantee high recall. As can be seen in Fig. 2 attempting to estimate landmarks without normalization may result in poor localization. Indeed, in Fig. 2a all nose landmarks are wrongly localized in the image processed without equalization.



(a) Landmarks estimated without equalization.

(b) Landmarks estimated with equalization.

Fig. 2. Face landmark detection without (a) and with (b) equalization on a challenging image. Nose landmarks, marked in yellow, are wrongly localized without equalization.



(a) Rotation alignment.

(b) Affine transform alignment.

Fig. 3. Alignment results with rotation compensation and with affine alignment. In the face marked in red the mouth is missing in the rotated image whilst using the affine compensation all important facial features are visible.

2.2 Face Detection

We use a very simple yet effective multi-pose linear classifier. The model is trained with structural SVM on ∼3000 faces with 5 poses: frontal, profile-left, profile-right, frontal left-tilted and frontal right-tilted. We used the structural SVM formulation of [10], this method allows very fast training and state-of-the art results even with linear classifiers.

2.3 Face Alignment

Our face representation exploits the joint statistics of pixel intensities and locations. In order to make the representation invariant to face pose we have applied a face alignment step. To do so we rescale and align the detected faces to a common reference square. The simplest face alignment approach consists to estimate the angle of the line intersecting the eye centers and tilting the face image. As can be seen in Fig. 3(a) for many faces the mouth is not always fully visible thus discarding important features. We instead apply an affinity based alignment. The affinity, performing a non uniform scaling along the two dimensions, allows to align the whole face in a common reference. We estimate the affine transformation matrix, i.e. rotation scale and translation, mapping the triangle defined by the eye and mouth centers and a canonical triangle defined as $(0.2 \cdot S, 0.2 \cdot S), (0.8 \cdot S, 0.2 \cdot S), (0.5 \cdot S, 0.5 \cdot S)$ where S is the square size. As highlighted by Fig. 3(b) all important facial features can be recovered.

To estimate the eye and mouth centers we firstly extract the 68 landmarks provided by [11] which implements a face shape estimation using a cascade of regression trees trained on pixel intensities. Robust estimates of eyes and mouth centers are obtained using the median of the 6(eye) and 20(mouth) landmarks describing these parts of the face. Finally, we remap detected faces in a square with 100 pixel side using the aforementioned affine transformation.Our method efficiently deals with poses with little yaw (±15°) for higher pose variation a full 3D approach should be used to improve results[12]. Our face detection and alignment solution runs at 30 FPS on a i5-2467M 1.60GHz CPU using a single thread.

3 Face Representation

Our face representation is inspired by recent image classification techniques based on local features [13] and face verification [14]. After face alignment we extract the face patch and we resize it to a fixed scale (as described in Sect. 2).We sample dense SIFT [15] descriptors without orientation and scale estimation. Even if faces are rescaled at a fixed size different features may appear at different patch scales, therefore we apply multi-scale sampling. Thanks to the face alignment we are able to exploit feature location. We compute Fisher vectors over SIFT descriptors augmented with their x,y coordinate rescaled in $[-1, 1]$. Before computing Fisher vectors we learn 64 PCA components on a set of 200K randomly sampled SIFT features. The final local feature is obtained concatenating the PCA reduced SIFT descriptor and the rescaled x,y coordinates. Considering the learned dictionary employing a Gaussian Mixture Model with parameters $\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \boldsymbol{\omega}_n$ and given soft-assignments $\gamma_m^{(n)}$ for each of the M augmented SIFT feature $\boldsymbol{x}_m \in \boldsymbol{X}$, the Fisher vector is computed concatenating the following gradients:

$$\mathcal{G}_n^{\mu}(\boldsymbol{X}) = \frac{1}{\sqrt{\boldsymbol{\omega}_n}} \sum_{m=1}^{M} \gamma_m^{(n)} \left(\frac{\boldsymbol{x}_m - \boldsymbol{\mu}_n}{\boldsymbol{\sigma}_n^2} \right), \tag{1}$$

$$\mathcal{G}_n^{\sigma}(\boldsymbol{X}) = \frac{1}{\sqrt{2\boldsymbol{\omega}_n}} \sum_{m=1}^{M} \gamma_m^{(n)} \left(\frac{(\boldsymbol{x}_m - \mu_n)^2}{\boldsymbol{\sigma}_n^2} - 1 \right), \tag{2}$$

where

$$\gamma_m^{(n)} = \frac{\boldsymbol{\omega}_n p_n(\boldsymbol{x}_m)}{\sum_{j=1}^{D} \boldsymbol{\omega}_j p_j(\boldsymbol{x}_m)}, \tag{3}$$

and p_n is the n^{th} Gaussian of the learned mixture and $\mathbf{X}$ is the feature set of a face image. Considering a vocabulary of size $D = 128$, the final image feature size is $66 \times 128 \times 2 = 16896$.

4 Large Scale Learning with SGD

Most of the best performing methods for age estimation rely on regression, this is indeed the natural approach to overcome quantization errors that occur for classification based approaches. Our feature representation is extremely high dimensional, therefore a linear regressor is likely to obtain good performance with very low evaluation cost.

Table 1. MORPH-II dataset gender and ethnicity statistics.

Race	Female	Male	Female and Male
Black	5,757	36,803	42,560
White	2,601	7,999	10,600
Hispanic	100	1,651	1,751
Asia	13	146	159
India	14	43	57
Other	2	3	5
Total	8,487	46,645	55,132

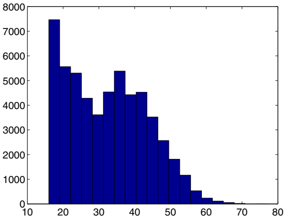


Fig. 4. MORPH-II age distribution.

From an applicative point of view, using a single linear regressor has many advantages. First, it reduces the memory footprint with respect to a multi-class classification approaches; second, avoiding kernels has also a strong impact in the evaluation time of the regressor allowing us to compare each detected face with just one hyperplane instead of computing a kernel evaluation per support vector.

Our aim is to estimate a weight vector $\boldsymbol{w}$ and a bias b given an image $\boldsymbol{I}$ and a feature function $\phi(\cdot)$ to produce an age estimate:

$$\text{age}(\phi(\boldsymbol{I})) = \langle \boldsymbol{w}, \phi(\boldsymbol{I}) \rangle + b \tag{4}$$

To efficiently train our regressor we apply stochastic gradient descent (SGD) to L2-regularized least square regression or ridge regression, optimizing the following equation:

$$\frac{1}{2}\lambda||\boldsymbol{w}||^2 + \frac{1}{n}\sum_{i=1}^{N} \left(\langle \boldsymbol{w}, \phi(\boldsymbol{I}) \rangle + b - y_i\right)^2 \tag{5}$$

Considering a vast amount of training samples SGD is efficient and accurate as also noticed in [16]. We set $\lambda = 1/(C \cdot N))$, where N are the training samples, and tune the parameter C by five fold cross-validation of MAE on the training set.

5 Experimental Results

We test our approach on the MORPH-II dataset that contains more than 55K facial images with different gender and ethnicity. In Table 1 the detailed statistics of gender and ethnicity are shown, whilst in Fig. 4 the age distribution is summarized.

5.1 Timing

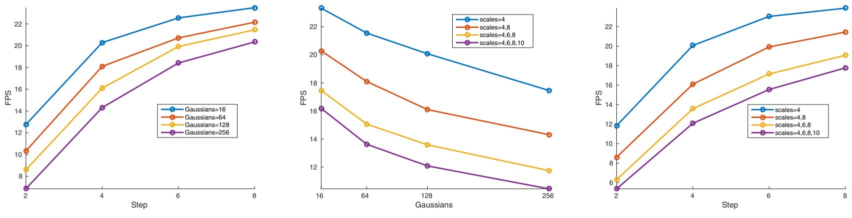
We run a set of benchmarks to evaluate the run time of our method. The system speed is mostly affected by the density of feature sampling both in scale and size as can be seen in Figs. 5a and 5c since the sampling step quadratically affects the amount of features extracted.

Furthermore, the number of Gaussians affects the computation time in two ways. First, with a large vocabulary single feature embeddings are slower to compute, since they need to calculate more derivatives. Second, increasing the final feature size the regression step is longer, even though the regression step time is negligible with respect to the feature computation step cost.

In Table 2 we have reported the FPS of some commercial systems presented in [17]. It is possible to notice that the best performing commercial frameworks obtain comparable performance results with those achieved by our approach but they are tested on a more powerful 6-cores Intel Xeon Processor X5690 CPU with respect to our 1.6GHZ i5-2467M processor, moreover they are implemented using multi threads. This results confirm that our method reaches state-of-the-art performance.

Table 2. FPS of commercial systems reported in [17]. Notice that the best performing commercial frameworks obtain comparable results with those achieved by our approach but they are tested on a more powerful 6-cores Intel Xeon Processor X5690 CPU and they are implemented using multi threads.

System	FPS
Our Approach	20
Junyu Tech.	15
Zhuhau-Yisheng	10
MITRE	27
Tsinghua University	11
NEC	19
Cognitech	5



(a) (b) (c)

Fig. 5. Frame rate of the proposed processing pipeline for different dictionary size, sampling in space and scale. We set scales=4,8 in (a) step=4 (b) and Gaussians=128 (c). Face detection and alignment is included.

5.2 Accuracy

We have assessed the quality of our method using the Mean Absolute Error or MAE $= \frac{1}{N}\sum_{i=1}^{N} |\text{age}(\phi(\boldsymbol{I})) - y_i|$.

In order to compare our results with those achieved by recently published methods we have ran a set of experiments with different experimental setups. Results are summarized in Table 4. To compare our results with [6,18] we have used the same photos used by the authors: a set of 5,492 images taken from people of Caucasian descent. We have reported the average of MAE over 30 trials.

To compare our results with those proposed in [2–4] we have followed the procedure specified in [3]. Given the whole dataset $\mathcal{W}$ we have defined a set $\mathcal{S} \subset \mathcal{W}$ of $\sim$ 21000 images of black and white individuals keeping all the women and an amount of men to keep the proportion between males and females 1:3. We have further split this set in $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ and $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$. Moreover, we have generated $\mathcal{S}_1$ and $\mathcal{S}_2$ so that identities of people in $\mathcal{S}_1$ are not allowed in $\mathcal{S}_2$ and vice versa. We have trained the regressor on $\mathcal{S}_i$ and we have reported the average of MAE obtained on $\mathcal{W} \setminus \mathcal{S}_i$ for $i = 1, 2$.

Finally since with our approach we can leverage a huge amount of data we have split the dataset using 80% of the identities for the training and 20% for testing and we have ran a 10-fold cross-validation. We have not stratified the sampling on gender and ethnicity but an empirical check has shown that randomly sampling identities keep the subsampled sets distribution of age, ethnicity and gender similar to the distribution on the whole set. This setup is the same proposed in [7].

In Table 3 we have shown how MAE varies depending on the feature extraction step using the setup of [7]. It is possible to notice that the only parameters affecting the MAE are the sampling step and the amount of Gaussians. A sufficiently tight sampling step is critical to ensure a wide coverage of all the facial features. At the same time a dictionary with too few Gaussians is unable to capture the SIFT descriptor statistics for faces. Instead, the amount of scales is not affecting the accuracy, this is mostly due to the fact that faces are all aligned and scaled at the same size so there is no need to match image patches representing the same structure at different scales.

In Table 4 we have compared our results with those achieved by some approaches tested on MORPH Album2 dataset. The first setup [2–4] is the easiest since it employes a single ethnicity. The second and third setups deal with multiple ethnicities and gender, with the second [6,18] using only black and white people and the third using the whole dataset [7].

These results show that our method is not limited to be trained on a single ethnicity or gender, nor require any strategy to deal with cross-racial or cross-gender influence in age estimation.

Table 3. Mean absolute error varying sampling step, scales and Gaussians. We used 128 Gaussians in (a) and step=4 and scales=4,8 in (b). The algorithm is mostly affected by the sampling step.

(a)

Scales	Sampling	MAE
4,6,8,10	2	3.7
4,8	2	3.7
4,8	4	3.7
4,8	8	4.0

(b)

Gaussians	MAE
16	4.2
64	3.8
128	3.7
256	3.6

Table 4. Mean Absolute Error (MAE) in years compared with recently published methods. Our method obtains state-of-the-art results with a very low-weight processing pipeline.

Approach	Features	Classifier	MAE [6,18]	MAE[2–4]	MAE[7]
Our approach	**SIFT+FV**	**L2L2 Regression**	**3.8**	**4.0**	**3.7**
Geng*et al.* [7]	AAM,BIF	CPDNN	-	-	4.9
Geng *et al.* [7]	AAM,BIF	IIS-LLD	-	-	5.7
Guo *et al.* [4]	Holistic BIF	Kernel PLS	-	4.2	-
Guo *et al.* [3]	Holistic BIF	3-Step	-	4.5	-
Guo *et al.* [2]	Holistic BIF	Linear SVM	-	5.1	-
Chang *et al.* [6]	AAM	Ordinal Hyperplane Ranker	6.1	-	
Chang *et al.* [18]	AAM	Ranking SVM	6.5	-	-

6 Conclusions

In this paper we have proposed a real-time age estimation system from face imagery. We have shown how a carefully engineered pipeline of efficient image analysis and pattern recognition techniques leads to state-of-the-art results. Our single threaded approach runs at 20 FPS on a 1.6GHZ i5-2467M processor, thus leaving room for further improvement. Furthermore, we have found that employing very densely sampled SIFT features and a large dictionary decreases the mean absolute age estimation error; nevertheless, this configuration conflicts with our real-time aim. With this in mind, we have identified another setting that obtains a low drop in performance (.1 years of MAE, for details see Sect. 5) but guaranteeing a real-time system.

To assess the quality of our framework we have tested our approach on a very large dataset of more than 55K images of people with different gender and ethnicity. We tested our method on different settings comprising the whole dataset or reducing it to a smaller single ethnicity version. Our method results compared with those achieved by other recently published approaches confirm the efficiency and the effectiveness of the proposed framework.

Acknowledgments. Lorenzo Seidenari is supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01_00034_23154_SMST.

References

1. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**, 681–685 (2001)
2. Guo, G., Mu, G., Fu, Y., Huang, T.: Human age estimation using bio-inspired features. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 112–119 (2009)
3. Guo, G., Mu, G.: Human age estimation: What is the influence across race and gender? In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 71–78 (2010)
4. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 657–664 (2011)
5. Han, H., Otto, C., Jain, A.K.: Age estimation from face images: Human vs. machine performance. In: International Conference on Biometrics, ICB 2013, June, 4–7, Madrid, Spain (2013)
6. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 585–592 (2011)
7. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 2401–2412 (2013)
8. Snchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision 105, 222–245 (2013)
9. Shan, D., Ward, R.: Wavelet-based illumination normalization for face recognition. In: 2005 International Conference on Pattern Recognition (ICPR) (2005)
10. King, D.E.: Max-Margin Object Detection. ArXiv e-prints (2015)
11. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR (2014)
12. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: Proc. of CVPR (2015)
13. Seidenari, L., Serra, G., Badanov, A.D., Del Bimbo, A.: Local pyramidal descriptors for image recognition. Transactions on Pattern Analisys and Machine Intelligence (2013)
14. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher Vector Faces in the Wild. In: British Machine Vision Conference (2013)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**, 91–110 (2004)
16. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Good practice in large-scale learning for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**, 507–520 (2014)
17. Ngan, M., Grother, P.: Face recognition vendor test (frvt) performance of automated age estimation algorithms. Technical report, NIST (2014)
18. Chang, K.Y., Chen, C.S., Hung, Y.P.: A ranking approach for human ages estimation based on face images. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 3396–3399 (2010)