This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

Age from faces in the deep learning revolution

Vincenzo Carletti, Antonio Greco, Gennaro Percannella and Mario Vento, *IAPR Fellow*

Abstract—Face analysis includes a variety of specific problems as face detection, person identification, gender and ethnicity recognition, just to name the most common ones; in the last two decades, significant research efforts have been devoted to the challenging task of age estimation from faces, as witnessed by the high number of published papers. The explosion of the deep learning paradigm, that is determining a spectacular increasing of the performance, is in the public eye; consequently, the number of approaches based on deep learning is impressively growing and this also happened for age estimation. The exciting results obtained have been recently surveyed on almost all the specific face analysis problems; the only exception stands for age estimation, whose last survey dates back to 2010 and does not include any deep learning based approach to the problem.
This paper provides an analysis of the deep methods proposed in the last six years; these are analysed from different points of view: the network architecture together with the learning procedure, the used datasets, data preprocessing and augmentation, and the exploitation of additional data coming from gender, race and face expression. The review is completed by discussing the results obtained on public datasets, so as the impact of different aspects on system performance, together with still open issues.

Index Terms—Age estimation, deep learning, face analysis, survey, review

1 Introduction

Between the end of the 80s and the first decade of the 2000s most of the research efforts in pattern recognition have been devoted to define novel and effective discriminant features, so as learning and classification models for maximizing the system accuracy; moreover at that time the main idea was that of proposing different recognition systems for the various problems to be solved: so, for example, many researchers approached the face recognition separately from face detection or gender recognition separately from ethnicity or race recognition. In the last decade, the technological progress, the availability of a large amount of data and the discovery of novel promising methodologies have jointly laid the basis to the deep learning revolution. The availability of powerful GPUs for massive parallel processing and the huge amount of available memory have significantly contributed to increase the number of neurons of a networks of three orders of magnitude. This stimulated the use of very large datasets, made of millions of samples that revealed to be a panacea for a very effective training of these deep nets; a never imagined performance has been so achieved and we assist everyday to remarkable results in almost all the applications using deep networks and huge datasets.

A further boost in performance was brought by the adoption of the Rectified Linear Unit (ReLU), a novel activation function, used in the neurons, able to adequately manage the numerical problem of the vanishing gradients and a consequent accurate convergence of the gradient descent algorithm. Then, the definition of novel heterogeneous layers, as the convolutional layers, used as feature detectors, and the pooling layers, for obtaining spatial/temporal invariance, created a new perspective; the process of defining handcrafted features has been progressively substituted by a much more effective method for automatically obtaining highly representative features starting from raw images.

The surprising results are everyday encouraging the scientific community to focus the attention on the design of effective deep architectures for facing complex problems, and more recently to design networks able to solve more problems at the same time; it is a consolidated conviction that this modern paradigm has a overwhelming superiority with respect to the classical machine learning approaches.

Among all the machine learning problems, the one that attracted a large part of the scientific community in the last two decades, is the analysis of facial images. The motivations behind that are numerous. A significant contribution was given by the progresses done at the beginning of this millennium as for the face detection, in particular with the method by Viola and Jones [1], that significantly improved the state of the art at that time. It is a common opinion that the availability of a method able to reliably find faces in images, has been the trigger for more advanced analyses of the detected faces. Person identification, recognition of gender, ethnicity, expression and age are examples of challenging problems dealing with face analysis. A further stimulus has been given by the several benchmarking initiatives [2], [3] and [4], that gave the opportunity to many researchers to design methods improving the existing ones and to compare their results on common data.

The growth of this research area determined a large amount of papers, surveyed in several manuscripts which analyse the solutions available in the literature for the various face analysis problems; in Table 1 we report the surveys on facial image analysis published in the last decade, with the indication of the specific treated problem.

Surprisingly, the only face analysis problem not yet surveyed in the era of deep learning is the age estimation that, among the others, is definitely the most challenging even for humans. In fact, the most complete review dates back to 2010, with the survey on age synthesis and age estimation by Fu et al. [5]. Recently, some papers reviewing

- *The Authors are at the Department of Computer and Electrical Engineering and Applied Mathematics, University of Salerno, Italy. E-mail: {vcarletti, agreco, pergen, mvento}@unisa.it*

Acknowledgements: The author warmly thank Prof. Pasquale Foggia for his useful comments and contributions at the discussions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1: Surveys regarding the analysis of facial images published in the last decade.

Ref.	Survey	Topic
[5]	Fu et al. 2010	Age estimation
[11]	Fu et al. 2014	Ethnicity recognition
[12]	Zafeiriou et al. 2015	Face detection
[13]	Ng et al. 2015	Gender recognition
[14]	Sarivanidi et al. 2015	Affect analysis
[6]	Dhimar et al. 2016	Age estimation
[9]	Dantcheva et al. 2016	Soft biometrics
[8]	Shu et al. 2016	Age progression
[15]	Ding et al. 2016	Face recognition
[16]	Wu et al. 2017	Facial landmarks detection
[10]	Sun et al. 2018	Demographic analysis
[17]	Li et al. 2018	Facial expression recognition
[7]	Osman et al. 2018	Age estimation

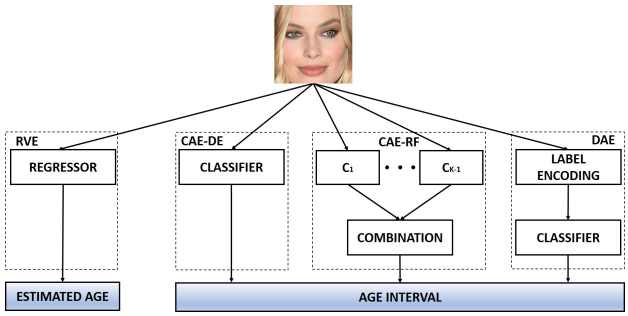


Fig. 1: Problem formulations. RVE uses a regressor providing the age. CAE gives an age interval using a single (CAE-DE) or many binary classifiers (CAE-RF). DAE provides the label distribution, centered on the estimated age.

age related tasks applied to face analysis, such as age estimation [6], [7], age progression [8], soft biometrics [9] and demographic analysis [10], have not considered the huge revolution produced by this paradigm.

Contributions: We provide a detailed analysis of the main advances in the last six years as for real age estimation (RAE), apparent age estimation (AAE) and age group classification (AGC), focusing on the deep methods. In particular, we analyse about fifty papers under different points of view, commenting the different design choices and the performance obtained so far. We put emphasis on the following fundamental aspects: the *formulation of the problem*, the *typology of the deep network*, the *ensembling* of multiple networks, if used, and the relative *learning procedures*. The latter regards the optimization of the learning process, by using suited techniques as: *preprocessing*, *data augmentation*, *training* and *learning using other soft biometrics*.

We believe that the analysis of the recent approaches, together with a review of the available datasets and a discussion of the achieved results, is useful for both researchers and practitioners interested in the deep learning paradigm for facing the age analysis problem in real applications.

Organization of the paper: Section 2 presents the methods, while Section 3 recalls the commonly used performance indices together with some useful considerations about their adoption. Section 4 presents the datasets used for performance assessment; each one is described in terms of size, age range, face image acquisition protocol and suitability for specific formulations of the AE problem. In Section 5, we analyse the results of the methods facing the three main face analysis problems: real age estimation, apparent age estimation and age group classification. Finally, in Section 5.3 we summarize the results of our analysis, present some open issues and delineate possible future research directions.

2 METHODS

In this section we review the methods for age estimation based on deep learning and presented in the last six years.

In particular, in Section 2.1) we discuss in details the different formalizations of the problem adopted in literature. Together with the latter other two relevant aspects have been analysed: the adopted typology of the deep network (Section 2.2) and the possible use of a plurality of networks, suitably ensembled each other (Section 2.3). Finally, Section 2.4 provides instead insights about the main aspects regarding the learning methodologies, namely the preprocessing, the data augmentation, the training and the use of other soft biometrics. The main information for each aspect are given into a specific column of the Table 2.

For completeness, we include a section 2.5 dedicated to the methods not based on deep learning; to this aim, we consider a selection of the recent methods achieving performance comparable to the ones of the deep systems.

2.1 Problem formulation

As for many pattern recognition problems, age estimation (AE) can be approached by using either a classifier or a regressor. The question "which formulation of the problem is the more suited to maximize the system performance?" is still under discussion, as reported in [48] and [50]. Surely, the simplest and most natural way for approaching the problem is to design a regressor able to provide an age label falling in a predefined range, f.i. [0,100]; this solution, known as **Real-Value Age Encoding (RVE)**, while makes the problem definition intuitive, on the other hand suffers of the fact that it is rather utopian to design a system able to estimate the age with the precision of a single year. So, the performance evaluation of these systems becomes a very complex task, as it should be taken into account that an age error of one year cannot be considered as equivalent to an error of ten or more years; consequently, the regressor should be integrated with another subsystem devoted to manage this matter creating some sort of classes of equivalence of the errors.

It has also to be considered that, as discussed in [48] and [50], the AE problem intrinsically defines an *order relationship* between the age labels, and that should be adequately taken into account; for instance, given a true age label equal to 40, the error of the system should be considered higher if the estimated age value is 30 instead of 41, being in both cases the estimated age label wrong.

Indeed an alternative formulation is based on the definition of age classes, i.e. groups of ages considered equivalent for practical applicative purposes; this approach, called **Classification Age Encoding (CAE)** turns the AE problem to a classification paradigm with a number of classes equal to the number of the considered age groups. The advantage of this approach is that the width of the age groups can be directly tailored at the input problem requirements; so, if the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 2: Classification of the methods on the basis of problem formulation, ensembling, used deep network and learning procedure (preprocessing, data augmentation, training and multi-task).

Ref.	Problem formulation	Deep network	Ensembling	Learning procedure			
				Preprocessing	Augmentation	Training	Multi-task
[18]	CAE-RF	ScatNet	No	Not specified	No	From scratch	MTL
[19]	RVE	CNN	No	Alignment	Yes	From scratch	MTL
[20]	RVE and DAE	GoogLeNet	E-DEC	Alignment	No	FR	No
[21]	CAE-DE	VGG-16	E-DEC	Best score	No	AE	No
[22]	RVE	GoogLeNet	E-DEC	Alignment	Yes	FR	No
[23]	RVE	VGG-16	E-DEC	Alignment	No	GT	MTL
[24]	DAE	VGG-16	E-DEC	Alignment	No	GT or AE	No
[25]	RVE	CNN	No	Not specified	No	From scratch	No
[26]	RVE	CNN	E-DES	Alignment	Yes	From scratch	No
[27]	RVE	LeNet	No	Alignment	No	From scratch	No
[28]	CAE-DE	GilNet	E-DEC	Alignment	Yes	From scratch	No
[29]	RVE	CNN	E-DEC	Alignment	No	FR	No
[30]	CAE-DE	VGG-16	E-DEC	Best score	No	AE	No
[31]	RVE	CNN	E-DEC	Alignment	No	FR	No
[32]	RVE	VGG-Face	E-DEC	Best score	No	FR	No
[33]	CAE-RF	CNN	No	Alignment	Yes	From scratch	No
[34]	RVE	VGG-16	No	Best score	No	GT	No
[35]	RVE	CNN	No	Alignment	Yes	AE	No
[36]	CAE-DE	VGG-Face, ResNet	No	Not specified	No	GT	No
[37]	CAE-DE	LeNet	E-DEC	Alignment	No	FR	No
[38]	CAE-DE	VGG-Face, AlexNet	No	Not specified	No	FR or GT	No
[39]	DAE and CAE-DE	VGG-16	E-DEC	Best score	Yes	AE	No
[40]	DAE	VGG-Face	E-DEC	Alignment	Yes	FR	No
[41]	CAE-DE	VGG-Face	E-DEC	Not specified	No	FR	No
[42]	CAE-DE	VGG-Face	E-DEC	Alignment	Yes	FR	No
[43]	CAE-DE	CNN	No	Alignment	No	From scratch	No
[44]	CAE-DE	VGG-16	No	Alignment	Yes	AE	No
[45]	RVE	Hyperface	No	Alignment	No	FR	MTL
[46]	CAE-RF	VGG-Face	No	Alignment	No	FR	No
[47]	CAE-RF	VGG-Face	E-DES	Alignment	No	FR	No
[48]	DAE	VGG-Face	No	Alignment	No	FR	No
[49]	DAE	GoogLeNet	No	Not specified	No	GT	No
[50]	RVE, CAE-DE, DAE	AlexNet, VGG-16	E-DEC	Alignment	No	From scratch	MTL
[51]	RVE	AlexNet	No	Alignment	No	From scratch	No
[52]	Not specified	CNN	No	Alignment	No	FR	No
[53]	RVE	VGG-16	E-DEC	Not specified	Yes	AE	No
[54]	CAE-DE	RoR	No	Not specified	No	GT and AE	BTL
[55]	CAE-DE	VGG-Face	No	Not specified	No	FR	No
[56]	RVE	Alexnet	No	Alignment	No	GT	MTL
[57]	RVE	CNN, VGG-16	E-DEC	Alignment	No	AE	BTL
[58]	CAE-RF	ResNet-101	No	Alignment	No	FR	No
[59]	RVE	Alexnet	No	Alignment	No	AE	MTL
[60]	RVE	CNN	E-DES	Best score	No	GT and AE	BTL
[61]	CAE-RF	ScatNet	E-DES	Not specified	No	AE	MTL
[62]	RVE	VGG-16	No	Alignment	Yes	FR	MTL
[63]	RVE	VGG-16	E-DEC	Alignment	Yes	AE	No

application requires to distinguish only three classes (young, middle age, elder) the classifier becomes very simple, while, in the most general case, the complexity of the classifier is strictly dependent on the requirements, given in terms of number of age classes.

These two formulations have been widely adopted by many methods in the literature; a rather deep comparison is reported in Antipov et al. [48] with the conclusion that CAE, under a plurality of aspects, is more effective than the others; on the other hand, Xing et al. [50] apply five different loss functions and conclude that the RVE achieves the best results when combined with the mean absolute error (MAE).

Independently on the use of a regressor or a classifier, the problem formulation should consider the metric for evaluating the error and in particular the nature of the order relationship on the solution space; to this concern, an effective approach has been proposed in [64] and is known as **Distribution Age Encoding (DAE)**. It is a modification of the CAE strategy obtained by substituting the one-hot encoding vector with a statistical distribution centered on the estimated age. The rationale lies in considering the neighbourhood of the different age classes; so, given a generic age label L, the interval of the ages adjacent to L is represented as a distribution where L is the average value.

Obviously, the methods adopting this paradigm may differ as for the used label distribution, even if the Gaussian one is the most adopted; in this case, being y_i the true age of the i-th sample, then the age label l_j for the j-th output neuron is expressed as follows:

$$l_j = \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{(j-L)^2}{2\sigma_j^2}} \quad (1)$$

being σ_j the standard deviation of l_j.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

Once introduced these problem formulations, the papers can adequately be categorized according to the taxonomy proposed in [48], as shown in Fig. 1.

Papers based on Real-Value Age Encoding (RVE) The methods in this category differ either for the specific loss functions or the regressive scheme.

As for the first aspect, the papers [19], [20], [22], [25], [35], [56] and [59] use the Euclidean loss function; differently, the Gaussian loss function is adopted in [29], [31], [45] and [63]. In particular, [29] uses a single stage optimizer, while [31] proposes an optimizer followed by an error correction stage; [45] combines the Gaussian and the Euclidean loss functions through a weighted sum. Other researchers propose their own loss functions [62] based on the L1 norm [51] or using an adaptive activation function [53]. Among all the loss functions, the experiments demonstrate (see Section 5) the effectiveness of the Gaussian one over the others.

Passing to the regressive schemes, the papers [23], [32], [34], [45], [57], [63] and [60] propose different solutions: an Extreme Learning Machine technique [65] is used in [32] and [60], a Support Vector Regressor [66] in [34], a regression output after two fully connected layers in [45], a Gaussian Process regression [67] in [57] and [63], while Kuang et al. in [23] combine two regression schemes, namely the Quadratic Regression with a Local Adjustment [68] and the Random Forest [69]. Finally, a logistic regression scheme is adopted in [26], while a regularized Canonical Correlation Analysis (CCA) regressor [70] is used in [27].

The experimental results reported in Section 5 highlight that among the many methods adopting this paradigm, only a few ones are able to achieve top performance on the available datasets [45], [63]; such approaches share the use of the Gaussian loss function, although using different regressors. So, the experimental evidence allows us to support the hypothesis that the success of these methods is much more depending on a careful selection of the loss function rather than to the adoption of a given regression paradigm.

Papers based on Classification Age Encoding (CAE): we include here the methods based on a classification paradigm; they require the introduction of age intervals, obviously non overlapped, each one representing a class.

We divide the methods falling in this category in two further subcategories, namely the CAE-DE (Direct Encoding) and the CAE-RF (Reduction Framework); as clarified in the following, while the former ones use a classifier for each age interval, the latter introduces a system architecture made of an aggregation of multiple binary classifiers.

The many methods in the CAE-DE class treat the age interval as separate and mutually exclusive classes; so, the age label is encoded by means of a one-hot vector, i.e. a vector having all zeros as component, except for the component i-th, associated to the i-th age interval, that assumes the value one. A significant aspect is the use of fixed age intervals instead of variable ones; it is noteworthy the use of 101 binary outputs, each encoding exactly one year, giving rise to the problem named *age estimation*.

Except for [38] and [41], in which the classification is performed through a SVM fed with the features produced by the hidden layers of the CNN, all the other approaches perform the classification directly with the output layers of the deep network; typically the latter are optimized by using a cross-entropy [21], [30], [44] or a softmax [42], [55] loss function. Dong et al. [37] complement the cross-entropy with a penalization term that accounts for the distance between the true age group and the guessed one; similarly Zhang et al. [54] weight the errors in dependence of the specific age group. Finally, Hou et al. [36] use the Earth Mover's Distance EMD^2, while in [28] and [43] a classic stochastic gradient descent is adopted.

The CAE-RF methods use an aggregation of multiple binary classifiers arranged according to the reduction framework proposed for facing ordinal ranking problems [71]. More specifically, once we have defined the K age intervals (or equivalently ranks), these systems adopt $K-1$ binary classifiers, as shown in Fig. 1; the generic k-th classifier is devoted to decide whether the rank of the age group associated to the input sample x_i is higher than k or not.

Under this assumption, given a face image with true age falling into the k-th rank, the expected label vector, having length $K-1$, has the first $k-1$ values equal to 1 and the remaining equal -1. So, the age rank of the input face x_i is predicted as:

$$r\left(x_i\right) = 1 + \sum_{k=1}^{K-1} [\![O_k\left(x_i > 0\right)]\!] \tag{2}$$

being $O_k\left(x_i\right)$ the output of the k-th classifier, and $[\![O_k\left(x_i > 0\right)]\!]$ equal to 1 if the k-th classifier classify x_i as belonging to the k-th rank or equal to 0 otherwise.

This framework is introduced in [18], [33] and [61]; these papers define age ranks of one year and use the cross-entropy as loss function. In addition, Yang et al. in [18] generalize the method so as to include further information as gender, race and expression, while Liu et al. in [46], [47] and [58] combine the cross-entropy with a square or a softmax loss function to optimize the classifier.

The CAE paradigm, adopted by several methods, generally achieves better performance with respect to the ones based on the regressive paradigm, as we will show in Section 5. Moreover, this success does not seem related to the specific CAE formulation, since remarkable results are achieved adopting both the formulation CAE-DE [30], [44] and CAE-RF [46], [49], [58].

So, the adoption of this paradigm, although less immediate with respect to a regressive method, is in practice paid back by better performance; such experimental finding explains the main reason why the majority of the recently proposed successful methods adopts this formulation.

Papers based on Distribution Age Encoding (DAE):

The DAE paradigm has been successfully used in [24], [40], [48] and [49]. Some papers present schemes combining different paradigms and results. This is the case of [20] that proposes a method which combines the output of two subsystems, one based on the model DAE and the other on RVE; the final age estimation is obtained by averaging the output given by the two estimations. Another proposal comes from Antipov et al. in [39] that use a DAE system for estimating the age in presence of an adult, and a further CAE-DE system for the estimation of a child.

From the applicative point of view, DAE requires higher efforts to collect labelled training data with respect to RVE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

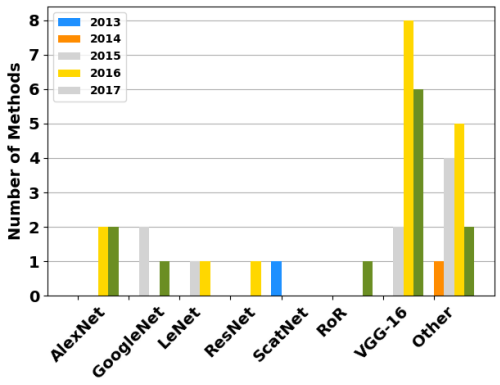


Fig. 2: Number of methods based on deep networks.

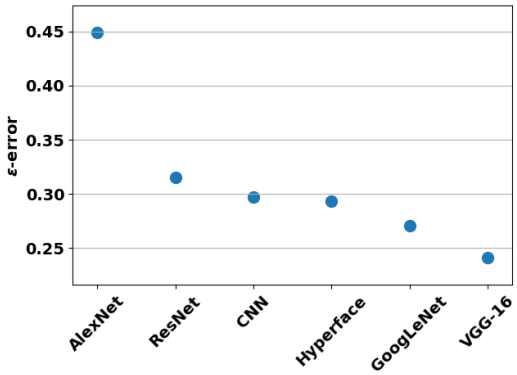


Fig. 3: Performance, in terms of ϵ-error (see Section 3), achieved by the different deep network architectures on the ChaLearn LAP databases.

and CAE due to the fact that, in this case, it is further required the distribution for each age label; anyhow, although more data are given in input to the training process, it does not significantly increase the needed training efforts in terms of time and learning cycles. The results achieved by [39], winner of the competition ChaLearn LAP 2016, and the analysis in [48] suggest that DAE is the most effective paradigm for AE; notwithstanding, the experimental analysis in Section 5 does not confirm this evidence. So, at the state of the knowledge, a definite conclusion about the best problem formulation is far to be reached and future investigations are desirable; this would surely be an interesting research direction.

2.2 Deep Network Design

As in the large majority of the application fields where deep learning methodologies have been used, also for AE there is a strong evidence that the achievement of better performance is strictly associated to the depth of the used network; this in turn implies the use of wider and wider datasets for an adequate training. Of course, the computational resources required during the training significantly increase and, as a matter of fact, only a few research laboratories in the world have enough computational power for training a new network from scratch in an acceptable time.

In this case, it is generally used the so called *fine tuning procedure* of an already trained net; it is based on the replacement of the final layers of an already trained net with others, specifically trained for solving the problem at hand; the front layers of the net, devoted to feature extraction remain the same. As shown in Fig. 2, we may recognize for the problem of age estimation the same evolution verified as for other applications, such as object recognition; in particular, the first methods used relatively small networks, usually trained from scratch on a relatively small dataset. This is the case of LeNet [72], the network that inspired many of the successive approaches; it is made of three convolutional layers followed by one or two fully connected.

Over time, such methods have progressively standardized the generalist network adopted for feature extraction; in particular, we have attended the affirmation of deep architectures such as AlexNet [73], VGG-16 [74], GoogLeNet [75] and Residual Networks [76], all of them already trained on huge and generalist image datasets. Among the other things, these nets revealed to be so effective that nowadays they are provided as off-the-shelf tools in most of the machine learning frameworks. This is particularly evident for VGG-16 and its specialized version VGG-Face [77], currently soaring as the most adopted architecture.

Furthermore, it is worth mentioning some other hybrid architectures tailored for increasing the effectiveness of the convolutional layers of a CNN, for instance by adding or substituting fully connected layers with other types of classifiers. Examples are given by the methods proposed in [18] and in [61] where the description is enriched with the features provided by a scattering network (ScatNet) [78]. Notable examples of the second case are given by Gurpinar et al. in [32], successively extended by Duan et al. in [60]; the underlying idea is the use of either Extreme Learning Machines (ELM) on top of a series of convolutional layers for age estimation or SVM [79], as in [22], [34], [41] and [25].

Among the wide variety of network architectures, VGG-16 is the one achieving the best performance on almost all the available datasets for age analysis, as witnessed by [44], [48] and [30] and shown in Fig. 3; this observation is also supported by the performance comparisons reported in various papers as [36], [38], [48], [50], [80] and [57], which theorize the superiority of this network, and by the experimental analysis that we report in Section 5. We believe that a relevant part of the success of this network is mostly due to VGG-Face, which benefits of a pre-training process on almost $1,000,000$ images for person identification, so allowing an effective transfer learning for AE. All these evidences, let us to be very confident in considering VGG-16 as the most effective deep learning architecture for age estimation among those currently available. We do not exclude that, in the future, similar results may be obtained by different novel network architectures as MobileNet [81], NasNet [82] and other networks reported in [83]; thus, we encourage the research in this direction.

2.3 Ensembling of Deep Classifiers

The ensembling, also called multi-classification, is a classification paradigm based on the use of a plurality of classifiers, suitably connected (ensembled); the outputs of the classifiers are combined at a certain level of the decision process, so as to obtain the final decision. The papers [84], [85],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

[86] and [87] describe the rationale behind this paradigm and give details about the rules to be used for combining at the best the decisions of any classifier. Many papers have demonstrated that an ensemble of classifiers allows the system to significantly improve the performance under the condition that the individual classifiers are, at least partially, complementary; the latter situation can be obtained by combining classifiers using different and uncorrelated features [88], different classification models [89] or classifiers trained on different datasets or different portions of the same dataset (bagging) [90]. The impressive results obtained by the use of a multi-classifier in many pattern recognition problems boosted it also for the AE.

The approaches specifically proposed for the problem at hand can be ascribed to two main classes, on the basis of the scheme adopted for ensembling the different deep networks constituting the system. In particular, we have methods implementing the so called ensembling at description level (E-DES), that are based on a plurality of deep networks made of convolutional layers, trained and working as feature extractors; their output converge in a final output layer that, as a matter of fact, works as a classifier with a feature vector obtained by arranging the features provided by the previous nets, see Fig. 4a. Another model is inspired by the idea that autonomous networks are used in parallel for obtaining an estimation of the age; the combination operates on the single decisions of these classifiers and provides its final decision, in most cases much more reliable, as shown in Fig. 4b in the particular case of three deep networks.

Ensembling at description level (E-DES). Examples of this paradigm are the methods presented in [26], [47], [61] and [60]. Liu et al. [47] that combine the description obtained by the well known VGG-16 with those obtained by two shallower networks fed with images obtained by reducing the resolution of the original input images. Another method which uses information at different resolution scales is presented in [26]; in this case the features are preliminarily extracted using small patches of the input image in parallel and then progressively aggregated on larger parts of the image. Yang et al. [61] use a concatenation of the features

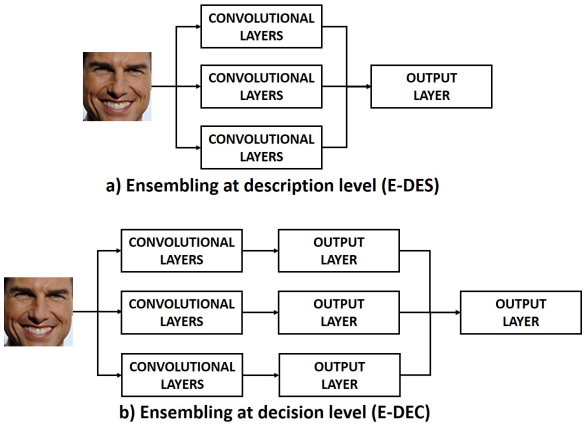


Fig. 4: Ensembling techniques. (a) At description level (E-DES): the output layer combines the features learnt by different nets. (b) At decision level (E-DEC): the output layer combines the decisions taken by different networks.

produced by a ScatNet [78] and the ones selected by the convolutional layers of another CNN. In [60] the combination is based on the features calculated by three different CNNs independently trained for age estimation, race and gender recognition. Although the ensembling at description level is very interesting from a theoretical point of view, it has not produced significant performance improvements with respect to the state of the art; on the other hand, as the combination at decision level revealed to be more effective, most of the ensembling solutions are based on this model.

Ensembling at decision level (E-DEC). Although many are the possibilities for designing an ensembling at decision level, we may mainly recognize two emerging trends: one based on the use of different networks, all of the same type, but trained on different data and another group, for sure more general, involving different network architectures.

Examples of methods belonging to the first group are the ensemble of four VGG-Face networks described in [40], where each network is independently fine-tuned for age estimation on a different dataset, and the one proposed by Malli et al. in [42], which trains several networks with the same architecture on samples with shifted age groups, combining the different outputs in the final fusion stage. We can ascribe to the same category the papers [21], [22], [30] and [41], where different classifiers are trained on different portions of the same dataset, or [28], [39], [53] and [37], in which the networks are fed with modifications of a same image, as flipping, cropping, rotations and scaling.

The approaches in [20], [23] and [24] belong to the second group. In particular, Liu et al. [20] use two different estimators, in particular, a regressor and a classifier with a resolution of one year, by taking the average of the two estimations for determining the age. Kuang et al. [23] estimate the age by averaging the output of two different regressors, a Random Forests and a Quadratic Regression with Local Adjustment, used as last stage of a deep network. Yang et al. [24] use a combination of classifiers based on two different architectures, VGG-16 and a shallower one, and introduce a priority between the two: in case there is a large disagreement between their estimations, the output of the whole system coincides with the output of the VGG-16.

It is worth pointing out that the combination at decision level may sometimes involve more complex schemes; indeed, some methods in this category are based on a series of classifiers arranged in a hierarchical structure. In this case, each classifier is selectively activated depending on the decision taken by the preceding classifier. The most typical approach of this group is adopted in [31] and [32] and produces the decision in two steps: in the former, a classifier provides the age group classification; in the latter, the final age estimation is done by a classifier specialized on the age group recognized by the first stage. Differently, Wan et al. [57] propose an even more articulated architecture that produces the final age estimation using a branch of the network specialized on the gender and the ethnicity recognized in the first levels of the hierarchy. Finally, [63] combines the decisions of classifiers trained with features coming from different intermediate layers of VGG-16.

The effectiveness of the ensembling techniques for AE has been widely experimentally demonstrated; although there are many techniques and variants for designing a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

multi-classification system, the availability of generalist, already trained, deep networks for face analysis is determining the rapid increasing of approaches, mainly based on the combination at decision level of these networks, with tuning procedures on portions of the dataset or in other ways. This is the reason why we are assisting to a progressive affirmation of ensembling methods using VGG-16 and/or Residual Networks; this trend is also supported by the fact that the ensembling according to these criteria allows the designer to obtain impressive results, without being involved in very huge processes of training from scratch.

Since the most successful networks [30], [39], [57] rely on classifier ensembling at decision level, there is evidence for considering this approach as the most promising one as for AE, differently from the fusion at description level that experimentally appears to be less effective [26], [47], [60], [61]. This aspect clearly comes out also from the results in Section 5. Moreover, among the E-DEC methods, the approaches adopting data-level combination are typically more effective than the ones using ensembles of different network architectures. In particular, it is worth noting that the method achieving the top-1 performance on the Adience dataset [53] and the winner of the ChaLearn LAP 2016 competition [39], are both ensembles of networks trained with images flipped, cropped, rotated and scaled; we deem that this design choice, strictly related to the data augmentation that we will treat in the following, has been essential for obtaining the best performance on the most challenging datasets currently available in the literature.

2.4 Learning methodologies

The achievement of good performance is heavily dependent on the procedures used for training the deep network, as already demonstrated by thousands of experimental data collected in the different application domains; it has to be clarified that the learning methodologies should not be considered as limited to the specific learning algorithm or to the values assigned to the meta-parameters adopted during the learning process but many other aspects are sometimes much more relevant.

For instance, the collection of the training set is crucial: data require to be preprocessed before their insertion in the training set and more frequently they are used for augmenting their representativeness (data augmentation). The methods for using the data during the training process require an adequate awareness; multi-task learning is a recent trend, attracting much attention as for the possibility of solving simultaneously a plurality of problems; as we will see in the following, when the problems are correlated, this may allow to obtain better performance by exploiting their interdependences. In the next we will consider these aspects, by presenting the solutions proposed so far.

2.4.1 Preprocessing

Preprocessing techniques are used for making the input data suited for training a neural network; we include in this category all the methods adopted for face detection, aimed at localizing and cropping the faces in the image, so as the algorithms for pose normalization, necessary for reducing the variability caused by different poses of the subjects.

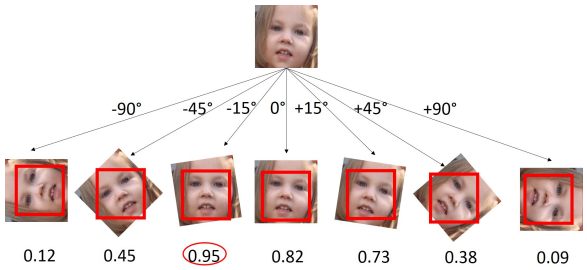


Fig. 5: Pose normalization by best score: the original image is rotated with six predefined angles and the detector applied; the face is normalized by using the angle with the best score.

Face detection. Except when the dataset already provides cropped face images, face detection is a mandatory step to localize the position of the face in the input image; to this aim, the use of a particular face detector is crucial, as any error at this level (false and missed detections) cannot be recovered, so having an heavy impact on the performance.

In recent years several face detector have been proposed [12]; it is worth mentioning the ones described in [91], [92], [93], [94], [95], [96], [97], [98], [99], [100] and [101] for their remarkable performance. Notwithstanding, the most used face detectors are the Viola-Jones algorithm [1], the Coc-DPM [102] and the "Head Hunter" [102]. Their success is mainly due, not only to the accuracy, but to their computational effectiveness, allowing to perform the detection of a plurality of faces in an image in real-time on standard CPUs.

The well known Viola-Jones algorithm [1] is a very fast detector based on the Haar features suitably selected by using a cascade of classifiers that effectively work by promptly stopping the search of faces when analysing unpromising regions. The results reported in [12] demonstrate that more recent methods outperform the Viola-Jones algorithm, as for the accuracy of the detection jointly with a consistent time saving [102]. The CoC-DPM method takes advantage of a weakly supervised deformable part model (DPM) [103] improved with a finer non-maxima suppression which allows to reduce to number of false positives. The "Head-Hunter" is very similar to the Viola-Jones algorithm, as it is based on the idea of using a cascade of (twenty two) classifiers acting as face detectors at different scales and orientations, even if it uses the Integral Channel Features (ICF) [104] applied to Histogram of Gradients (HOG) [105] and to the LUV color channels. This approach guarantees a very good trade-off between accuracy and processing speed, being able to run on 640×480 images at 50 fps.

Finally, it is worth citing some recently proposed detectors based on deep paradigms, as [106], [107], [108], [109] and [110]. Although they revealed to be particularly effective, their use is still limited; it is probably due to the fact that they are time consuming with respect to other available approaches and the accuracy improvement does not adequately compensate the heavy computational overhead.

Pose normalization. The most important normalization applied to faces before feeding them to a network is the adjustment of pose variations; in fact, different orientations of the input faces, even the smaller ones, negatively affect the system performance. The methods for pose normalization

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

here surveyed can be divided in two groups, depending on the fact that they leverage or not facial landmarks.

In the first group we number the methods using facial landmarks [16] for horizontally aligning the face and for computing an affine transformation which puts eyes, nose and mouth into canonical coordinates [20], [22], [23], [24], [27], [29], [31], [33], [35], [37], [40], [42], [43], [44], [46], [47], [57], [58], [62], [111] and [63].

The above mentioned methods require a preliminary facial landmark detection; a comprehensive survey presenting the different methods is [16], very useful for the reader to select the most suited algorithms and the relative expected performance. It is important to say that the most used tool for effectively detecting the fiducial points in a few milliseconds on traditional workstations, is dlib [112]: a cascaded regression method, similar to the Supervised Descent Model (SDM) [113], that, starting from an initialization of the positions of the landmarks, gradually corrects them by using a regression function fitting a pre-learned model.

The methods for detecting landmarks based on deep networks, as CNN, have been widely used in recent years; among them, Hyperface [114], CFAN [115] and the method by Sun et al. [116] are known for the accuracy achieved on the standard benchmarks [16]; they represent currently a very effective solution when the computational load is not a problem.

Finally, Active Shape Model (ASM) [117] and Active Appearance Model (AAM) [118], which model the shape and the global appearance of the face, are still used although they have been the first techniques proposed for facial landmarks detection [16], now dated.

As for the approaches belonging to the second category, namely [21], [30], [32], [34], [39] and [60], their rationale is to apply rotations of the input image and to use the score provided by a face detector (e.g. [102]) as a measure of the quality of the result, i.e. the pose of the face after the rotation. As shown in Fig. 5, once calculated the scores corresponding to any imposed rotation, the original face image is rotated using the angle corresponding to the highest obtained detection score. The methods working according to this rationale differ for the range of considered angles (e.g. [-60°,60°], [-90°,90°]), the step (in most cases 5 degrees) and the portion of image used to extend the face width (to the left and to the right) and height (above and below) .

Finally, for both the categories, the normalization ends with a rescaling of the resulting image to a given resolution; in some cases, an intensity normalization is performed for reducing the brightness variability as in [35] and [77].

In conclusion, it has been demonstrated in [15] that face detection and normalization techniques make the training procedure more effective. For this reason, they can be considered as mandatory preliminary steps of any method for AE; there is no face detector that stands over the others with respect to the achieved accuracy; on the contrary, from the experimental analysis in Section 5 we observe that the use of fiducial points for face normalization allows to achieve higher accuracy with respect to the ones based on the best score, mainly thanks to very effective algorithms for canonical alignment.

2.4.2 Data augmentation

Data augmentation is a technique that allows to significantly extend the size of the training set with the introduction of artificial samples obtained by applying transformations of the images in the original dataset; the rationale behind the data augmentation lies in the fact that, according to the curse of dimensionality, as the size of the network increases, a proportional amount of training data is necessary to achieve good performance and adequate generalization ability [73]. It has to take into account that the size of the network is in turn related to the complexity of the considered problem, so being imposed by the application domain. A further reason why a training set needs to be augmented subsists when the available data are not adequate to effectively represent real situations; an example is when we have a dataset of faces all in the perfect pose: even if the number of faces could be potentially sufficient to have an effective training, in presence of real situations we may have wrong results, for instance when an input face is rotated, even slightly. It happens because the network has learnt only perfectly posed faces. Analogous considerations may be done by considering other variable conditions of the input data, related to noise, resolution, illumination, contrast, and so on. Consequently, data are augmented by adding noise or, more importantly, different orientations of faces, position in the image, resolution scale and so on; the expected result is a deep network much more insensitive to real conditions.

Random cropping is an easy way to augment the data for making the system insensitive to relative variations of the position and size of the bounding box, obtained by the detection stage, with respect to the face. It is based on the selection of random patches from the original faces, as proposed in [28], [33] and [42]. These two methods, although inspired by the same idea, differ each other for the size and the scale of the considered random patches. Other approaches may also include different transformations as translation, scaling, flipping, rotation, sharpening, random brightness changes and addition noise, either salt and pepper or Gaussian one; in this way, the augmentation gives rise to effective representative datasets, as described in [22], [35], [39], [44], [46], [53], [57], [62] and [63].

The success obtained by the method proposed in [44], which achieves state of the art accuracy on most of the available datasets, is surely due also to the enrichment, by data augmentation, of the training set with variations (orientation, location, scale, brightness, noise) [119] that occur in real world. This result can be considered as the proof that data augmentation on normalized samples significantly reduces the sensibility of the resulting network to translation, viewpoint, size and illumination. Although there is a strong experimental evidence that data augmentation may significantly increase the accuracy and the generalization capability of the systems, only a small percentage of the existing methods, around 20%, adopts this important technique; in spite of this and according to the experimental findings, we strongly recommend the use of data augmentation.

2.4.3 Training

From 2013 to 2016 most of the authors used to train their CNNs from scratch, using the available datasets [18], [19],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

[25], [26], [27], [28], [33] and [43]. Of course, training from scratch a deep network is not trivial, because of the size of the datasets; the process may require weeks of computation even using powerful systems with GPU accelerators.

Whether the computational resources are limited and the number of available samples is not sufficiently high (see for example [50] and [51]), the usual training procedure can be avoided by reusing the intermediate-level features of a pre-trained network; indeed, the first hidden layers of the network learn low-level features, as simple relations of the raw image pixels with their neighbourhood (e.g. points, edges, contours), while moving closer to the output layers the net learns more complex structures, up to the features that better describe the objects of interest.

This property allows the reuse of the features across similar applications, since different problems can be addressed by only replacing the output layers with new fully connected layers or other standard classification techniques (SVM, Random Forests and so on); of course, only these layers must be finally trained to face the new problem. Such a technique, known as *transfer learning* or *fine tuning*, is gaining popularity for two main reasons: (i) the possibility to use smaller datasets, being the features already provided by the front layers of the pre-trained network; (ii) the time saved during the training step, since a smaller number of weights and samples are involved in the learning procedure.

According to this paradigm very large datasets need to be used only for training networks devoted to solve general problems (i.e. object recognition and face recognition); the obtained networks are then fine tuned for the specific task, in our case age estimation, on the available smaller datasets. This approach has been studied in [48], where the authors experimentally evaluate the importance of two different pre-training strategies, namely the *general task (GT)* and the *face recognition (FR) pre-training*; in addition, we discuss here a further recent strategy which makes use of age annotations, the so called age estimation (AE) pre-training. In the following we describe the rationale of these techniques.

GT pre-training. The general task pre-training is a typical solution adopted for transfer learning. The deep network is pre-trained for solving the general problem of image classification with a large number of classes, typically ImageNet [120]. Examples of networks pre-trained on Imagenet and fine tuned with specific age estimation databases are VGG-16 [23] [24], [34], AlexNet [36], [38], [56], GoogleNet [49], or some other proprietary CNNs as [54], [60]. As shown in [48], the features learnt by the hidden layers of these deep networks are still very general and, consequently, the performance on the problem at hand are only weakly affected by adopting the fine tuning procedure instead than a complete learning from scratch.

FR pre-training. The face recognition pre-training consists of a training phase from scratch by using suited datasets annotated with the identity of the subject. The rationale behind this choice is that the deep network learns a face representation that is more specific for face analysis than that obtained by using the GT pre-training [48]. Of course, in order to learn a good representation it is necessary to feed the network with a big and representative dataset.

VGG-Face is the most used network in the field, pre-trained with a dataset of about $1,000,000$ face images, as described in [32], [36], [38], [40], [41], [42], [46], [47] and [55]. Another deep network, Slightound Face API [52], has been even pre-trained with $4,000,000$ face images. Finally it is important to cite a popular dataset for FR pre-training is the CASIA-WebFace [121], composed by about $500,000$ face images and has been used in [20], [22], [62] to pre-train GoogleNet and in [29], [31] and other proprietary CNNs.

AE pre-training. The age estimation pre-training is a preliminary training of the network by using a suited database annotated with age labels; it is a very common technique used in the recent years to learn more specialized features for the problem at hand. The publication of the IMDB-WIKI dataset [30], which consists of more than $500,000$ face images annotated with age labels, encouraged the researchers to its use, as described in [21], [30], [39], [53], [57] and [59]; it has been also augmented for increasing its representativeness by flipping, rotating and adding Gaussian noise on the images [35], [44], [63]. The most common network architecture for this kind of training is VGG-16 [21], [30], [39], [44], [53], [63]; it has also been effectively applied to AlexNet [59] or proprietary CNNs [61].

The results of the recent methods suggest that the AE pre-training [30], [44], [48], [53], [57] and FR pre-training [20], [45], [52], [58] are significantly more effective than the GT pre-training [34], [36], [60] and the training from scratch with small datasets [50].

It is worth saying that the use of wide face datasets, annotated for face recognition (VGG-Face DS2, Casia-WebFace) and age estimation (IMDB-WIKI) and extended with data augmentation, allows to effectively learn face representations. The features learned with the GT pre-training are probably too general for face analysis and thus inadequate to guarantee good performance by transfer learning on the age datasets, as shown in [48]. Consequently, it is strongly encouraged the application of FR or AE pre-training; among these two, the latter is surely more promising, especially if in the future a dataset, significantly wider than IMDB-WIKI, will be made available.

2.4.4 Learning using other biometrics

The accuracy of an age estimation system may be also affected by many biometric factors, as gender, ethnicity and expression of the person whose face is under analysis. This dependence is experienced and is going to be treated adequately by the researchers; for instance, as for the dependency from gender, described by Zhang et al. in [54], the age estimation in real conditions achieves different performance for men and women for at least two reasons. On the one hand, the aging process of men and women is different because of the longevity, hormones, skin thickness, and so on; on the other hand, the make-up may significantly alter the perception of the true age, even for humans.

The impact of different factors on age estimation has been experimentally verified by Guo et al. in [122]; consequently, several methods have been designed so as to take advantage from these aforementioned factors. The solutions adopted by the different proposed methods can be categorized in two main groups. A first group includes the methods that still continue to provide just the estimated age as output and use the biometrics during the training procedure, for instance by applying fine tuning or transfer

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

learning with images annotated with them; hereinafter these methods are denoted as *biometric task learning* (BTL).

Viceversa, the other approaches are based on the multi-task-learning (MTL) paradigm [123], i.e. using a system that is able to solve at the same time a plurality of tasks; in the considered case, for instance, in addition to the age estimation, typical additional tasks are gender recognition, or ethnicity recognition. To this concern, the paper by Xing et al. in [50] proposes a comprehensive taxonomy of multi-task learning, and introduces two further subcategories, namely the *parallel-multi-task learning* (P-MTL) and the *deeply supervised multi-task-learning* (D-MTL).

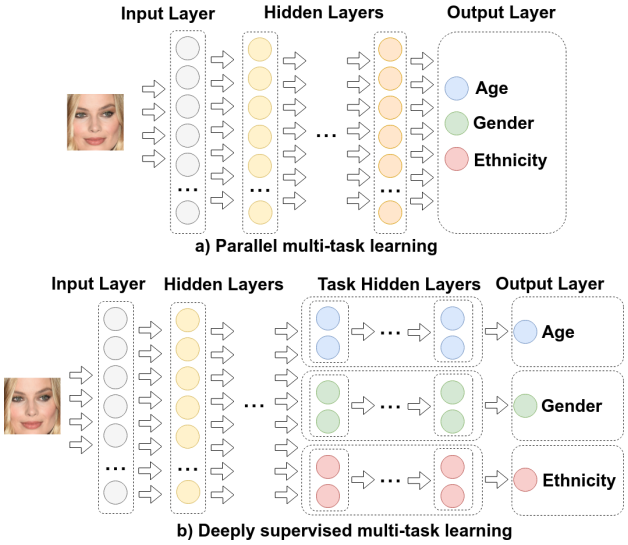


Fig. 6: (a) P-MTL: the network is trained for age estimation in parallel with nets for other biometrics; all the layers are in common and each output neuron is associated to a task. (b) D-MTL: the different tasks may share some front layers, but are solved with specific nets, working in parallel.

Methods based on BTL. In this category we mainly count the three methods [54], [60] and [57], which adopt very different approaches for taking into account the further biometric factors. In particular, Zhang et al. propose a system for age group classification presented in [54]; the authors adopt a transfer learning methodology by initially training the network using images labelled with gender data and successively apply a fine tuning with age data. They experimentally show that this approach allows to improve the accuracy with respect to the case when gender data are not exploited. Duan et al. in [60] propose to use a global descriptor that includes both the age descriptor and those adopted for gender and race classification: this is practically done by firstly training three independent classifiers respectively designed for age, gender and race classification problems. Successively, these networks are used without their last layer; the output of the penultimate layers are fed to an age classifier suitably trained for this purpose. Finally, Wan et al. in [57] propose the use of an age estimator specialized on single face categories (white male, black female, and so on) arranged in a hierarchical way.

Methods based on MTL. The methods belonging to P-MTL are characterized by the fact that all the layers of the network are in common among the different tasks, while only the output layer, and sometimes the preceding fully connected layers, are differentiated among the tasks, as shown in Fig. 6(a). This architecture has been adopted in several papers, as [18], [19], [62] and [61].

On the other hand, the systems based on the D-MTL paradigm fuse the different tasks progressively in the net by allocating different layers for the different tasks at hand, as shown in Fig. 6(b); this means that some layers are in common among all the tasks, while at some points specific layers are dedicated to the different tasks. The rationale behind this paradigm is that the different tasks may need, for their intrinsic complexity, some dedicated layers; on the contrary the initial layers, devoted to extract low level features, may be shared among them. Relevant examples are given in [45], [56] and [59].

Gender and race are the biometrics most used in addition to age; good examples are given in [18], [19], [23], [50], [56] and [59]. Moreover, Yang et al. in [61] focused on age estimation and expression recognition. The papers in [45], [56] and [59] propose networks for extracting a large number of descriptors from faces; in particular [45] uses identity independent information, as face detection, visibility, key-points, pose, smile; similarly, [56] and [59] adopt identity dependent information as age and gender, ordinal descriptors as hair length and nominal descriptors as race.

Data reported in Table 2 show that the exploitation of additional biometrics extracted from facial images, as gender, race, expression, have not been widely explored so far; only one out of four paper takes into account such relevant sources of information during the learning and/or the inference phases. However, it has to say that, looking at the distribution over time of papers using additional biometrics, we may note an increasing trend, with roughly half of approaches published in 2017 and 2018.

Experimental results highlight that the use of additional biometric information allows, in some cases, to improve the accuracy; this is particularly effective in the case of expression invariant age estimation as evident in [18], [54], [61] and [45]. It is evident that the use of biometrics as gender, ethnicity and expression determines significant performance improvements of the AE, as witnessed by [45], [52], [54], [61] and [18]; but it cannot be neglected that outstanding results have been also achieved without the use of these additional information, as evident by reading [44], [48] and [30]. So, although these techniques can be profitably used for improving the results, a real evaluation on the relative impact on performance improvements is not simply determinable; hopefully, in the future, further experimental comparisons will be helpful to better characterize this matter.

2.5 Traditional machine learning approaches

Although in the last years most of the proposed methods are based on deep learning architectures, it is worth to account for those approaches based on traditional machine learning paradigms. In particular, in this subsection, we briefly describe the most promising approaches which achieved high performance on commonly used benchmarks; we characterize them in terms of the features used for face description, so as the classification architecture and the relative training techniques.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

Most of the methods rely on known typologies of features: for example [124], [125], [126] use Biologically Inspired Features (BIF), originally proposed in [127]; Lou et al. [128] adopt Local Binary Pattern (LBP) histograms, described in [129], while [125] and [130] use histograms of oriented gradients (HOG), proposed in [105], eventually fused with SIFT features [130]; Wang et al. [131] extract DSIFT histograms from 7×7 ROIs, see [132].

In other cases, the researchers propose their own handcrafted features tailored for age estimation. For instance Lu et al. [133] design a cost-sensitive local binary feature learning (CS-LBFL) that allows to learn discriminant features for face representation directly from raw pixels. The method learns specific hashing functions which transform face patches in binary codes, that are similar whether the age difference is small, rather different when the age difference increases. The face is described by collecting the codes into histograms. To further improve the performance of age estimation, the authors also propose a local binary multi-feature learning (CS-LBMFL), which learns a set of hashing functions for face patches extracted at multiple scales. Iqbal et al. [134] start from the characteristics of craniofacial growth and skin aging reported in [5] and design a local face descriptor, called directional age-primitive pattern (DAPP), which encodes aging by relating it to some texture primitives in a compact code histogram. It takes into account both texture and shape features and achieves remarkable results for real age estimation and age group classification.

As for the choice of the classifier, the Support Vector Machine (SVM) is the most used as, for instance, in [125], [126], [133], [134], [135] and [128]. Han et al. [124] instead use a multi-class AdaBoost algorithm [136] for feature selection and classification, while Sawant et al. [130] adopt a Gaussian Process Classifier (GPC) and a Gaussian Process Regressor (GPR). Wang et al. [131] propose a new method for learning and classification, called raSVM+, which defines relative attributes for SVM using skin smoothness, shape, face size and wrinkles to separate outliers from inliers at the training stage. Despite such information are not available during the test, the authors demonstrate the capability of the raSVM+ to achieve good results.

Furthermore, we provide some insights about the training process adopted by the various methods. In [135], Chang et al. propose a method for real age estimation which uses the ordering property of the age labels for improving the performance of the SVM classifier by changing its weights. Similarly, Feng et al. [126] define an age estimation algorithm independent of the face representation. The prediction models for different age labels are not learnt independently, but simultaneously; the authors capture the dependence among different age labels and control the model complexity. Sawant et al. [130] use two levels of classification/regression, by adopting different GPRs specialized in estimating the age within a certain range; so, after the application of a GPC for determining the age group of a person, a further GPR specialized for the age group provided at the previous stage makes the final age estimation.

To conclude, it is worth mentioning that the method proposed in [125] performs feature selection with the ANOVA algorithm [137] and ensembling at decision level to combine 12 age estimators trained with various types of features.

The brief analysis of the best performing methods recently presented and based on traditional machine learning paradigms puts in evidence the use of more and more complex features (as LBP, SIFT, HOG, CS-LBMFL, BIF, ST) over time, indirectly confirming that the features are the main weaknesses of these systems with respect to ones based on deep learning. It comes up that the exploitation of additional soft biometrics is increasing, as it happens for the raSVM+. Nevertheless, even the best traditional machine learning approaches are not able to achieve performance comparable with those achieved by deep based methods; this is shifting, particularly in the last two years, most of the researchers working on this topic toward deep based methods.

3 Evaluation Metrics

Various evaluation metrics have been proposed for characterizing the performance of facial age analysis methods. In the following we briefly recall the ones commonly used.

Mean Absolute Error (MAE) is the most used evaluation metric for RAE. Given x_i the age estimated on the i-th sample and y_i the corresponding true label, the MAE is the average error over the n test samples. Denoting with $e_i = |x_i - y_i|$ the error on the i-th sample:

$$MAE = \frac{\sum_{i=1}^{n} e_i}{n} \tag{3}$$

Cumulative Score (CS) [138] is another performance metric used for RAE. The CS is computed as:

$$CS(\theta) = \frac{n_{e_i \leq \theta}}{n} \times 100 \tag{4}$$

where $n_{e_i \leq \theta}$ is the number of test samples with e_i less or equal than a threshold θ. Since CS is a function of θ, the results with this metric are given as curves, as θ varies.

ϵ-error is a metric specifically proposed for the Chalearn Looking at People challenge [139], [140] for apparent age estimation, where no ground truth is available; the age labels are computed by asking people to guess them by watching the images. Let be μ_i and σ_i^2 respectively the average and the variance of the distribution of the estimates provided by the persons for the i-th sample of the dataset, the classification error ϵ_i on the i-th sample, is calculated as:

$$\epsilon_i = 1 - e^{-\frac{(x_i - \mu_i)^2}{2{\sigma_i}^2}} \tag{5}$$

The idea behind this index is to weight the estimation errors by taking into account the complexity of the sample under consideration as experienced by people; in particular, the error on the i-th sample is normalized by the corresponding variance, according to (6); samples with high variance give less contribution to the error. The ϵ-error is the average of the ϵ_i over the n samples of the TS.

Accuracy is the evaluation metric generally adopted for age group classification, defined as the ratio between the number of correct classifications and the total number of test samples. This index is used in two forms: the **top-1** and the **1-off**. In the first case, a classification is considered correct if and only if exactly corresponds to the true age group; in the second case, the evaluation metric is more tolerant and considers correct also the classifications for age groups that are adjacent to the true age group.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

4 DATASETS

In recent years several datasets have been collected for age estimation, apparent age estimation and age group classification; the information about these databases are provided in Table 3, Table 4 and Table 5, together with the most relevant characteristics as the size, the age range, the presence of additional annotations as gender, race, expression, the experimental protocol and the evaluation metric. A few examples of face images extracted from these datasets are shown in Fig. 7. An idea of the size and the popularity of the various datasets along the years, in terms of the number of methods using them, is given in Fig. 8. A more detailed description of each dataset is reported in the next subsections, while some comments and insights for future research are reported in Section 4.4.

Fig. 7: Images from the most used datasets for real and apparent age estimation and age group classification.

4.1 Datasets for Real age estimation

IMDB-WIKI [30] is the largest available dataset annotated with age and gender labels. It comes from the fusion of the $460,723$ face images of $20,284$ people from IMDB, the famous web portal of the celebrities, and $62,328$ from Wikipedia. Being composed by pictures of celebrities, mostly looking towards the camera, the images do not present very challenging variations. It is important to mention that the authors do not assure the accuracy of the identities and of the age annotations. Indeed, they took the images and the birth date from the personal profiles of the celebrities and assumed that the timestamp of the photo was correct. In a recent study [48], it has been verified that less than $200,000$ face images are correctly annotated (the so called "cleaned" version, which is not publicly available), especially the ones belonging to the IMDB subset. Despite the lack of significant pose variability and the uncertainty of the annotations, IMDB-WIKI is currently the most used dataset for pretraining networks for AE, which probably allows to achieve remarkable performance thanks to its size.

MORPH-II [141] is the dataset proposed by the Benchmarking Facial Image Analysis Technologies (BeFIT) [2] for age estimation in controlled laboratory conditions, obtained with collaborative people that look towards the camera. So, the images do not have significant pose variations and the quality and image resolution is rather poor (between 200×240 and 400×480 pixels). Notwithstading, this dataset is characterized by a high variability in terms of the age range, being [16,99], ethnicity, with a black/white ratio of about 4:1, and a male/female ratio of 5.5:1; it is created so as to preserve the distribution of age, gender and ethnicity of the whole dataset for each of its 5 folds.

Researchers who want to use this dataset must know that only a few authors adopted the proposed experimental protocol. In other cases, the dataset is partitioned in three non overlapping subsets, one used for the training phase and the union of the remaining two for the testing; the final results are then computed as the average over the two experiments. Alternatively, the dataset is divided in two subsets: 80% for training and 20% for testing. In the following we respectively denote the three protocols as **MORPH-5-CV**, **MORPH-S** and **MORPH-80/20**. Results achieved with different protocols are not taken into account.

Face and Gesture Recognition Research Network aging database (FG-NET) [142] has been proposed by the BeFIT for age estimation in uncontrolled real-life conditions. In our opinion this dataset is not particularly challenging, since the people are aware they are in the field of view of the camera. In addition, the images are not equally distributed over age and only a few images of individuals older than 40 are available. Moreover, its limited size imposes the leave-one-person-out (**LOPO**) evaluation protocol.

FACES [143] database has been collected with the aim of evaluating facial expressions in young, middle-aged and older women and men. People are framed with six different expressions: happiness, anger, fear, sadness, disgust, neutrality. Commonly, the results achieved on this benchmark are reported for each expression but in this paper, for the sake of comparison with the different methods, we provide evaluations averaged all over the expressions. This dataset is particularly suited for characterizing the impact of facial expression on the performance of age estimation algorithms; of course, the fact that the images are collected in controlled laboratory conditions makes it not particularly challenging in terms of both poses and image quality.

LIFESPAN [144] consists of $1,142$ face images of 575 subjects (218 in the range [18,29], 76 in [30,49], 123 in [50,69], and 158 in [70,93]) of different ethnicities, with two different expressions: neutral and happy. The resolution of the faces is 640×480. For this dataset we can do the same considerations already carried out for FACES, with the further limitation of having only two classes of emotions available.

Cross-Age Celebrity Dataset (CACD) [145] is mainly used for face recognition and retrieval purposes, but it is one of the largest publicly available dataset of celebrities images, with age annotations. Its main problems are on the one hand the uncertainty of the annotations (collected with the same protocol proposed for IMDB-WIKI) and on the other hand the reduced age range [14,62]. For these reasons, the authors themselves discourages its use for age estimation purposes.

WebFace dataset [146] contains $59,930$ web photos, labelled with age and gender, in the age range [1,80]. With respect to other publicly available databases, the authors declare a better coverage of face poses and ethnicities. Its main drawback lies in the fact that it has been collected for the experiments of a PhD thesis, so making it difficult to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3: Publicly available datasets for real age estimation.

Ref.	Dataset	Year	# Images	# Subjects	Age range	Other info	Experimental protocol	Metric
[30]	IMDB-WIKI	2016	523,061	82,612	[0,99]	Gender, Identity	Used for pre-training	MAE
[141]	MORPH-5-CV MORPH-S MORPH-80/20	2006	55,608	13,673	[16,99]	Gender	5-CV S1-S2-S3 TR:80% TS:20%	MAE and CS
[142]	FG-NET	2002	1,002	82	[0,69]	Identity	LOPO	MAE and CS
[143]	FACES	2010	2,052	171	[19,80]	Expression	5-CV	MAE and CS
[144]	LIFESPAN	2004	1,142	575	[18,93]	Expression	5-CV	MAE and CS
[145]	CACD	2014	163,446	2,000	[14,62]	Identity	TR:89% TTS:5% TS:6%	MAE
[146]	WebFace	2012	59,930	N/A	[1,80]	Gender	4-CV	MAE

TABLE 4: Publicly available datasets for apparent age estimation.

Ref.	Dataset	Year	# Images	# Subjects	Age range	Other info	Experimental protocol	Metric
[139]	LAP 2015	2015	4,691	4,691	N/A	N/A	TR:52% TTS:24% TS:24%	ϵ-error
[140]	LAP 2016	2016	7,591	7,591	N/A	N/A	TR:54% TTS:20% TS:26%	ϵ-error

TABLE 5: Publicly available datasets for age group classification.

Ref.	Dataset	Year	# Images	# Subjects	Age range	Other info	Experimental protocol	Metric
[147]	Adience	2014	26,580	2,284	[0,60+]	Gender	5-CV	Top1 and 1off
[148]	GROUPS	2009	28,231	5,080	[0,66+]	Gender	TR:80% TS:20%	Top1 and 1off

get; probably that's why only few experimental results are available. Anyway, we report it for the sake of completeness.

4.2 Datasets for Apparent age estimation

ChaLearn Looking at People (LAP) since 2011 organizes several computer vision competitions aimed at recognizing people in images. One of the most successful events is the Age Estimation Challenge, organized for the first time during ICCV 2015 [139] and CVPR 2016 [140]. A dataset for estimating the apparent age of people, based on the opinion of web users, was in these occasions made available to the community. Although these datasets do not provide large number of images, they are rightly considered the most challenging ones in terms of face variations and consequently the most commonly adopted in recent years. These reasons make them the most suited for benchmarking the performance of AE methods both in terms of completeness and of available performance data.

4.3 Datasets for Age group classification

Adience [147] is a dataset for age and gender classification collected in real-world conditions, including variations in appearance, pose, lighting and image quality. The whole dataset consists of $26,580$ face images with different yaws from frontal angle and only $13,649$ almost frontal. The faces are divided in eight not balanced age categories: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+. These characteristics make the dataset particularly suited for the benchmark of age group classification algorithms in challenging conditions.

Images of Groups (GROUPS) [148] is the dataset proposed by the BeFIT for age group classification. The images of the dataset are divided in seven age categories: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. Considering that the class 20-36 contains more than $15,000$ faces, this dataset is significantly unbalanced. Therefore, the set up proposed by the authors [148] is made of $3,500$ images used for training and $1,050$ for testing, being careful that training and testing images are equally distributed across the seven age groups.

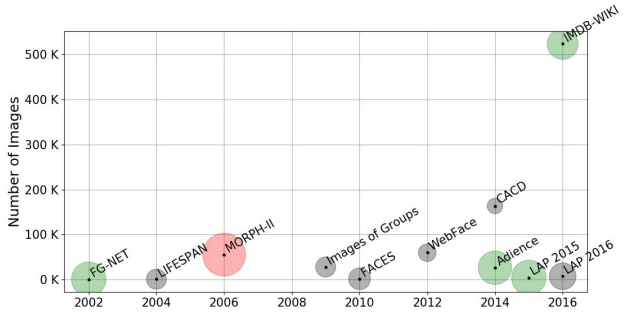


Fig. 8: Datasets for age analysis: the larger is the radius of the circle, the higher is the number of methods using it.

It is worth pointing out that the complexity of this database is similar to that of Adience; Images of Groups exhibits less variations in terms of pose, consists of less images in its balanced version and has a single age group 37-65 for the adulthood. Consequently, Adience is widely better and this is the reason why GROUPS is less used.

4.4 Analysis of the datasets

We give here some insights and recommendations that may guide the researchers to select the datasets more suited for their specific needs; in particular, we discuss which datasets should be used for training a deep network for age analysis and which ones are more challenging and exhibit more adequate variations for taking into account real situations. Finally, we discuss the lacks of the currently available datasets.

4.4.1 Datasets for training a deep network for AE

To perform an effective training of a network for age estimation it is fundamental to have a very large and heterogeneous dataset. Currently, the dataset most suited to this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

need is the IMDB-WIKI, used by almost all the researchers in their experimentations. Despite the remarkable effort done for the collection of IMDB-WIKI, the authors themselves point out that the procedure used for gathering images and labels is not adequate to guarantee the accuracy of the age annotations; nevertheless, this aspect does not have a dramatic impact on the performance of the deep networks pre-trained with it, as we will show in Section 5, as the label imprecisions fall inside the human uncertainty of the true age perception. Moreover, being available the identity and the gender annotations, IMDB-WIKI can be used likewise for face recognition pre-training and for BTL and MTL.

4.4.2 Datasets for performance benchmarking

Fig. 8 points out that MORPH-II and FG-NET are the most used datasets for evaluating the performance of AE algorithms, since they were the first datasets available (2006 and 2002 respectively) and recommended by the BeFIT. The availability of several results on those datasets in the last five years is surely a stimulus to their use, but, as we have previously observed, their images are not particularly challenging; they may be profitably used in conjunction with applications planned to run in controlled conditions.

On the contrary, the recent ChaLearn LAP datasets, collected in more heterogeneous scenarios have a intrinsic wide variability that, jointly with the availability of many experimental results, make them particularly suited for benchmarking AE methods; our experience suggests their use despite they are smaller than MORPH-II and FG-NET, as for the robustness of the obtained systems.

Whether the reader needs a larger dataset, we suggest to consider CACD and WebFace ($163,446$ and $59,930$ images); as for their use, it is important to consider that the former is not recommended for testing purposes due to the inaccuracy of the age annotations, while the latter is less popular and not easy to find at the point that only a few experimental results are currently available. Adience and Images of Groups are datasets as challenging as the ones proposed for the ChaLearn LAP competitions, since the available images exhibit strong variations in terms of image resolution and quality and facial poses. They may be considered as good alternatives, even because the age group annotation is available, but not the real age.

Recently, many researchers have experienced the benefits of using facial soft biometrics; a good dataset for analysing the impact of the facial expression on the perceived age is FACES. Indeed the other datasets typically contain people smiling or with a neutral expression (including LIFESPAN).

MORPH-II and IMDB-WIKI, instead, may be used for evaluating the impact of gender on the age estimation accuracy or the performance of methods based on multi-task learning. However, they are acquired in controlled conditions and the available images have very good quality, with the face in frontal pose that occupies almost the entire image; therefore, the AE is significantly less challenging, as demonstrated by the results in Section 5.

4.4.3 Future trend

The evidence that arises from the analysis of the datasets is the unavailability of face images extracted from real video sequences (”in the wild”), which have very challenging and heterogeneous variations in terms of poses, illumination conditions, image quality, face size and occlusions. These situations, and their combinations, are mostly absent also in the most challenging as ChaLearn LAP, Adience and Images of Groups. This is surely a big lack if we consider that in the recent years there is a growing interest for systems able to estimate the age of people by using classic surveillance cameras in uncontrolled conditions.

5 Performance analysis

Here we present and discuss the results of the methods on standard datasets, highlighting the state of the art performance for each specific age analysis task.

5.1 Real age estimation

Giving a first look at the Table 6, we note that the large variability of the benchmarking datasets together with the variety of experimental protocols makes it difficult the identification of research trends. It is possible to note that a large amount of results are available on MORPH-II and FG-NET with respect to FACES, LIFESPAN, CACD and WebFace. Therefore, the comparison of the results obtained using the first two datasets may be considered more significant as it involves a great amount of different methods, even if the images in these benchmarks are not particularly challenging compared with the ones available in the ChaLearn LAP datasets, in Adience and in Images of Groups. Nevertheless, the analysis of the results on FACES more than LIFESPAN is interesting from another point of view; in fact, thanks to the variability of the combinations of expressions and ages, it allows to characterize the robustness of the methods with respect to different facial deformations and expressions. These two aspects are discussed in the following.

The methods [44], [48] and [63] stand out as the most effective ones on MORPH-II, and represent currently the state of the art on this database. It has to be considered that they are very recent, being proposed in the last two years; each of them achieves the top results on one of the experimental protocols on MORPH-II, respectively MORPH-S, MORPH-80/20 and MORPH-5-CV; among these three it is worth highlighting that [44] is the only one that provides the experimental results on two MORPH-II protocols, getting the top and the second top results on MORPH-S and MORPH-80/20. So, it is the most experimented method on MORPH-II and its results can be surely considered statistically significant and impressive as for the error obtained.

The top three performance on FG-NET are achieved respectively by [45], [125] and [48]; among these, while [45] presents the results only on FG-NET, the other two have been experimented also on MORPH-II and gets respectively the 4th-top on MORPH-S and top-1 on MORPH-80/20. It has to be pointed out that [44], the best on MORPH-II is ranked as 4th-top on FG-NET.

So, considering the experimental analysis of the two databases jointly, we may conclude that nine are the methods that achieve one of the top-3 positions on these two databases, say [30], [44], [45], [48], [50], [57], [59], [125] and [62]; four of them achieve at least top-1 rank, say [44], [45], [48] and [63], while only two of them achieve two ranks in the top-3, in particular [44] and [48].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 6: Performance over different datasets in terms of MAE. Top-3 results are highlighted. The methods are sorted per increasing year of publication; those marked with * are not based on deep networks.

Method	MORPH-II			FG-NET	FACES	LIFESPAN	CACD	WebFace
	5-CV	S	80/20					
Yang et al. [18], 2013		8 (3.48)			5 (7.01)	1 (3.87)		
Yi et al. [19], 2014		10 (3.63)						
Li et al. [26], 2015		9 (3.61)						4 (7.27)
Huerta et al. [27], 2015	5 (3.88)							
Wang et al. [25], 2015			8 (4.77)	13 (4.26)				
Chang et al. [135]*, 2015		11 (3.82)		17 (4.48)				
Han et al. [124]*, 2015	4 (3.80)			18 (4.80)				
Lu et al. [133]*, 2015				15 (4.36)	3 (5.49)	5 (5.44)		
Liu et al. [125]*, 2015		4 (2.97)		2 (2.81)				
Niu et al. [33], 2016			5 (3.27)					
Rothe et al. [34], 2016			7 (3.45)					
Rothe et al. [30], 2016			3 (2.68)	7 (3.09)			3 (6.52)	
Yang et al. [35], 2016		7 (3.23)						
Wang et al. [131]*, 2016	6 (5.05)			12 (4.07)				
Xing et al. [50], 2017		3 (2.96)						1 (5.75)
Li et al. [51], 2017		6 (3.06)						3 (6.04)
Liu et al. [47], 2017				11 (3.93)				
Antipov et al. [48], 2017		5 (2.99)	1 (2.35)	3 (2.84)				
Ranjan et al. [45], 2017				1 (2.00)				
Tan et al. [44], 2017		1 (2.70)	2 (2.52)	4 (2.96)			1 (4.68)	
Hou et al. [53], 2017				5 (3.01)				
Liu et al. [46], 2017				10 (3.71)	2 (3.90)	3 (4.25)		
Feng et al. [126]*, 2017				14 (4.35)				2 (6.03)
Han et al. [59], 2018	3 (3.00)							
Wan et al. [57], 2018		2 (2.93)	6 (3.30)				2 (5.22)	
Yang et al. [61], 2018					4 (5.95)	2 (4.01)		
Liu et al. [58], 2018			4 (2.89)	8 (3.31)	1 (3.82)			
Lou et al. [128]*, 2018					6 (7.41)	4 (5.26)		
Yoo et al. [62], 2018	2 (2.91)		4 (2.89)	9 (3.43)				
Sawant et al. [130]*, 2019				16 (4.41)				
Taheri et al. [63], 2019	**1 (2.81)**			6 (3.08)				
Best results	**2.81**	**2.70**	**2.35**	**2.00**	**3.82**	**3.87**	**4.68**	**5.75**

The analysis of the results obtained on the other databases has been limited to methods achieving the top performance, as the total number of experiments is much lower; the top results on FACES, LIFESPAN, CACD and WebFace have been respectively achieved by [18], [44], [58] and [50]. Except for [44], no one of these methods has get a rank among the top-3 in the previously considered databases MORPH-II and FG-NET.

It emerges that [44] appears among the best methods also on these databases; it is the most widely experimented method and overall the most frequently winning.

Finally, it is important to note that the methods not based on deep learning [124], [125], [126], [126], [128], [130], [131], [133], [135] are in the average less competitive with respect to those using deep learning; they do not reach any top results and, among all the results in the considered databases at most get a 4th-top and a 2nd-top rank, say [124] and [125], respectively on MORPH-5CV and FG-NET.

It's worth understanding the similarities of the methods that obtained the top performance, so as to realize the strength points that are at the basis of their success.

To this aim, we highlight that the methods proposed in [63], [44] and [30] are very similar. In particular, the last two formulate the problem as a CAE, using the cross-entropy loss function, while the first [63] uses the RVE; moreover all of them perform the AE pre-training of a VGG-16 net, using the database IMDB-WIKI. The better performance of [44] and [63] probably derives from the preprocessing of the input face images introducing a massive data augmentation. It is reasonable to consider this data augmentation as one of the strength points that allows this method to obtain the best performance also on the CACD dataset.

Differently, Antipov et al. [48] formulate the problem as a DAE and use the VGG-Face net, pre-trained for face recognition on almost $1,000,000$ face images. Ranjan et al. [45] use their own network pre-trained for face recognition and fine tuned applying RVE with a Gaussian loss function.

In conclusion, except for the latter, the approaches that obtained the best performance use VGG-16 with AE pre-training or FR pre-training (VGG-Face). This trend is over time confirmed by the results obtained by other approaches based on VGG-16 or VGG-Face, as [46], [47], [53], [57] and [34] that achieve good performance on MORPH-II and FG-NET and state of the art results on WebFace [50].

Looking at the performance on FACES and LIFESPAN it emerges the unexpected fact that the winners are methods not based on VGG-Face; in particular, on those databases, the best results have been respectively obtained by a network ResNet-101, as [46], and by the method [58] based on ScatNet that provides a face description very robust with respect to deformations. The latter property reveals particularly effective as FACES contains face images with many different expressions, with a wide variety of natural face deformations affecting significantly the performance of deep approaches, when not adequately taken into account.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 7: ϵ-error on LAP data. Top-3 results are highlighted.

Method	LAP 2015 Validation	LAP 2015 Test	LAP 2016 Test
Rothe et al. [21], 2015	5 (0.2951)	2 (0.2650)	
Liu et al. [20], 2015	3 (0.2923)	3 (0.2707)	
Kuang et al. [23], 2015	10 (0.3273)	4 (0.2873)	
Zhu et al. [22], 2015	9 (0.3163)	5 (0.2948)	6 (0.3405)
Yang et al. [24], 2015	12 (0.3806)	6 (0.3058)	
Ranjan et al. [29], 2015		7 (0.3733)	
Antipov et al. [39], 2016			1 (0.2411)
Huo et al. [40], 2016			4 (0.3214)
Uricar et al. [41], 2016			5 (0.3361)
Malli et al. [42], 2016			7 (0.3668)
Gurpinar et al. [32], 2016	14 (0.4833)	8 (0.5240)	9 (0.3740)
Chen et al. [31], 2016	6 (0.2970)		
Tan et al. [44], 2017	1 (0.2800)	1 (0.2635)	2 (0.3100)
Ranjan et al. [45], 2017	4 (0.2930)		
Liu et al. [46], 2017	7 (0.3120)		
Liu et al. [47], 2017	11 (0.3690)		
Dehghan et al. [52], 2017			3 (0.3190)
Wan et al. [57], 2018	2 (0.2900)		
Liu et al. [58], 2018	8 (0.3150)		
Han et al. [59], 2018	13 (0.4490)		
Duan et al. [60], 2018			8 (0.3679)
Best results	**0.2800**	**0.2635**	**0.2411**

TABLE 8: Performance over Adience and Groups in terms of top-1 and 1-off. Top-3 results are highlighted. Methods marked with * are not based on deep networks.

Method	Adience top-1	Adience 1-off	Groups top-1	Groups 1-off
Eidinger et al. [147]*, 2014	13 (45.1)	12 (80.7)	2 (66.6)	3 (95.3)
Levi et al. [28], 2015	12 (50.7)	9 (84.7)		
Rothe et al. [30], 2016	3 (64.0)	3 (96.6)		
Hou et al. [36], 2016	7 (61.1)	5 (94.0)	3 (64.6)	2 (96.1)
Ozbulak et al. [38], 2016	10 (57.9)			
Chen et al. [31], 2016	11 (52.9)	8 (88.4)		
Dong et al. [37], 2016			5 (54.0)	5 (91.0)
Zhang et al. [54], 2017	1 (67.3)	1 (97.5)		
Hou et al. [53], 2017	1 (67.3)	2 (97.4)	4 (54.2)	4 (93.0)
Lapuschkin et al. [80], 2017	4 (62.8)	4 (95.8)		
Iqbal et al. [134]*, 2017	5 (62.2)			
Dehghan et al. [52], 2017	6 (61.3)		1 (70.5)	1 (96.2)
Qawaqneh et al. [55], 2017	9 (59.9)	7 (90.6)		
Duan et al. [60], 2018	2 (66.5)			
Liu et al. [58], 2018	8 (60.2)	6 (93.7)		
Best results	**67.3**	**97.5**	**70.5**	**96.2**

5.2 Apparent age estimation

The competition organized by ChaLearn LAP, together with the provided datasets, stimulated the researchers to design and test their methods in more challenging situations with respect to the datasets available for RAE; this aspect determined an undeniable significant advance of the state of the art in apparent age estimation. The experimental results are reported in chronological order in Table 7.

It is noticeable that the best results on these competitions are obtained by the same best methods for real age estimation, as [30], [44], [48] and [45]. In particular, Tan et al. [44] got the best ϵ-error on both the datasets of 2015, while Antipov et al. [39] won the competition of 2016 with an approach similar to the one presented in [48]. We can also note that Rothe et al. [21] won the challenge of 2015 with the same deep network described in [30]. The method [45] also achieves one of the best results on the validation set of 2015.

The participants who achieved the best results during the competition of 2015 used the VGG-16 net, say [21], [23], [24], or the GoogLeNet, as [20], [22] with various types of ensembling at decision level; the choice of using a different kind of CNN with a combination at decision level [29], have not appreciably improved the performance.

In the ChaLearn LAP 2016 challenge, most of the participants, starting from the evidence that in the previous edition the most effective methods were those based on VGG-16, decided to adopt VGG-16 or VGG-Face, as [39], [40], [41], [42] and [32]; most of those methods, including the winner, used an ensembling at decision level. The winner adopted a combination of VGG-16 networks. Such a choice, together with a suited preprocessing (combining face detection score and facial landmarks information), made the algorithm so effective to achieve the state of the art result on this dataset.

For completeness, it should be pointed out that other recent methods have been evaluated on the ChaLearn LAP 2015 and 2016 datasets. A few ones, as [31], [45] and [52] got performance close to the top ones; others outperform them on the validation [44], [57] and on the test set [44].

The methods that achieve similar results are [31], [45] and [52] that are based on proprietary CNNs; top results are also got by [57] and [44] that are based on VGG-16.

5.3 Age group classification

The results in Table 8 point out that the datasets acquired in uncontrolled environments, such as Adience and Images of Groups, are significantly challenging being the top-1 performance around $65-70\%$ and 1-off above 95%.

Also for this problem we may see that most of the methods are based on VGG-16 or VGG-Face, for instance [30], [36], [53], [55] and [38]; sometimes an ensemble at decision level is adopted, as done in [53] and [30].

However, the best performance is achieved by [54] that uses RoR [149], while the third rank is obtained by [60] with a proprietary CNN. It is worth noting that both the approaches proposed in [54] and [60] are based on BTL and use gender or ethnicity to improve the performance.

It is important to say that here, the best methods based on VGG-Face (see [36], [53]) are outperformed by the proprietary CNN described in [52]; we justify this result by considering that the dataset used by the authors of [52] for FR pre-training is so wide (about $4,000,000$ images) that the net is too specialized to deal with a so high data variability.

This experimental evidence confirms the observation done in Section 2.2; only further experimentation of novel network architectures may allow the researchers to conclude whether VGG-16 may be overcome in the future, especially when using challenging datasets in the wild.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2910522, IEEE Transactions on Pattern Analysis and Machine Intelligence

6 Where are we going?

Age estimation is an important topic in the area of face analysis and in the last years many researchers have contributed to the growth of the field proposing a variety of novel approaches, organizing competitions and publicly distributing wide datasets of images.

As in many fields, also in this research area deep learning has had a significant impact, determining impressive performance improvements with respect to traditional machine learning approaches. Nevertheless, we deem that there is still much research work to do; despite the advancements that are under the public eye, the currently available methods and datasets are not yet completely ready for real world exploitation. Moreover, the large disparity of performance among the various datasets is symptomatic of the fact that the methods are still dependent at a given extent on training data that are probably not yet enough representative of the real world complexity, especially due to the intrinsic variability of the problem. A significant experimental evaluation in real world conditions is needed, despite it requires a very wide dataset of real video sequences to be collected and annotated. This is not a trivial task, especially if we consider that, in the real world, the actual age of the acquired faces is not easy to collect.

Anyway, although the number of AE methods is very wide, as for network architecture, based either on a single or ensemble of classifiers, and learning procedures, some relevant aspects occur in most of the the state of the art methods. They allow to draw some conclusions, surely useful to identify research trends and open issues. Among the various aspects, we consider the most relevant ones for achieving effective systems, the following:

- VGG-16 revealed to be the best network architecture for the problem at hand, especially if the system uses different source of data, for example relative to both age estimation and face recognition; its success is related to the capability of extracting effective face features that allow the last layers, the ones devoted to provide the final decision, to obtain state of the art performance for age estimation. Anyway, it cannot be excluded that recently proposed deep network architectures not yet considered for age analysis, especially applied on more challenging datasets as ChaLearn LAP and Adience (or even over datasets in the wild not yet available), can outperform the current state of the art methods.
- The ensemble of different nets, appears to be not strictly necessary, as the results do not definitely overcome the ones achieved with a single net: this unexpected result is surely a stimulus for further scientific investigations aimed at understanding why the superiority of a multi-classification paradigm, experimentally demonstrated in many fields, here does not exhibit the effectiveness one would expect; anyway, it emerges that the combination at decision level is more effective than other kinds of ensemble strategies as the fusion at description level.
- Data augmentation is one of the keys of the success of deep approaches and age estimation does not constitute an exception; indeed this problem, among all the others on face images, is even more challenging, at the point that human beings are only partially able to estimate the ages. The augmentation of data is here promising because many factors may influence the performance; in particular we may expect variations on data as changes in orientation, resolution, contrast and brightness of face images, but more importantly, much more heavy variations in the faces, related to the face expression, race, and so on. Consequently, it is crucial to provide the deep network the adequate capability to generalize with respect to the above mentioned heterogeneous face variations and can be even more effective after face normalization.
- The pre-training of the network with age estimation data (AE pre-training), supported by a careful data augmentation, is the procedure that experimentally demonstrated to be the best suited one, allowing to achieve excellent results when a sufficiently large training set is available. On the other hand, the pre-training with data for face recognition (FR pre-training), followed by a fine tuning process for the specific age estimation task, is a good solution when the available dataset for age estimation is not big enough, even if not so competitive as the AE pre-training with sufficiently large training sets. To this concern, it has to clarify that multi-task learning and FR pre-training are not complementary each other, so they should be used alternatively; currently, the impact of the use of these techniques on the age estimation accuracy has not experimentally demonstrated and more extensive future experiments may be much needed.

In conclusion, although the impressive performance improvements obtained up to now, age estimation is still likely to be further investigated so as to achieve the performance needed for an applicative exploitation. Nevertheless, we can claim that deep learning revealed to be very effective and there is large evidence that the traditional approaches, based on handcrafted features, are evidently less competitive, so enshrining the definitive adoption of deep based approaches.

The deep learning revolution is exploded also for estimating the age!

References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of IEEE Conf. on CVPR*, 2001, pp. 511–518.

[2] K. I. of Technology, "Befit - benchmarking facial image analysis technologies," Available: http://fipa.cs.kit.edu/412.php, 2011.

[3] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Tech. Rep., 2007.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2018, pp. 67–74.

[5] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. on PAMI*, pp. 1955–1976, 2010.

[6] T. Dhimar and K. Mistree, "Feature extraction for facial age estimation: A survey," in *Proc. of IEEE Int. Conf. WISPNET*, 2016, pp. 2243–2248.

[7] O. Osman and M. Yap, "Computational intelligence in automatic face age estimation: a survey," *IEEE Trans. on ETCI*, 2018.

[8] X. Shu, G.-S. Xie, Z. Li, and J. Tang, "Age progression: Current technologies and applications," *Elsevier Neurocomputing*, pp. 249–261, 2016.

[9] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? a survey on soft biometrics," *IEEE Trans. on IFS*, pp. 441–467, 2016.

[10] Y. Sun, M. Zhang, Z. Sun, and T. Tan, "Demographic analysis from biometric data: Achievements, challenges, and new frontiers," *IEEE Trans. on PAMI*, pp. 332–351, 2018.

[11] S. Fu, H. He, and Z.-G. Hou, "Learning race from face: A survey," *IEEE Trans. on PAMI*, pp. 2483–2509, 2014.

[12] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: past, present and future," *Elsevier CVIU*, pp. 1–24, 2015.

[13] C.-B. Ng, Y.-H. Tay, and B.-M. Goi, "A review of facial gender recognition," *Springer PAA*, pp. 739–755, 2015.

[14] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. on PAMI*, pp. 1113–1133, 2015.

[15] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Transactions on Intelligent Systems and Technology*, p. 37, 2016.

[16] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Springer IJCV*, pp. 1–28, 2017.

[17] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*, 2018.

[18] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen, "Automatic age estimation from face images via deep ranking," *Networks*, pp. 1872–1886, 2013.

[19] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. of Springer ACCV*, 2014, pp. 144–158.

[20] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, "Agenet: Deeply learned regressor and classifier for robust apparent age estimation," in *Proc. of IEEE ICCV Workshops*, 2015, pp. 16–24.

[21] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proc. of IEEE ICCV Workshops*, 2015, pp. 10–15.

[22] Y. Zhu, Y. Li, G. Mu, and G. Guo, "A study on apparent age estimation," in *Proc. of IEEE ICCV Workshops*, 2015, pp. 267–273.

[23] Z. Kuang, C. Huang, and W. Zhang, "Deeply learned rich coding for cross-dataset facial age estimation," in *Proc. of IEEE ICCV Workshops*, 2015, pp. 96–101.

[24] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, and X. Geng, "Deep label distribution learning for apparent age estimation," in *Proc. of IEEE ICCV Workshops*, 2015, pp. 102–108.

[25] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. of IEEE Winter Conf. on Applications of Computer Vision*, 2015, pp. 534–541.

[26] S. Li, J. Xing, Z. Niu, S. Shan, and S. Yan, "Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition," in *Proc. of IEEE Conf. on CVPR*, 2015, pp. 222–230.

[27] I. Huerta, C. Fernndez, C. Segura, J. Hernando, and A. Prati, "A deep analysis on age estimation," *IEEE PRL*, pp. 239 – 249, 2015.

[28] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. of IEEE Conf. on CVPR Workshops*, 2015, pp. 34–42.

[29] R. Ranjan, S. Zhou, J. Cheng Chen, A. Kumar, A. Alavi, V. M. Patel, and R. Chellappa, "Unconstrained age estimation with deep convolutional neural networks," in *Proc. of IEEE ICCV Workshops*, 2015, pp. 109–117.

[30] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. Journal of Computer Vision*, Aug 2016.

[31] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," in *Proc. of IEEE Int. Conf. on BTAS*, 2016, pp. 1–8.

[32] F. Gurpinar, H. Kaya, H. Dibeklioglu, and A. Salah, "Kernel elm and cnn based facial age estimation," in *Proc. of IEEE Conf. on CVPR Workshops*, 2016, pp. 80–86.

[33] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proc. of IEEE Conf. on CVPR*, 2016, pp. 4920–4928.

[34] R. Rothe, R. Timofte, and L. Van Gool, "Some like it hot-visual guidance for preference prediction," in *Proc. of IEEE Conf. on CVPR*, 2016, pp. 1–9.

[35] Y. Yang, F. Chen, X. Chen, Y. Dai, Z. Chen, J. Ji, and T. Zhao, "Video system for human attribute analysis using compact convolutional neural network," in *Proc. of IEEE ICIP*, 2016, pp. 584–588.

[36] L. Hou, C.-P. Yu, and D. Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint arXiv:1611.05916*, 2016.

[37] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," *Neurocomputing*, pp. 4 – 10, 2016.

[38] G. Ozbulak, Y. Aytar, and H. K. Ekenel, "How transferable are cnn-based features for age and gender classification?" in *Proc. of IEEE Int. Conf. on BIOSIG*, 2016, pp. 1–6.

[39] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Apparent age estimation from face images combining general and children-specialized deep learning models," in *Proc. of IEEE Conf. on CVPR Workshops*, 2016, pp. 96–104.

[40] Z. Huo, X. Yang, C. Xing, Y. Zhou, P. Hou, J. Lv, and X. Geng, "Deep age distribution learning for apparent age estimation," in *Proc. of IEEE Conf. on CVPR Workshops*, 2016, pp. 722–729.

[41] M. Uricar, R. Timofte, R. Rothe, J. Matas, and L. V. Gool, "Structured output svm prediction of apparent age, gender and smile from deep features," in *Proc. of IEEE Conf. on CVPR Workshops*, 2016, pp. 730–738.

[42] R. C. Malli, M. Aygun, and H. K. Ekenel, "Apparent age estimation using ensemble of deep learning models," in *Proc. of IEEE Conf. on CVPR Workshops*, 2016, pp. 714–721.

[43] B. Hebda and T. Kryjak, "A compact deep convolutional neural network architecture for video based age and gender estimation," *Federated Conf. on Computer Science and Information Systems*, pp. 787–790, 2016.

[44] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, and S. Z. Li, "Efficient group-n encoding and decoding for facial age estimation," *IEEE Trans. on PAMI*, 2017.

[45] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Proc. of IEEE Int. Conf. on Automatic Face & Gesture Recognition*, 2017, pp. 17–24.

[46] H. Liu, J. Lu, J. Feng, and J. Zhou, "Ordinal deep learning for facial age estimation," *IEEE Trans. on CSVT*, 2017.

[47] H. Liu, J. Lu, J. Feng, and J. Zhou, "Group-aware deep feature learning for facial age estimation," *Elsevier PR*, pp. 82–94, 2017.

[48] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Elsevier PR*, pp. 15–26, 2017.

[49] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Trans. on IP*, pp. 3087–3097, 2017.

[50] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," *Elsevier PR*, pp. 106–116, 2017.

[51] K. Li, J. Xing, W. Hu, and S. J. Maybank, "D2c: Deep cumulatively and comparatively learning for human age estimation," *Elsevier PR*, pp. 95 – 105, 2017.

[52] A. Dehghan, E. G. Ortiz, G. Shu, and S. Z. Masood, "Dager: Deep age, gender and emotion recognition using convolutional neural network," *arXiv preprint arXiv:1702.04280*, 2017.

[53] L. Hou, D. Samaras, T. Kurc, Y. Gao, and J. Saltz, "Convnets with smooth adaptive activation functions for regression," in *Int. Conf. on Artificial Intelligence and Statistics*, 2017, pp. 430–439.

[54] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, "Age group and gender estimation in the wild with deep ror architecture," *IEEE Access*, pp. 22 492–22 503, 2017.

[55] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep convolutional neural network for age estimation based on vgg-face model," *arXiv preprint arXiv:1709.01664*, 2017.

[56] F. Wang, H. Han, S. Shan, and X. Chen, "Deep multi-task learning for joint prediction of heterogeneous face attributes," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2017, pp. 173–179.

[57] J. Wan, Z. Tan, Z. Lei, G. Guo, and S. Z. Li, "Auxiliary demographic information assisted age estimation with cascaded structure," *IEEE Trans. on Cybernetics*, 2018.

[58] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *IEEE Trans. on Information Forensics and Security*, pp. 292–305, 2018.

[59] H. Han, A. K. Jain, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. on PAMI*, 2017.

[60] M. Duan, K. Li, and K. Li, "An ensemble cnn2elm for age estimation," *IEEE Trans. on IFS*, pp. 758–772, 2018.

[61] H.-F. Yang, B.-Y. Lin, K.-Y. Chang, and C.-S. Chen, "Joint estimation of age and expression by combining scattering and convolutional networks," *ACM Trans. on Multimedia Computing, Communications and Applications*, 2018.

[62] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim, "Deep facial age estimation using conditional multitask learning with weak label expansion," *IEEE Signal Processing Letters*, 2018.

[63] S. Taheri and Ö. Toygar, "On the use of dag-cnn architecture for age estimation with multi-stage features fusion," *Neurocomputing*, 2019.

[64] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. on PAMI*, pp. 2401–2412, 2013.

[65] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Elsevier Neurocomputing*, pp. 489–501, 2006.

[66] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, pp. 203–224, 2007.

[67] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *Journal of Machine Learning Research*, pp. 1939–1959, 2005.

[68] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. on IP*, pp. 1178–1188, 2008.

[69] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *ACS JCICS*, pp. 1947–1958, 2003.

[70] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, 2004.

[71] H.-T. Lin and L. Li, "Reduction from cost-sensitive ordinal ranking to weighted binary classification," *Neural Comput.*, pp. 1329–1367, 2012.

[72] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, pp. 2278–2324, 1998.

[73] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[75] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. of IEEE Conf. on CVPR*, 2015, pp. 1–9.

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conf. on CVPR*, 2016, pp. 770–778.

[77] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." in *Proc. of BMVC*, 2015, p. 6.

[78] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. on PAMI*, pp. 1872–1886, 2013.

[79] C. Cortes and V. Vapnik, "Support-vector networks," *Springer Machine Learning*, pp. 273–297, 1995.

[80] S. Lapuschkin, A. Binder, K.-R. Müller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," in *Proc. of IEEE ICCV*, 2017.

[81] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[82] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *CoRR*, 2017.

[83] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, 2018.

[84] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. on SMC*, 1992.

[85] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento, "Reliability parameters to improve combination strategies in multi-expert systems," *Springer PAA*, pp. 205–214, 1999.

[86] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Elsevier Artificial Intelligence*, pp. 239–263, 2002.

[87] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Comp. Int. Mag.*, pp. 41–53, 2016.

[88] A. Tsymbal, S. Puuronen, and D. W. Patterson, "Ensemble feature selection with the simple bayesian classification," *Elsevier Information Fusion*, pp. 87–100, 2003.

[89] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Comp. Int. Mag.*, pp. 41–53, 2016.

[90] L. Breiman, "Bagging predictors," *Springer Machine Learning*, pp. 123–140, 1996.

[91] S. Yan, S. Shan, X. Chen, and W. Gao, "Locally assembled binary (lab) feature with feature-centric cascade for fast and accurate face detection," in *Proc. of IEEE Conf. on CVPR*, 2008, pp. 1–7.

[92] R. Ranjan, V. M. Patel, and R. Chellappa, "A deep pyramid deformable part model for face detection," in *Proc. of IEEE Int. Conf. on BTAS*, 2015, pp. 1–8.

[93] D. Oro, C. Fernández, J. R. Saeta, X. Martorell, and J. Hernando, "Real-time gpu-based face detection in hd video sequences," in *Proc. of IEEE ICCV Workshops*, 2011, pp. 530–537.

[94] Microsoft, "Microsoft project oxford api," https://dev.projectoxford.ai/docs/services/, 2018.

[95] SeetaFaceEngine, "Seetafaceengine," https://github.com/seetaface, 2018.

[96] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *Proc. of Springer ECCV*, 2014, pp. 109–122.

[97] B. Jun, I. Choi, and D. Kim, "Local transform features and hybridization for accurate face and human detection." *IEEE Trans. on PAMI*, pp. 1423–1436, 2013.

[98] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *Proc. of IEEE Int. Joint Conf. on Biometrics*, 2014, pp. 1–8.

[99] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, "Efficient boosted exemplar-based face detection," in *Proc. of IEEE Conf. on CVPR*, 2014, pp. 1843–1850.

[100] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic part model for unsupervised face detector adaptation," in *Proc. of IEEE ICCV*, 2013, pp. 793–800.

[101] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Elsevier IVC*, pp. 790–799, 2014.

[102] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Proc. of Springer ECCV*, 2014, pp. 720–735.

[103] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. on PAMI*, pp. 1627–1645, 2010.

[104] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. of BMVC*, 2009.

[105] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Conf. on CVPR*, 2005, pp. 886–893.

[106] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. of Springer ECCV*, 2016, pp. 21–37.

[107] R. Girshick, "Fast r-cnn," in *Proc. of IEEE ICCV*, 2015, pp. 1440–1448.

[108] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Trans. on PAMI*, pp. 1137–1149, 2017.

[109] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of IEEE Conf. on CVPR*, 2016, pp. 779–788.

[110] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint*, 2017.

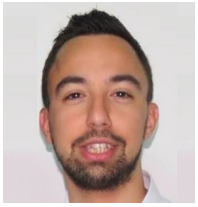
[111] N. Srinivas, H. Atwal, D. C. Rose, G. Mahalingam, K. Ricanek, and D. S. Bolme, "Age, gender, and fine-grained ethnicity prediction using convolutional neural networks for the east asian face dataset," in *Proc. of IEEE Int. Conf on Automatic Face & Gesture Recognition*, 2017, pp. 953–960.

[112] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. of IEEE Conf. on CVPR*, 2014, pp. 1867–1874.

[113] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. of IEEE Conf. on CVPR*, 2013, pp. 532–539.

[114] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. on PAMI*, 2019.
[115] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *Proc. of Springer ECCV*, 2014, pp. 1–16.
[116] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. of IEEE Conf. on CVPR*, 2013, pp. 3476–3483.
[117] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Elsevier CVIU*, pp. 38–59, 1995.
[118] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on PAMI*, pp. 681–685, 2001.
[119] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Proc. of Springer ECCV*, 2016, pp. 579–596.
[120] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Springer IJCV*, pp. 211–252, 2015.
[121] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.
[122] G. Guo and G. Mu, "Joint estimation of age, gender and ethnicity: Cca vs. pls," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2013, pp. 1–6.
[123] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.
[124] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. on PAMI*, pp. 1148–1161, 2015.
[125] K.-H. Liu, S. Yan, and C.-C. J. Kuo, "Age estimation via grouping and decision fusion," *IEEE Trans. on IFS*, pp. 2408–2423, 2015.
[126] S. Feng, C. Lang, J. Feng, T. Wang, and J. Luo, "Human facial age estimation by cost-sensitive label ranking and trace norm regularization," *IEEE Trans. on Multimedia*, pp. 136–148, 2017.
[127] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *Springer IJCV*, pp. 93–104, 2008.
[128] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu, and T. Gevers, "Expression-invariant age estimation using structured learning," *IEEE Trans. on PAMI*, pp. 365–375, 2018.
[129] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on PAMI*, pp. 2037–2041, 2006.
[130] M. M. Sawant and K. Bhurchandi, "Hierarchical facial age estimation using gaussian process regression," *IEEE Access*, 2019.
[131] S. Wang, D. Tao, and J. Yang, "Relative attribute svm+ learning for age estimation," *IEEE Trans. on Cybernetics*, pp. 827–839, 2016.
[132] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense sift," *Elsevier Information Fusion*, pp. 139–155, 2015.
[133] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. on IP*, pp. 5356–5368, 2015.
[134] M. T. B. Iqbal, M. Shoyaib, B. Ryu, M. Abdullah-Al-Wadud, and O. Chae, "Directional age-primitive pattern (dapp) for human age group recognition and age estimation," *IEEE Trans. on IFS*, pp. 2505–2517, 2017.
[135] K.-Y. Chang and C.-S. Chen, "A learning framework for age rank estimation based on face images with scattering transform," *IEEE Trans. on IP*, pp. 785–798, 2015.
[136] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *Proc. of ICML*, 1996, pp. 148–156.
[137] S. Glantz, "How to analyze rates and proportions," *Primer of Biostatistics*, pp. 126–178, 2005.
[138] G. Box and J. Ramirez, "Cumulative score charts," *Wiley Quality and Reliability Engineering International*, pp. 17–27, 1992.
[139] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. of IEEE ICCV*, 2015, pp. 1–9.
[140] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri *et al.*, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proc. of IEEE Conf. on CVPR Workshops*, 2016, pp. 1–8.
[141] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 341–345.
[142] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. on PAMI*, pp. 442–455, 2002.
[143] N. C. Ebner, M. Riediger, and U. Lindenberger, "Faces: A database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Springer BRM*, pp. 351–362, 2010.
[144] M. Minear and D. C. Park, "A lifespan database of adult facial stimuli," *Springer BRM*, pp. 630–633, 2004.
[145] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. of Springer ECCV*, 2014.
[146] S. ZHENG, "Visual image recognition system with object-level image representation," Ph.D. dissertation, 2012.
[147] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. on IFS*, pp. 2170–2179, 2014.
[148] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. of IEEE Conf. on CVPR*, 2009, pp. 256–263.
[149] K. Zhang, M. Sun, X. Han, X. Yuan, L. Guo, and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Trans. on CSVT*, 2017.

Vincenzo Carletti received the Ph.D. with European label in Computer Engineering in 2016, from the University of Salerno, Italy, where is currently Assistant Professor. His research activity is focused on exact and inexact graph matching for structural pattern recognition, applications in computer vision and pattern recognition.



Antonio Greco received the Ph.D. in Computer Science and Computer Engineering from the University of Salerno in 2018, where he is currently a Research Fellow. His research activity is focused on computer vision and machine learning techniques for video surveillance applications. He serves as a referee for many journals and international conferences.

Gennaro Percannella received the Ph.D. in electronic and computer engineering in 2002, from the University of Salerno, where he is currently associate professor of Computer Engineering. He authored more than 80 papers in international journals and conference proceedings in the field of computer vision and pattern recognition. His current interests include Pattern Recognition and Machine Learning in Artificial Vision. He is currently Associate Editor for IEEE Transactions on Medical Imaging.

Mario Vento received the Ph.D. in Computer Engineering in 1989 from the University of Napoli. He is currently a Full Professor of Artificial Vision, Machine Learning and Cognitive Robotics at the University of Salerno, Italy, where he is the Coordinator of the Artificial Vision Lab (MIVIA Lab). His research activities cover real-time video analysis and interpretation, cognitive robotics, classification techniques, exact and inexact graph matching, and learning methodologies for structural descriptions. Prof. Vento is a Fellow Scientist of the International Association Pattern Recognition (IAPR). He served as the Chairman of the IAPR Technical Committee 15 on Graph-Based Representation in Pattern Recognition from 2002 to 2006. He is currently Associate Editor of the Pattern Recognition journal.