Neural Networks 130 (2020) 238–252

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

# Feature fusion via Deep Random Forest for facial age estimation

O. Guehairia [a], A. Ouamane [b], F. Dornaika [c,d,*], A. Taleb-Ahmed [e]

[a] Laboratory of LESIA, University of Biskra, Biskra, Algeria
[b] Laboratory of LI3C, University of Biskra, Biskra, Algeria
[c] University of the Basque Country UPV/EHU, San Sebastian, Spain
[d] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[e] IEMN DOAE UMR CNRS 8520 Laboratory, Polytechnic University of Hauts-de-France, Valenciennes, France

ARTICLE INFO

*Article history:*
Received 29 September 2019
Received in revised form 27 May 2020
Accepted 6 July 2020
Available online 14 July 2020

*Keywords:*
Age estimation
Cascade of classification trees ensembles
Deep Random Forest
Face descriptors
Deep features

ABSTRACT

In the last few years, human age estimation from face images attracted the attention of many researchers in computer vision and machine learning fields. This is due to its numerous applications. In this paper, we propose a new architecture for age estimation based on facial images. It is mainly based on a cascade of classification trees ensembles, which are known recently as a Deep Random Forest. Our architecture is composed of two types of DRF. The first type extends and enhances the feature representation of a given facial descriptor. The second type operates on the fused form of all enhanced representations in order to provide a prediction for the age while taking into account the fuzziness property of the human age. While the proposed methodology is able to work with all kinds of image features, the face descriptors adopted in this work used off-the-shelf deep features allowing to retain both the rich deep features and the powerful enhancement and decision provided by the proposed architecture. Experiments conducted on six public databases prove the superiority of the proposed architecture over other state-of-the-art methods.



## 1. Introduction and motivation

Prior knowledge of Age improves, in general, the quality of interactions with individuals. In the digital era, the automatic age estimation becomes a hot research topic. Its valuable role in modern society grows year by year to become a salient one.

Automatic age estimation in face images has proven to be of a great value. It was thus adopted in many areas like Human Computer Interaction (HCI), security and management applications. Furthermore, many companies and advertisers rely on age categories when they recommend products and services to their clients.

Automatic age estimation can provide a range of utilities that can enhance surveillance systems and ease police investigation.

Automatic age estimation in facial images has some difficulties since the human aging process is very complicated (see Fig. 1). Indeed, there are many factors that make this problem challenging. These can be classified into two groups: intrinsic and extrinsic factors (Angulu, Tapamo, & Adewumi, 2018).

Intrinsic factors are related to health conditions. But, the extrinsic factors are external to the health conditions. These factors can be related to the living style and the working environment. Other challenges include the sample origin (i.e., society and region) which can influence the estimation precision.

Furthermore, the automatic age estimation should overcome the challenge of close cross age correlation, i.e., a man of 40 years of age looks almost the same as in his 39 and 41 years of age. The latter encouraged a lot of researchers to lean toward the regression solutions (Dornaika, Arganda-Carreras, & Belver, 2018; Niu, Zhou, Wang, Gao, & Hua, 2016; Shen, Guo, Wang, Zhao, Wang, & Yuille, 2018). Other works view age estimation either as a multi-class classification problem (Hu et al., 2017; Huerta, Fernández, Segura, Hernando, & Prati, 2015), or as a combination of regression and multi-class classification (Guo, Fu, Dyer, & Huang, 2008a, 2008b).

Automatic facial age estimation is an age approximation that stands on a numerical analysis of the person's facial image, either by a determination of the age group or by the exact scalar value of the age. Our work focuses on the estimation of the exact age value, in which a face image was automatically labeled with the estimated age through a learning process.

Pioneering works on facial age estimation faced a lot of challenges due to the scarcity of annotated image databases. The first work that showed interest in the subject was done by Kwon and da Vitoria Lobo (1999). They classified subjects into babies,

* Corresponding author at: University of the Basque Country UPV/EHU, San Sebastian, Spain.
*E-mail addresses:* oussama_guehairia@hotmail.fr (O. Guehairia), ouamaneabdealmalik@yahoo.fr (A. Ouamane), fdornaika@gmail.com (F. Dornaika), Abdelmalik.Taleb-Ahmed@uphf.fr (A. Taleb-Ahmed).

https://doi.org/10.1016/j.neunet.2020.07.006
0893-6080/

**Fig. 1.** Example of aging effect on a subject from FG-NET database. The appearance of the subject's face was affected considerably by aging.

adults and senior adults. Their method utilizes the analysis of skin wrinkles and craniofacial shape evolution. Lanitis, Taylor, and Cootes (2002), later, used the Active Appearance Model (AAM) by considering both face anthropometries and texture. Many methods exploited hand-crafted features such as: Local Binary Pattern (LBP) (Choi, Lee, Lee, Park, & Kim, 2011), Biologically Inspired Feature (BIF) (Guo, Mu, Fu, & Huang, 2009) and Haar-like features (Zhou, Georgescu, Zhou, & Comaniciu, 2005). Encouraged by the efficiency of gait representation (Lee, 1996) and Gait Energy Image (GEI) (Han & Bhanu, 2006), Gabor features were utilized in Lu and Tan (2010). A different approach can be found in Geng, Yin, and Zhou (2013). The authors considered each facial image as an instance linked with an age label distribution.

More recently, the use of deep neural networks in computer vision has made a lot of progresses. Its proven efficiency has become well known among specialists. In Shen et al. (2018), an end-to-end model for age estimation, named DRF (Deep Random Forest), has been presented. In this work, the authors connected the split nodes to the output of a fully connected layer of a convolutional neural network (CNN). They dealt with heterogeneous data by jointly learning input-dependent data partitions at the split nodes and data abstractions at the leaf nodes. In 2015, Chalearn organized the Looking At People (LAP) challenge with a focus on apparent age estimation (Escalera et al., 2015).

This paper is inspired by the work of Zhou and Feng (2017). These authors proposed a novel decision trees ensembles approach to a broad range of classification tasks. The performance and predictive accuracy of their proposed approach can be close to that of deep neural networks. They named it gcForest or Deep Random Forest. The main idea is the cascading structure generated by forests ensembles. Similarly to the deep neural networks, the learning is represented in the layer-by-layer processing of the raw features. The authors claimed that the training of these models is much easier than training deep CNNs. They applied their models to classic recognition problems.

To the best of our knowledge, the gcForest has not been exploited yet in Facial Age Estimation tasks. Table 1 shows the results of the direct application of gcForest on facial images, as explained in Zhou and Feng (2017), with six different databases. In this case, each year in the age scale is considered as a class. As it can be appreciated in that table, the obtained Mean Absolute Errors (in years) are worse than those obtained by many proposed approaches in the literature that are shown in Table 17. This suggests that gcForest does not suit the age estimation problems. Indeed, in gcForest structure the first layer of random forests take as input rawbrightness patches (obtained by sliding windows as it is shown in Fig. 4) in the original image. The patches are fed to the random forests to get an encoding. While this trick produced remarkable results for classic object recognition and classification problems, it failed to tackle the problem of age estimation in a satisfactory manner.

Thus, the above observation motivated us to propose a method that can be more adequate with the concerned problem. We introduce two main modifications: (i) we use image descriptors that can be either hand-crafted or provided by a pretrained deep CNN, and (ii) we propose a novel architecture for the deep random forests allowing to fuse any types of image descriptors. In the current work, although we use the DEX Chalearn pretrained CNN model to extract image's features, any other types of image features can be used and fused.

**Table 1**
Results of the direct use of gcForest on six public datasets. For each dataset, the Mean Absolute Error (years) is depicted using the standard evaluation protocols explained in Section 5.

| Method | Database | | | | | | |
|---|---|---|---|---|---|---|---|
| | FG-NET | PAL | MORPH Caucasian | LFW+ | APPA-REAL | FACES | average |
| gcForest | 8.43 | 5.49 | 6.78 | 9.43 | 7.89 | 5.86 | |

For each type of image features, we create an ensemble of Random Forests as in Zhou and Feng (2017) (that will be later called Deep Random Forest-Fusion (DRF-Fusion)) to extract a representation vector with more information. In a first phase, we carried out a fusion on the outputs of each type to get one fused representation vector. In a final stage, we applied a different form of DRF, namely fd-DRF (final Decision-Deep Random Forest) to the fused representation vector and generated the predicted age. In fd-DRF, a modified decision function is adopted by imitating some deep learning based models. This decision concerns the final output of the proposed architecture and uses the $N_{max}$ probabilities parameter (provided by the user), to select the $N_{max}$ ages having the largest probabilities. It then calculates the final prediction age through their arithmetic mean.

The main contributions of this paper are summarized as follows:

- A novel deep architecture for Random Forests that is applied to the facial age estimation problem.
- The architecture contains two main parts:
  1. The first part encodes and fuses the features of data representations.
  2. The second part is a Deep Random Forest structure that provides final age prediction using the largest $N_{max}$ probabilities.
- The proposed architecture allows the integration and fusion of different types of image descriptors.

To advocate our work, we empirically evaluated it against the state of the art methods using six challenging databases namely FG-NET, PAL, MORPH Caucasian, FACES, APPA-REAL and LFW+. In most of them we provided better results than those presented in the state of the art works in terms of Mean Absolute Error (MAE) and time complexity which prove the efficiency of our deep approach.

The structure of this paper is as follows. Section 2 presents some related work. Section 3 reviews the principles of Deep Random Forest. Section 4 details our proposed deep approach. Section 5 presents the experimental results. Section 6 analyzes the complexity and running time of the proposed approach. Section 7 concludes the paper. Table 2 summarizes the main acronyms and notations used in this paper.

## 2. Related work

Several works on facial age estimation had been proposed. The age estimation error in terms of mean absolute error (MAE) metric has been decreased by a massive margin from the appearance of this task. where, in Kwon and da Vitoria Lobo (1999), the authors proposed to predict age category in images where three

**Table 2**
Main acronyms and notations used in the paper.

| Acronym and notation | Description |
|---|---|
| AAM | Active Appearance Model |
| BIF | Bio Inspired Feature |
| CNN | Convolutional Neural Network |
| DEX | Deep Expectation |
| DCNN | Deep Convolutional Neural Network |
| DRF | Deep Random Forest |
| DMTL | Deep Multi Task Learning |
| ERT | Ensemble of Regression Trees |
| GEI | Gait Energy Image |
| HCI | Human Computer Interaction |
| LAP | Looking At People |
| LBP | Local Binary Pattern |
| LPQ | Local Phase Quantization |
| MAE | Mean Absolute Error |
| MSG | Multi Grained Scanning |
| LSDML | Label Sensitive Deep Metric Learning |
| ODLA | Ordinal Deep Learning Approach |
| SVM | Support Vector Machine |
| SVR | Support Vector Regression |
| GB | Gabor Wavelets |
| $C$ | Class number |
| $D$ | Original Feature vector size |
| $d$ | Number of features |
| $F$ | Number of forests |
| $V$ | Number of feature vectors |
| $n$ | Number of samples |
| $L$ | Number of layers (levels) |
| $n_{trees}$ | Number of trees |

categories were considered: babies, adults, and senior adults. In Lanitis et al. (2002), the authors used both anthropometries and texture as the main cues. The works described in Choi et al. (2011), Guo et al. (2009), and Zhou et al. (2005) used hand-crafted features, and achieved modest results.

In recent times, deep learning has bloomed significantly and gained popularity after being validated experimentally in a variety of fields in artificial intelligence, mainly in image recognition. Researchers have used Convolutional Neural Networks (CNN) extensively in different image-based tasks. The excellent performances in pose invariant face recognition tasks have led to its adoption in many demographic attributes estimation studies dealing with ethnicity, gender, and age estimation.

Authors in Shakeel and Lam (2019) have investigated the problem of age estimation through the deep learning techniques. Their diagnosis included three different kinds of formulations for the age estimation problem. They used the five most representative loss functions. In the work done by Huerta et al. (2015) a deep learning scheme have been proposed to upgrade the state of the art. A robust deep feature encoding-based discriminative model for age-invariant face recognition has been suggested in Xing, Li, Hu, Yuan, and Ling (2017). Researchers in this paper used a pre-trained Deep CNN model to extract high-level deep features. The extracted features were then encoded by learning a codebook, which converts each of the features into a discriminant S-dimensional code-word for image representation. They used canonical correlation analysis to fuse the pair of training features. For the recognition purposes, they uses a linear regression-based classifier. Authors in Abdulnabi, Wang, Lu, and Jia (2015) used a multitask CNN model to extract features corresponding to attributes in images before the application of the SVM models.

In other studies, an end-to-end solutions have also emerged in age estimation. For instance, the works described in Kontschieder, Fiterau, Criminisi, and Rota Bulo (2015) and Shen et al. (2018) relied on the use of a CNN and decision trees. Tree-based models treated as a chart-topping model due to its natural interpretability property. It is considered as a powerful method in decision tasks. Authors in Kontschieder et al. (2015) presented Deep Neural Decision Forests as a novel approach, which brings a DNN representation learning functionality together with classification trees by training them in an end-to-end manner. This model differs from conventional deep networks because the final predictions are provided by a decision forest. In Shen et al. (2018), an approach for age estimation under the name of Deep Regression Forest (DRFs) has been implemented. In this endeavor, researchers connected the split nodes of a decision tree to a fully connected layer of a CNN, and dealt with heterogeneous data by jointly learning input-dependent data partitions at the split nodes and data abstractions at the leaf nodes. A new deep ranking framework for age estimation has been proposed by Chen, Zhang, Dong, Le, and Rao (2017),in which they presented a model, that includes a set of basic CNNs, where each of these CNNs has been initialized with the pre-trained base CNN and fine-tuned with ordinal labels. In order to provide the final age prediction, authors aggregated the binary output of the basic CNNs.

Standing on the fact that age labels are chronologically correlated, the age estimation is an ordinal learning problem. In Liu, Lu, Feng, and Zhou (2019), the authors has presented a method to learn feature descriptors for face representation directly from raw pixels. Their method is termed Ordinal Deep Learning approach (ODFL). In ODFL, two criteria were enforced on the descriptors, which were learned at the top of the deep networks. These criteria are: topology-preserving ordinal relation, which was used to exploit the order of information in the learned feature space and age difference cost information.

In Liu, Lu, Feng, and Zhou (2018), the authors have also considered age estimation as an ordinal learning problem. They exploited the label correlation among face samples in the transformed subspace. Their approach was named Label Sensitive Deep Metric Learning (LSDML) for facial age estimation. LSDML differs from the recent deep metric methods (Hu, Lu, & Tan, 2014) and Hu, Lu, and Tan (2015), which used hand-crafted feature to feed deep network, LSDML leverages deep residual network to learn series of nonlinear features transformation, where the feature similarity is smoothly sensitive to the degree of age difference.

In Lou, Alnajar, Alvarez, Hu, and Gevers (2018), the authors have introduced a new graphical model where age is jointly learnt with expression, in comparison to expression-independent age estimation. The proposed model aims to learn the relationship, which ties the age and the expression, by including a latent layer between the age expression's labels and features.

The efforts in Han, Jain, Wang, Shan, and Chen (2018) have been focused on the attribute correlation and heterogeneity. The authors included an estimation of the multiple face attributes, in the form of deep multi-task learning approach in age estimation problem. They allowed shared feature learning among all attributes, and category-specific feature learning for heterogeneous attributes, by modeling all attributes in a single network.

Fusion strategies have been considered as a popular technique in biometrics. They were used in some facial age estimation works. The basic idea is to fuse decisions or features in a hierarchical learning system. A typical example is given by the Deep EXpectation (DEX) of apparent age method (Rothe, Timofte, & Van Gool, 2018). The authors detected the facial images first prior

to the extraction of CNN prediction from a network ensemble as a fusion method. In 2015 DEX won the apparent age Chalearn LAP competition.

Convinced by the advantages of the classification and regression trees, Zhou and Feng (2017) have proposed a new approach (named gcForest) for image classification.

Their method consists of a decision tree ensemble arranged in a cascade of layers where each layer (level) is composed of several random forests. The resulting performance is highly competitive to that of deep neural networks in a broad range of classification tasks (Zhou & Feng, 2017).

The differences between our proposed method and the existing works are that we use deep features from pre-trained models as input features, and we integrate the paradigm of deep learning that is similar to the deep neural networks, by cascading a simple machine learning tool that is provided by Random Forests (RF).

The work described in Zhou and Feng (2017) was dedicated to generic classification problems. The proposed classifier is a cascade of tree ensembles.

The major differences between the approach of Zhou and Feng (2017) and the CNN methods concern the hyper-parameters and the training process. Indeed, the approach of Zhou and Feng (2017) needs much fewer hyper-parameters than deep neural networks, and its model complexity can be automatically determined in a data-dependent way. The applications shown in Zhou and Feng (2017) targeted classical recognition problems.

The main similarity between our work and the work described in Zhou and Feng (2017) regards the use of Random Forests that generate feature representations. But, our proposed approach contains several novel modules which are different from Zhou and Feng (2017). These are as follows: the use of different deep feature vectors as input, the use of a mid-level fusion module, and the targeted application (facial age estimation), which is different from the classic classification tasks. In detail, we were inspired by the main idea of Zhou and Feng (2017) to create random forest ensembles with a cascade structure. However, this structure will be used twice in our proposed architecture. First, it is used for encoding and fusing the individual input features (i.e., generating a fused-representation). Second, the proposed architecture exploits the generated vector (fused-representation) for the final decision. More importantly, the work in Zhou and Feng (2017) does not contain a fusion module, and its input is composed of raw brightness of a sliding window (Multi-Grained-scanning). Random forest ensembles with cascade structures are then used for the final classification.

Table 3 summarizes the similarities and differences between our proposed method and the one presented in Zhou and Feng (2017).

It is worth noting that our proposed method and the work presented in Shen et al. (2018) are not similar. In Shen et al. (2018), the authors proposed a method where the split nodes of a regression tree are directly linked to a fully connected layer of a convolution neural network.

In their work, DRF refers to Deep Regression Forest and the deep concept is tied to the use of deep Convolutional Neural Networks.

They used trees model conditional probability over the ages where each leaf node can store a given trainable probability distribution.

They described how to learn a single differentiable regression tree. Also, they described how to learn an ensemble of trees to form a forest. In our work, for facial image features, we use adequate deep features (retrieved via some pre-trained CNNs) and integrate the deep concept by deploying a cascade of classic Random Forests ensembles. Furthermore, our architecture allows the fusion of different types of features. We created many

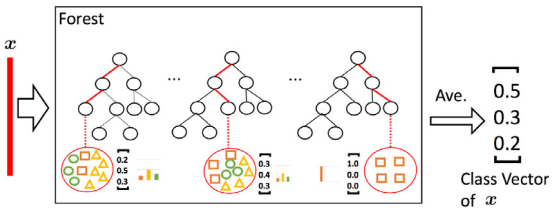


**Fig. 2.** Illustration of Random Forest classifier. Each class vector is generated by counting the percentage of different classes of training examples at the leaf node where the concerned instance falls and then averaging across all trees in the same forest (Zhou & Feng, 2017).

RFs with different settings. Those random forests (the ensemble) will be constructed layer by layer (level). The training in Shen et al. (2018) alternates between learning a CNN and learning a set of differentiable trees, which increases the computational complexity of the algorithm.

Moreover, the major difference between our proposed method and the methods that use Deep CNNs (e.g., Shen et al., 2018, Liu et al., 2019, Liu et al., 2018, Rothe et al., 2018) is the training time complexity. Indeed, our proposed method has less computational cost than that of the CNN based approaches.

Table 4 gives a brief overview of some recent works on age estimation.

## 3. Review of deep random forest

Some recent works used a Probabilistic Random Forest to tackle the age estimation problem Shen et al. (2018). They showed some interesting results. The method based on Random Forests RF Zhou and Feng (2017) was considered as a good competitor to Deep Neural Networks for classic classification problems. Deep Random Forests have many advantages such as the implementation simplicity and the reasonable time complexity associated with the training phase. These have encouraged us to explore and to adopt this strategy to the facial age estimation problem. To the best of our knowledge, our work, which is partially inspired by the idea presented in Zhou and Feng (2017), is the first work that addresses the age estimation task using deep Random Forests.

The specific characteristic that makes RF suitable for such applications is the reasonable cost of training and the robustness to over-fitting. Besides, RF has the advantage that it's few parameters are easy to set. We can use many RFs with different parameters. Diversity enhances the final performance. We can create a cascade structure with RF to create more layers like in deep neural networks where each layer can produce a piece of different information.

### 3.1. Random forest

Random Forest (RF) is a method that provides predictive models for classification and regression operations. RF uses binary decision trees that include CART trees proposed by Breiman (2001). The general idea behind the RF method is to generate several predictors before pooling their different predictions instead of trying to get an optimized procedure at once (see Fig. 2). More details about Random Forests can be found in Louppe (2015).

### 3.2. Deep random forest for classification

Relying on the advantages of random forest, Zhou and Feng (2017) have introduced a new approach that included many ensembles of random forests. By creating more than one level, the ensembles of random forests act as a cascade structure. In

**Table 3**
A comparison between our work and the method in Zhou and Feng (2017).

| Phases | Method in Zhou and Feng (2017) | Our method |
|---|---|---|
| Multi-Grained Scanning of raw images | ✓ | × |
| Ensembles of random forests | ✓ | ✓ |
| Cascade structure using random forests | ✓ | ✓ |
| Fusion representations using random forest ensembles. | × | ✓ |
| Final decision based on the max probabilities class | ✓ | ✓ |
| Final decision using the average of first largest probabilities | × | ✓ |
| Problem tackled | Classic classification | Facial age estimation |

**Table 4**
Overview of age estimation approaches.

| Publication | Approach | Face databases | Year |
|---|---|---|---|
| Guo et al. (2009) | Biological inspired feature BIF a pyramid of Gabor filters are used at all position of the input image | YGA FG-NET | 2009 |
| Choi et al. (2011) | Hierarchical classifier method based on both global and local facial features | FG-NET PAL BERC | 2010 |
| Geng et al. (2013) | Regards each face images as an instance associated with a label distribution | FG-NET MORPH Yeast Gene | 2013 |
| Huerta et al. (2015) | Fusing local feature appearance and use of deep learning scheme for age estimation | MORPH FRGC | 2015 |
| Pontes, Britto, Fookes, and Koerich (2016) | New framework integrate (AAM) (LBP),(GW) and (LPQ), age group classification using SVM and SVR for age estimation | MORPH II FG-NET | 2016 |
| Antipov, Baccouche, Berrani, and Dugelay (2017) | A fully synthetic age normalization algorithm based on Generative Adversarial Network and Local Manifold Adaption | LFW, IMBD-WIKI FG-NET | 2017 |
| Han et al. (2018) | Joint estimation of multiple heterogenous attributes | LFWA MORPH II Celeb A FotW LFW+ CLAP 2015 | 2017 |
| Liu et al. (2019) | Ordinal deep feature learning with two criterions on the descriptors (Topology-preserved and age difference) | FG-NET MORPH II | 2017 |
| Liu et al. (2018) | Deep Metric Learning to exploit label correlation among face sample in the transformed subspace | FACES FG-NET MORPH II Adience | 2018 |
| Lou et al. (2018) | Joint learning the age and expression by using latent layer between age expression | FACES LifeSpan NEMO | 2018 |
| Rothe et al. (2018) | Deep learned model from large data. Expected value formulation for age regression | MORPH FG-NET CACD | 2018 |
| Liu and Liua (2019) | Structure-based framework for facial age estimation under a constrained condition. Four stage fusion framework for facial age estimation 1- gender-recognition, 2-gender specific age groups, 3- age estimation within age, 4- fusion. groups | MORPH II FG-NET CLAP2016 | 2019 |
| Zeng, Ding, Wen, and Tao (2019). | A novel age encoding method (Soft Ranking). that simultaneously encodes both ordinal information and the correlation between adjacent ages. | MORPH II AgeDB ChalearnLAP 1015 ChalearnLAP 2016 | 2019 |

this paper, we will not distinguish between "level" and "layer". In this structure, each level is composed of an ensemble of random forests as illustrated in Fig. 3.

This structure was partially inspired by the layer-by-layer (or level-by-level) processing of a learning representation in the deep neural networks. Each level (or layer) of the Deep Random

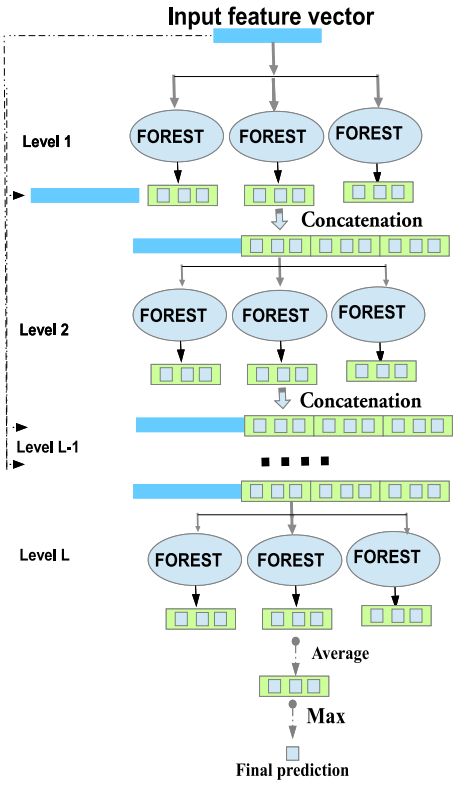


**Fig. 3.** Illustration of the deep random forest (DRF) structure where each level of the cascade receives feature information processed by its preceding level and outputs its processing results to the next level. Assume that each level of the cascade consists of three forests, and that there are three classes to predict. Thus, each forest will output a three-dimensional class vector, which is then concatenated for re-representation of the original input.

Forest is an ensemble of forests, precisely an ensemble of decision trees ensembles. The first level receives the feature vector as a given input, each forest of the same level will generate a class probability distribution as in Fig. 3. Suppose there are $C$ classes to predict, then a C-dimensional class vector will be the output of every single forest. The input vector for the next levels is obtained by concatenating the original input vector with the generated class vectors of each forest (resulting from the previous level). The dimension of the representation vector will be given by Eq. (1).

$$Dim = D + F \times C \tag{1}$$

where:

- $D$: the original feature size.
- $F$: the number of forests.
- $C$: the number of classes.

In the $L$-level (the last level), the RF generated class vectors will be averaged via arithmetic mean to produce the final class vector, the max value index of which, will be the prediction class.

$$FinalClass = \frac{1}{F}\sum_{f=1}^{F} Class(f) \tag{2}$$

where:

- *FinalClass*: the final probabilities class vector.
- *Class (f)*: the probabilities class vectors of a single forest $f$.
- $F$: the number of forests.

The capability of treating feature relationships using Deep neural networks encouraged the authors in Zhou and Feng (2017)

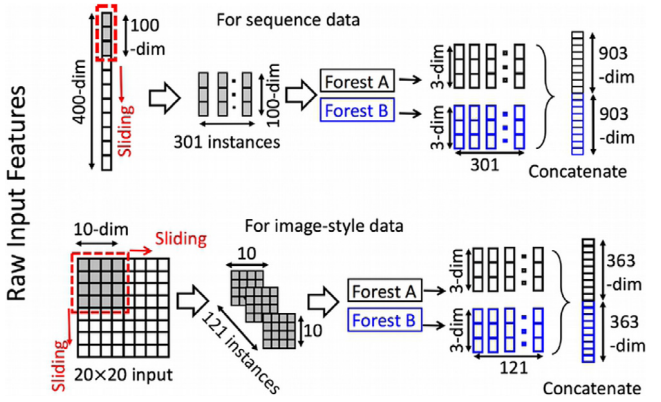


**Fig. 4.** Illustration of Multi-grained scanning in both case sequence data and image style data (Zhou & Feng, 2017).

to create a procedure for feature re-representation. This procedure aims to replace the convolution operation applied to the pixels of an image. It also aims to enhance the cascade forest and is named Multi-grained scanning (MGS) (see Fig. 4). The MGS uses a sliding window to scan the raw feature of sequence data or image style data, and it creates an ensemble of instances that have the same scan window size, those instances will be used to train two different types of forests to generate a class vector (for each) as elucidated before. The resulting class vectors will be concatenated to be transformed features.

## 4. Proposed approach

The DRF introduced in Zhou and Feng (2017) was proposed and used for the classic recognition and classification tasks. It was used for identity discrimination and object identification. The method in Zhou and Feng (2017) inspired us to develop an approach that will be applied to the problem of age estimation. The original method might face some difficulties if the human age nature is not taken into account. Indeed, the human age follows a uni-modal distribution, and the associated classes (if each year is considered as a class) can be fuzzy. We included an arithmetic function to improve the original final decision. This function influences the final decision to be more suitable to the nature of the human age. Although the DRF has proved its good results (Cao, Li, Ge, Wu, & Jiao, 2018; Zhang et al., 2018) in classic recognition problems, we think that there is still a room for better results. The estimation can be improved through the enhancement of either the prediction criteria, the initial input features, and the intermediate fusion scheme. We propose a method that uses all these three items and applies the resulting architecture to the problem of age estimation. Fig. 5 illustrates the overall architecture of our proposed model. This architecture is composed of two principal parts in addition to the pre-processing and feature extraction phases. The first part is represented by several individual DRF whose output is fused and handed out to the second part represented in the bloc fd-DRF (final decision-Deep Random Forest).

First, a process that serves as an enrichment of the initial input vectors should be added. We propose a different use of the original DRF that aims, in addition to final class prediction, to extract a vector through the previously explained concatenation of the original input vector with the random forest generated class vectors of a chosen level. The resulting vector will be larger and richer in information than the original one.

To enrich the input feature vector, even more, we took advantage of the efficiency of the feature fusion, which is considered

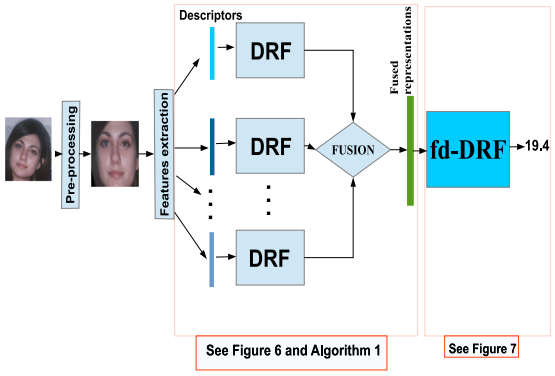


**Fig. 5.** Illustration of the overall architecture of our method.

as a popular technique in biometric that attracted the focus of researchers. We opted to fuse feature vectors obtained from several original descriptors in the DRF architecture arithmetically. Thus, for each type of original features, one DRF is designed to process it and to produce its DRF representation. Those original descriptors can be of any type: hand-crafted, deep features, and scanned windows like in gcForest (Zhou & Feng, 2017). Algorithm 1 and Fig. 6 show how the new input vector is computed.

$$Fused-representation = \frac{1}{V}\sum_{v=1}^{V} \mathbf{DRFF}(v) \tag{3}$$

where:

- **DRFF**(v) is the output of each individual DRF.
- *Fused-representation* is the output of the DRF-Fusion.
- $V$ is the number of original input feature vectors.

The proposed fusion (average of all DRF representations) assumes that the original features input in DRF has the same dimension. To overcome the case of different sizes of input vectors, we use zero paddings for the shortest vectors before the averaging process. The proposed fusion scheme aims to provide an averaged input vector for the final prediction process to minimize the influence of extreme values.

Algorithm 1 takes as input the face descriptor vectors $\mathbf{FV}(v)(v = 1, \ldots, V)$ (the feature vectors), the number of cascade levels (layers) $L$ (level or a layer contains several random forests), the number of input feature vectors $V$ and the number of forests $F$ in a given level. For the first level or if the number of levels is set to one, the input feature vectors to this level are the original feature vectors, else it will be another vector generated by the previous level as follows: Each forest $f \in 1, \ldots, F$ in the current level $level \in 1, \ldots, L$ will generate a probabilities class vector; those class vectors will be concatenated with the original input vector, as illustrated in Fig. 6. Each original feature vector has been encoded using the deep Random Forests. We call this generated code **DRFF**. The $V$ **DRFF** representations will be fused using the arithmetic average. Other fusion schemes can be adopted. In our work, we have tested two fusion schemes: the average and the concatenation. We have found that the average fusion has provided almost the same performance that is obtained with the concatenation (see section 5.3.4.), yet the average scheme provided much more compact representations.

Second, another process that enhances the prediction criteria is also integrated and is called fd-DRF. This one is similar to the cascade structure presented in the section of Random forest only in the last level (level $L$) which contains the final class vectors probabilities. We propose two ways for the final decision (Ages having the largest probability and Ages with $N_{max}$ highest probabilities). Ages with $N_{max}$ highest probabilities considered as new decision function, distinguished from the original work. Instead of picking the age having the highest probability, the new decision function takes in consideration the other ages having high probability (chosen in descending order) values. The new decision function uses the $N_{max}$ probabilities and their associated ages to produce the final age prediction (The mathematical process is the arithmetic mean of the $N_{max}$ ages having the highest probabilities where $N_{max}$ is a given parameter.). Fig. 7 illustrates the fd-DRF.

**Algorithm 1** DRF-Fusion

**Input:**
Face descriptors: $\mathbf{FV}_1, \mathbf{FV}_2, \ldots, \mathbf{FV}_V$;

Number of input feature vectors $V$;
Number of levels $L$;
Number of forests $F$.

**Output:**
Fusion: **Fused − representation**

**For** $v = 1 : V$

- **For** $level = 1 : L$
  **if** $(level = 1)$ **input** = **FV**$(v)$
  **Else input** = **current-input**
  **End if**
  - **For** each *FOREST* in $level\ f = 1 : F$
    **class** $(f)$ = generate-class-probabilities (*FOREST* $(f)$, **input**)
    **End for**

  **current-input** = concatenate (**FV**(**v**), **class**(1), ...., **class**($F$))
  **End for**

- **DRFF**$(v)$ = **current-input**

**End For**
**Fused-representation** = $\frac{1}{V}\sum_{v=1}^{V}$ **DRFF**$(v)$

## 5. Experiments

In this section, we will present the details of the algorithm implementation. We also provide a comparison against other similar studies. Our implementation contains many parts in which our main goal is to test various methods on a few given feature vectors. This allows us to assess the performance of the proposed model. We used the original Deep Random forest algorithm. We then compare its results with both the proposed method and the SVM classifier after the fusion phase. More explanation will be provided in the following subsections.

### 5.1. Implementation details

#### 5.1.1. Preprocessing

In this work, we localized the facial landmarks using the Ensemble of Regression Trees (ERT) algorithm (Kazemi & Sullivan, 2014) which is a robust and very efficient algorithm for facial landmarks localization. Facial landmarks help us to get eyes coordination, building on those points, we applied the face alignment, which is considered as an important step in image-based age estimation. After performing the alignment, the face region should be cropped (aligned face).

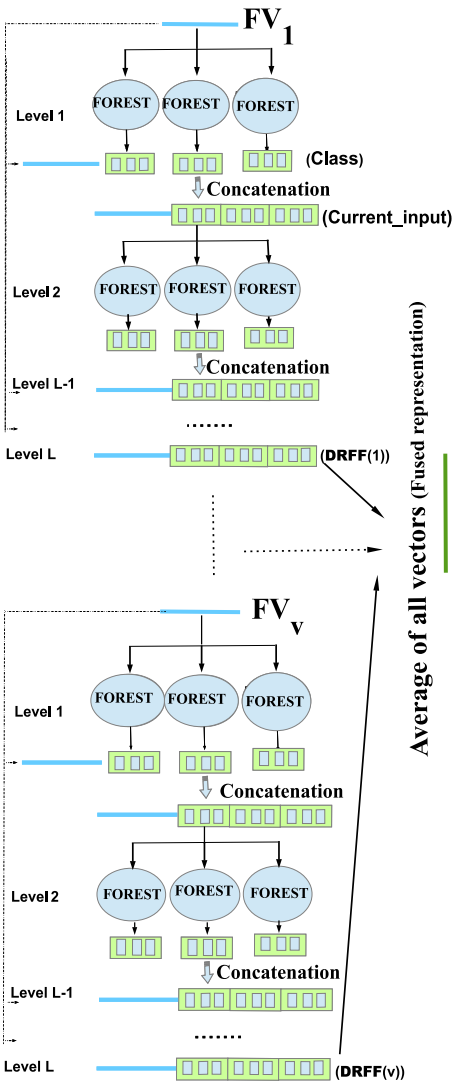


**Fig. 6.** An illustration of the DRF-Fusion scheme. Many DRFs with various input feature vectors are used to produce richer representations, which are later fused to obtain the Fused-representation.

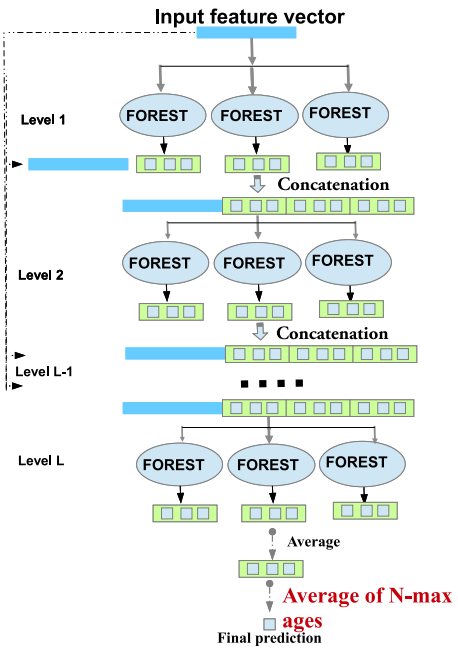


**Fig. 7.** Illustration of the final decision method fd-DRF.

### 5.1.2. Face features extraction

For face feature, we use the Deep Expectation (DEX) –Chalearn ICCV2015. DEX-Chalearn ICCV2015 is the winner (1st place) of the Chalearn LAP 2015 challenge on apparent age estimation. More than 500,000 images of celebrities from IMDb and Wikipedia labeled with age were assembled by authors of DEX-Chalearn to fine-tune the VGG-16 architecture used in DEX, the VGG-16 pre-trained on ImageNet for image classification. DEX-Chalearn is a powerful deep learning model for age estimation. It provides tools to generate deep features suitable for age characteristics due to the large data used to fine-tune it. We extract the last two fully connected layer vectors of DEX-Chalearn pre-trained model FC6 and FC7 of the input preprocessed images with a size of 224 × 224. The vectors FC6 and FC7 are later considered as the input features to the proposed architecture.

### 5.1.3. Parameter setting

Each level of the Deep Random Forest (DRF) (in both DRF-Fusion and fd-DRF) contains 10 forests. To encourage diversity, we used two types of forests. Thus, we used 5 completely random trees forests and 5 random forests. For both types, the five forests contain 500, 1000, 1500, 2000, 2500 trees, respectively. Selecting a feature at each tree node was randomly generated.

### 5.1.4. Evaluation metric

To evaluate the performance of the proposed age estimation method, we used the Mean Absolute Error (MAE). It is one of the most known indicators for age estimator performance evaluation in literature. MAE calculates the average of absolute error between the predicted and the ground truth ages. It is given by:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|p_t - g_t|, \tag{4}$$

where $n$ is the number of tested images, $p_t$ is the predicted age of image $t$, and $g_t$ is the ground-truth age of this image.

## 5.2. Benchmark databases

### 5.2.1. MORPH

This database contains images of 13,618 individuals (males and females). It contains more than 55000 unique images. Each facial image is annotated with chronological age. Ages are between 16 and 77 years. MORPH can be divided into more than ethnicity: African, European, and others. Following (Chang, Chen, & Hung, 2011; Chen, Gong, Xiang, & Change Loy, 2013; Rothe, Timofte, & Van Gool, 2016), we use the Caucasian subset, which contains 5492 images from the original MORPH database. We use the random split evaluation protocol on the Caucasian images. The 5,492 images are randomly partitioned to 80% training set and the other 20% testing set. It is repeated five times. The average of the five different splits will be the final performance.

### 5.2.2. FG-NET

This is a widely known database in age estimation. This database has a large variation in lighting conditions, pose and expression. FG-NET contains 1002 facial images associated with 82 individuals. Each individual has more than 10 photos taken at different ages. The FG-NET age range is from zero to 69. As in Shen et al. (2018) and Liu et al. (2018), we use the "Leave One Person Out" cross-validation on FG-NET. We leave one individual image out for testing and the other 81 individuals images for training.

**Table 5**
MAE (years) obtained with two different hand-crafted features (HOG and LBP) using DRF on the MORPH Caucasian dataset. We used $L_2$ normalization for LBP vector in the fusion part.

| Descriptor | Number of layers | |
|---|---|---|
| | 1 layer | 2 layers |
| LBP | 7.56 | 7.98 |
| HOG | 5.20 | 4.93 |
| LBP+HOG | 6.11 | 5.23 |
| LBP+HOG+FC6 | 5.24 | 5.13 |

**Table 6**
MAE obtained with two different hand-crafted feature using the DRF method with $N_{max} = 5$ ($N_{max}$ is the number of the ages having the highest probabilities) of MORPH Caucasian dataset. We used $L_2$ normalization for LBP vector in the fusion part.

| Descriptor | Number of layers | |
|---|---|---|
| | 1 layer | 2 layers |
| LBP | 6.23 | 6.12 |
| HOG | 5.83 | 5.27 |
| LBP+HOG | 5.12 | 5.98 |
| LBP+HOG+FC6 | 4.86 | 5.37 |

#### 5.2.3. PAL

The Predictive Aging Lab face is another database from Texas university. It contains 1046 frontal face images (430 males, 616 females). PAL contains faces with different expressions. We perform the random partition as in Dornaika et al. (2018) and Günay and Nabiyev (2016), where we randomly partition images in 80% training and the other 20% for testing. It is repeated five times. The average of the five different splits will be the final performance.

#### 5.2.4. LFW+

The MSU LFW+ database (Han et al., 2018) was created by extending the LFW database to study the joint attribute learning/estimation (age, gender, and race) from unconstrained face images "the images were taken in different positions and conditions and that what makes this database hard in the test". The extended LFW database (LFW+) contains 15,699 unconstrained face images of about 8,000 subjects. For each face image, three MTurk workers were asked to provide their estimates of age, gender, and race. The apparent age is determined as the average of the three estimates. we use the five-fold cross validation used in Han et al. (2018).

#### 5.2.5. FACES

The FACES database contains 2052 face images from 171 persons. For each person, there are 6 expressions: neutral, sad, disgust, fear, angry, and happy. For evaluation, we used the five random split protocol as in MORPH, Caucasian, and PAL. We conducted the experiments on image subset having the same facial expression (Liu et al., 2018; Shen et al., 2018).

#### 5.2.6. APPA-REAL

The APPA-REAL database (Agustsson et al., 2017) contains 7591 images. It has a default split into train, test, and validation. We used the same setting that is described in Agustsson et al. (2017). This database contains two types of age labels: Real Age and Apparent Age label. The Apparent Age labels are gathered from around 300,000 votes. On average, around 38 votes per each image, and this makes the average apparent age very stable.

### 5.3. Experimental results

#### 5.3.1. Performance evaluation

In this section, we will quantify the performance of the proposed method as a function of different factors. These include: (i) fused and non-fused features, (ii) type of features, (iii) number of layers, and (iv) number of highest probabilities.

We have used the hand-crafted features (LBP and HOG) as presented in Tables 5 and 6 with the MORPH Caucasian dataset. We relied on the same fusion strategy presented in our work. Tables 5 and 6 show the results of LBP, HOG, a fusion of HOG and LBP (LBP+HOG) and finally the fusion of HOG, LBP and FC6 feature vectors. The results presented in Table 5 correspond to the use of the age associated with the largest probability. The results in Table 6 were obtained using the mean of the five ages associated with the largest probabilities. The third row presents the fusion of the two hand-crafted features (LBP+HOG). In the fourth row, we included the deep feature vector FC6 of the DEX-Chalearn pre-trained model. The use of the deep feature aims to demonstrate its impact on the results when performing such tasks. As it can be seen, the MAEs depicted in Tables 5 and 6 are different. This is not surprising since the image feature type and the number of layers all affect the final performance. We remind that the proposed architecture is composed of two modules (fusing process and final decision process) that cannot be separated. The depicted architecture in Fig. 5 presents the general case where we have more than one type of features. The fusion module has several DRFs whose output are fused. The resulting fused vector feeds another DRF (named fd-DRF) for the final decision (see Fig. 5).

We emphasize that the presented comparisons in Tables 5 and 6 aim at studying several cases (fused features vs. individual features) as well as several types of features. The presented comparison aims to observe the advantages of the fusion process itself in the final stage (fd-DRF). This comparison elucidates what we can gain with such fusion processes. In some cases in which the decision is based on the highest probability (see Table 5), the fusion has not given an MAE that is better than the best one obtained by the individual features. This is the case where the LBP and HOG descriptors were used. The explanation can be as follows. Since the used LBP descriptor is not very relevant to the problem of age estimation, its fusion with HOG and FC6 features was not able to get a better result than what can be obtained by HOG alone. On the other hand, when the decision is based on the use of the highest probabilities (Table 6), the fusion of LBP and HOG gave better results than that of the individual features. Moreover, as it will be shown in Table 7, the fusion scheme of FC6 and FC7 features has not given the best results for some datasets. Based on the above observations, we can see clearly that the performance of the fusion depends on many factors that include the image feature type, the number of layers, the decision scheme, and the dataset. Thus, future work would investigate the fusion of many types of features as well as automatic feature weighting.

We have used the deep features FC6 and FC7 as input vectors to the DRFs (first part of Fig. 5). We perform two groups of experiments. In the first group, the DRF adopts one layer. In the second group, the DRF adopts two layers. The final output vector of this process is named $Fused - representation1$ when using DRF with one layer and $Fused - representation2$ when using the output DRF with two layers.

For the representations $Fused - representation1$ and $Fused - representation2$, we evaluated two solutions: (i) the first one is given by the DRF (i.e., the predicted age is estimated by the full architecture of Fig. 5), the second one uses the SVM multi-class classifier which is applied on the representation generated by the DRF module. We emphasize that, for the individual features, the SVM is applied to the output of the DRF module.

**Table 7**
MAE (years) obtained by the proposed architecture on seven datasets.

| Descriptor | Databases | | | | | | |
|---|---|---|---|---|---|---|---|
| | FG-NET | PAL | MORPH Caucasian | LFW+ | APPA-REAL Real Age | APPA-REAL Apparent Age | FACES |
| FC6 | 3.80 | 3.54 | 4.50 | 6.00 | 5.25 | 3.11 | 1.49 |
| FC7 | 4.00 | 4.80 | 4.26 | 6.16 | 5.71 | 3.60 | 2.77 |
| Fused-representation1 | 3.77 | 3.07 | 4.07 | 5.99 | 5.39 | 3.47 | 1.35 |
| FC6 | 3.84 | 3.68 | 5.80 | 6.12 | 5.96 | 3.12 | 1.30 |
| FC7 | 4.20 | 4.71 | 6.66 | 6.46 | 6.30 | 3.16 | 2.67 |
| Fused-representation2 | 3.90 | 3.09 | 6.11 | 6.11 | 6.52 | 3.57 | 1.85 |

**Table 8**
MAE (years) obtained by the proposed architecture on the FACES dataset.

| Descriptor | Face Expression | | | | | | |
|---|---|---|---|---|---|---|---|
| | Neutrality | Happiness | Disgust | Fear | Sadness | Angry | Average |
| FC6 | 1.10 | 1.26 | 1.92 | 1.52 | 1.34 | 1.82 | 1.49 |
| FC7 | 2.21 | 2.54 | 3.38 | 2.87 | 2.65 | 3.01 | 2.77 |
| Fused-representation1 | 0.86 | 1.15 | 1.73 | 1.47 | 1.18 | 1.71 | 1.35 |
| FC6 | 0.90 | 1.14 | 1.71 | 1.29 | 1.16 | 1.60 | 1.30 |
| FC7 | 2.11 | 2.43 | 3.25 | 2.83 | 2.51 | 2.91 | 2.67 |
| Fused-representation2 | 0.88 | 1.69 | 2.63 | 2.10 | 1.51 | 2.29 | 1.85 |

**Table 9**
MAE (years) obtained by the proposed architecture (without the fd-DRF) and the SVM multi class classification on seven datasets.

| Descriptor | Databases | | | | | | |
|---|---|---|---|---|---|---|---|
| | FG-NET | PAL | MORPH Caucasian | LFW+ | APPA-REAL Real Age | APPA-REAL Apparent Age | FACES |
| FC6 | 3.80 | 3.54 | 4.50 | 6.00 | 5.25 | 3.11 | 1.49 |
| FC7 | 4.00 | 4.80 | 4.26 | 6.16 | 5.71 | 3.60 | 2.77 |
| Fused-representation1 | 4.67 | 3.11 | 4.99 | 8.02 | 6.89 | 4.30 | 1.08 |
| FC6 | 3.84 | 3.68 | 5.80 | 6.12 | 5.96 | 3.12 | 1.30 |
| FC7 | 4.20 | 4.71 | 6.66 | 6.46 | 6.30 | 3.16 | 2.67 |
| Fused-representation2 | 4.33 | 2.99 | 4.05 | 5.95 | 5.31 | 3.20 | 1.03 |

**Table 10**
MAE (years) obtained by the proposed architecture (without the fd-DRF) and the SVM multi class classification on the FACES dataset.

| Descriptor | Face expression | | | | | | |
|---|---|---|---|---|---|---|---|
| | Neutrality | Happiness | Disgust | Fear | Sadness | Angry | Average |
| FC6 | 1.10 | 1.26 | 1.92 | 1.52 | 1.34 | 1.82 | 1.49 |
| FC7 | 2.21 | 2.54 | 3.38 | 2.87 | 2.65 | 3.01 | 2.77 |
| Fused-representation1 | 0.90 | 1.07 | 1.25 | 1.17 | 1.05 | 1.09 | 1.08 |
| FC6 | 0.90 | 1.14 | 1.71 | 1.29 | 1.16 | 1.60 | 1.30 |
| FC7 | 2.11 | 2.43 | 3.25 | 2.83 | 2.51 | 2.91 | 2.67 |
| Fused-representation2 | 0.85 | 1.01 | 1.20 | 1.06 | 0.96 | 1.15 | 1.03 |

We compare them with the original FC6 and FC7, which allowed us to evaluate the possible benefits offered by the fused representations generated by the proposed architecture.

Tables 7 and 8 summarize the results obtained with the proposed architecture using the deep features FC6 and FC7. Table 7 contains all used datasets and Table 8 presents the detailed results on the FACES dataset with all facial expressions. The first three rows of each table present the results obtained with DRFs adopting one layer (one level), where the remaining three rows present results obtained with DRFs adopting two layers (2 levels). In those two tables, we can see that the best results were obtained by *Fused — representation*1.

Tables 9 and 10 summarizes the results obtained by the SVM multi-class classifier. For the five datasets, the use of SVM with *Fused — representation*2 gives better results than the other representations. The SVM classifier with *Fused — representation*2 gives more accurate results than the SVM classifier that used the DRF representation of the individual FC6 or FC7 except for the FG-NET database.

Using SVM with the fused representations (provided by the first part of the proposed architecture) can reduce the final MAE in particular when two layers are used. This demonstrates the efficiency of the fusion method.

Tables 7 and 9 summarize the results of three types of comparisons: (i) individual feature vs. fused features; (ii) one layer vs. two layers for the individual DRFs, and (iii) SVM classifier on fused representations versus the proposed architecture.

The results have shown that whenever SVM is used the fusion has not improved the results compared with individual features (in particular in the case of one layer). On the other hand, when the proposed architecture is used, the fusion scheme adopting one layer for the individual DRFs has improved the performance with respect to the individual features.

Actually, the effectiveness of DRF depends on the number of layers. There is no evidence that by increasing the number of layers in the individual DRFs the final performance would necessarily increase. As in the original work that proposed the DRF for object recognition and classification the number of layers

**Table 11**
MAE obtained with the DRF using the highest probabilities method with Fused-representation1.

| Databases | $N_{max}$ Probabilities | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| FG-NET | 3.77 | 3.90 | 3.81 | 3.72 | 3.70 | 3.67 |
| PAL | 3.07 | 2.79 | 2.78 | 2.80 | 2.73 | 2.80 |
| MORPH Caucasian | 6.11 | 5.51 | 5.16 | 4.98 | 4.86 | 4.78 |
| MORPH Caucasian | 4.07 | 3.98 | 3.93 | 3.89 | 3.88 | 3.88 |
| LFW+ | 5.99 | 5.86 | 5.82 | 5.82 | 5.82 | 5.83 |
| APPA-REAL Real Age | 5.39 | 5.25 | 5.28 | 5.30 | 5.33 | 5.34 |
| APPA-REAL Apparent Age | 3.47 | 3.36 | 3.37 | 3.39 | 3.40 | 3.43 |
| FACES | 1.35 | 1.24 | 1.21 | 1.24 | 1.24 | 1.24 |

**Table 12**
MAE obtained with the DRF using the highest probabilities method on FACES database with Fused-representation1.

| Face expression | $N_{max}$ Probabilities | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Neutrality | 0.86 | 0.75 | 0.73 | 0.73 | 0.72 | 0.72 |
| Happiness | 1.15 | 1.02 | 0.98 | 1.01 | 1.03 | 1.02 |
| Disgust | 1.73 | 1.637 | 1.58 | 1.62 | 1.65 | 1.64 |
| Fear | 1.47 | 1.38 | 1.43 | 1.46 | 1.45 | 1.44 |
| Sadness | 1.18 | 1.07 | 1.00 | 0.98 | 1.02 | 1.04 |
| Angry | 1.71 | 1.60 | 1.56 | 1.57 | 1.57 | 1.57 |
| Average | 1.35 | 1.24 | 1.21 | 1.24 | 1.24 | 1.24 |

should be determined by a cross-validation scheme, and adopt the one that provides the best performance.

Thus, it is normal that results can be influenced by the number of layers and by the final classifier that output the predicted age (SVM or DRF). We recall that our method that we compared its results with the state of the art results (Table 17) is the fd-DRF (averaged predictions of several ages).

In the remainder of this section, we will present the results of the proposed architecture when the predicted age is set to the mean of ages having $N_{max}$ highest probabilities. We used the full proposed architecture with *Fused* − *representation*1 and *Fused* − *representation*2. The final age prediction is given as the average of $N_{max}$ ages that have the $N_{max}$ highest probabilities in the final output as explained in Fig. 7. We studied the effect of several values of $N_{max}$. Tables 11 and 12 illustrate the MAEs obtained by the DRF estimator on the vector *Fused*−*representation*1. Tables 13 and 14 illustrate the MAEs obtained by the fd-DRF estimator on the vector *Fused* − *representation*2. The results depicted in Tables 11 and 12 show the benefit of using $N_{max}$ ages with the highest probabilities. We can observe that the MAE decreases in all datasets as $N_{max}$ increases from one to six. In our work, the best results were, in general, obtained with the six highest probabilities.

Figs. 8.(a), (b), (c), and (d) illustrate graphically the MAE as a function of $N_{max}$ (the results were also depicted in Tables 11, 12, 13, and 14). Using the average of $N_{max}$ ages allowed the reduction of the final MAE by exploiting the strength of decision trees that can provide a distribution of the estimates. Thus, this scheme helped to get more accurate age prediction.

In Tables 13 and 14, the obtained MAEs are better than those obtained by many existing methods. In Tables 11 and 12, we can observe a constant decrease of the MAE as $N_{max}$ increases. However, in Tables 13 and 14, there is no constant decrease. Nevertheless, the averaging process shows that the optimal $N_{max}$ is either 5 or 6. For the PAL database, results obtained with *Fused-representation2* were better than those obtained with *Fused-representation1*.

#### 5.3.2. Fusion schemes

We have also used the concatenation method in the intermediate fusion stage to show the differences between various fusion strategies. The MAE obtained by the concatenation (concatenation of FC6 and FC7) vectors with DRF and SVM are shown in Tables 15 and 16, respectively. The *fused-representation1* and the *fused-representation2* refer to the fusion of FC6 and FC7 obtained by DRF with one level and two levels, respectively. Results indicate that the intermediate fusion by concatenation or by average gives almost the same results. However, using the average method leads to less memory space use and less computational cost since the size of the fused vectors is half that obtained with the concatenation.

### 5.4. Comparison with state-of-art methods

We compared our method, in term of the MAE, with the state-of-the-art methods depicted in Table 17. This table presents the results associated with MORPH Caucasian, FG-NET, PAL, LFW+, FACES and the APPA-REAL database with both label types (real age and apparent age). Table 18 shows the comparison of our method with state of the art with FACES database in detail of expression folds. Our work outperforms all the state of the arts in FACES and PAL with a large difference and in FG-NET too. Table 18 shows that our method is better than the compared methods in any face expression fold.

LFW and LFW+ exist in Chang et al. (2011) and Han et al. (2018). The work done by Chang et al. (2011), they used just the frontal face images (4211 images) of the LFW. The other

**Table 13**
MAE obtained with the DRF using the highest probabilities method with Fused-representation2.

| Databases | $N_{max}$ Probabilities | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| FG-NET | 3.90 | 3.80 | 3.79 | 3.78 | 3.79 | 3.86 |
| PAL | 3.09 | 2.86 | 2.85 | 2.86 | 2.85 | 2.86 |
| LFW+ | 6.11 | 5.96 | 5.92 | 5.90 | 5.89 | 5.89 |
| APPA-REAL Real Age | 6.52 | 6.45 | 6.45 | 6.50 | 6.60 | 6.63 |
| APPA-REAL Apparent Age | 3.57 | 3.50 | 3.53 | 3.53 | 3.56 | 3.60 |
| FACES | 1.85 | 1.63 | 1.59 | 1.58 | 1.56 | 1.55 |

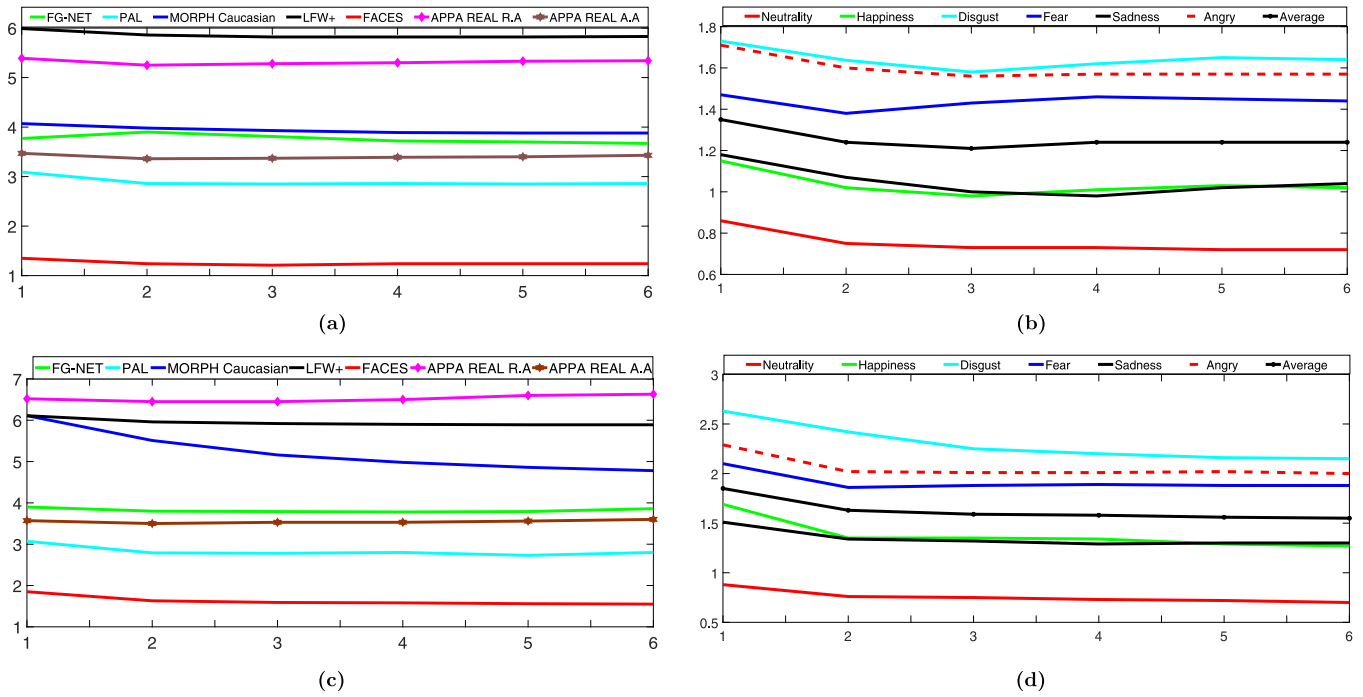


**Fig. 8.** Performance as a function of $N_{max}$. (a): MAE variation with DRF using the input Fused-representation1 on six databases,(b): Six subsets of the FACES dataset, (c) MAE variation with fd-DRF using the input Fused-representation1 on six databases, (d): Six subsets of the FACES dataset.

**Table 14**
MAE obtained with the DRF using the highest probabilities method on FACES dataset with Fused-representation2.

| Face expression | $N_{max}$ Probabilities | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Neutrality | 0.88 | 0.76 | 0.75 | 0.73 | 0.72 | 0.70 |
| Happiness | 1.69 | 1.35 | 1.35 | 1.34 | 1.29 | 1.27 |
| Disgust | 2.63 | 2.42 | 2.25 | 2.20 | 2.16 | 2.15 |
| Fear | 2.10 | 1.86 | 1.88 | 1.89 | 1.88 | 1.88 |
| Sadness | 1.51 | 1.34 | 1.32 | 1.29 | 1.30 | 1.30 |
| Angry | 2.29 | 2.02 | 2.01 | 2.01 | 2.02 | 2.00 |
| Average | 1.85 | 1.63 | 1.59 | 1.58 | 1.56 | 1.55 |

**Table 15**
MAE with DRF of fused representation vectors obtained using two fusion strategies.

| Scheme | Datasets | | | |
|---|---|---|---|---|
| | Concatenation method | | Average method | |
| | PAL | FG-NET | PAL | FG-NET |
| Fused-representation1 | 2.92 | 3.81 | 3.07 | 3.77 |
| Fused-representation2 | 3.35 | 4.11 | 3.09 | 3.90 |

**Table 16**
MAE with SVM of fused representation vectors obtained using two fusion strategies.

| Scheme | Datasets | | | |
|---|---|---|---|---|
| | Concatenation method | | Average method | |
| | PAL | FG-NET | PAL | FG-NET |
| Fused-representation1 | 3.08 | 4.73 | 3.11 | 4.67 |
| Fused-representation2 | 3.14 | 4.52 | 2.99 | 4.33 |

work (Han et al., 2018) authors create the LFW+ and they find the best results due to the many advantages offered by their Multi-task learning approach in which the age and other face attributes are simultaneously predicted. The training of their method requires more auxiliary attributes in addition to the age labels. They proposed Deep Multi-Task Learning (DMTL) network and they use a modified layer, with batch normalization (BN) layer inserted after each Convolution layer for shared feature learning. In Agustsson et al. (2017), the authors introduced the APPA-REAL database that contains both real and apparent age labels. We emphasize that the work of Agustsson et al. (2017) presented for each type of ages two solutions: (i) Fine-tuned DEX (DEX), and (ii) Fine-tuned DEX followed by a network-based residual estimation (Residual DEX). Experimental results in the two last columns, in Table 17, have shown that our proposed method outperforms the work of Agustsson et al. (2017) on both types of ages. It also outperforms the two solutions. For the APPA-REAL dataset, although the DEX CHALEARN was fine-tuned, the final performance is still inferior to that obtained by our proposed scheme.

It is obvious that our proposed method is neither a CNN-based approach nor a hand-crafted approach. This improvement is determined by various factors in our architecture.

Since our proposed method uses the pre-trained DEX-Chalearn network for extracting two types of image features, it would be interesting to compare the performance of our proposed approach with that obtained by the direct use of this network. Table 19 presents the results obtained from a direct use of the DEX-Chalearn network. The pre-trained CNN model is used for a direct age prediction of the face images. The comparison results show that our method provides better results than those obtained by the DEX-Chalearn estimator. Special attention can be drawn to PAL and FACES databases, where we get a significant difference in performance.

## 6. Complexity and running time

The computational complexity for training the proposed architecture which is composed of two parts based DRF will be split in

**Table 17**
Comparison of our method with some of state-of-the-art method using six datasets FG-NET, MORPH Caucasian, PAL, LFW+, FACES and APPA-REAL.

| Method | Database | | | | | | |
|---|---|---|---|---|---|---|---|
| | FG-NET | MORPH Caucasian | PAL | LFW+ | FACES | APPA-REAL Real Age | APPA-REAL Apparent Age |
| Human workers (Han, Otto, Liu, & Jain, 2015) | 4.70 | 6.30 | / | / | / | / | / |
| Rank (Chang, Chen, & Hung, 2010) | 5.79 | / | / | / | / | / | / |
| DIF (Han et al., 2015) | 4.80 | / | / | 7.8 | / | / | / |
| AGES (Geng, Zhou, & Smith-Miles, 2007) | 6.77 | 8.38 | / | / | / | / | / |
| IIS-LLD (Geng et al., 2013) | 5.77 | / | / | / | / | / | / |
| CPNN (Geng et al., 2013) | 4.76 | / | / | / | / | / | / |
| CA-SVR (Chen et al., 2013) | 4.67 | 5.88 | / | / | / | / | / |
| OHRank (Chang et al., 2011) | 4.48 | 6.07 | / | / | 7.14 | / | / |
| Pontes et al. (2016) | 4.50 | / | / | / | / | / | / |
| CAM (Luu, Seshadri, Savvides, Bui, & Suen, 2011) | 4.12 | / | / | / | / | / | / |
| Rothe et al. (2016) | 5.01 | 3.45 | / | / | / | / | / |
| Liu, Lu, Feng, and Zhou (2017) | 3.93 | / | / | / | / | / | / |
| LSDML (Liu et al., 2018) | 3.92 | / | / | / | 4.14 | / | / |
| Liu and Liua (2019) | 3.92 | / | / | / | / | / | / |
| DRFs (Shen et al., 2018) | 3.85 | **2.91** | / | / | / | / | / |
| Günay and Nabiyev (2016) | / | / | 5.40 | / | / | / | / |
| Nguyen, Cho, Shin, Bang, and Park (2014) | / | / | 6.50 | / | / | / | / |
| Luu et al. (2011) | / | / | 6.00 | / | / | / | / |
| Bekhouche, Ouafi, Dornaika, Taleb-Ahmed, and Hadid (2017) | / | / | 5.00 | / | / | / | / |
| Dornaika et al. (2018) | / | / | 3.79 | / | / | / | / |
| (DMTL) Han et al. (2018) | / | / | | **4.50** | / | / | / |
| Structured learning (Lou et al., 2018) | 3.89 | / | / | / | 7.40 | / | / |
| Agustsson et al. (2017) (DEX) | / | / | / | / | / | 5.46 | 4.08 |
| Agustsson et al. (2017) (Residual DEX) | / | / | / | / | / | 5.35 | 4.45 |
| Proposed method | **3.65** | 3.67 | **2.73** | 5.82 | **1.24** | **5.25** | **3.36** |

**Table 18**
Comparison of our method with some state-of-the art methods on FACES database detailed in facial expression.

| Method | Face expression | | | | | | |
|---|---|---|---|---|---|---|---|
| | Neutral | Happy | Disgust | Fearful | Sad | Angry | Average |
| BIF+OHRANK (Chang et al., 2011) | 5.16 | 7.64 | 8.31 | 7.00 | 6.87 | 7.87 | 7.14 |
| LBP+OHRANK (Chang et al., 2011) | 6.36 | 8.88 | 9.20 | 7.30 | 9.09 | 8.86 | 8.28 |
| BIF (Guo & Wang, 2012) | 9.50 | 10.70 | 13.26 | 12.65 | 10.78 | 13.26 | 11.69 |
| BIF+MFA (Guo & Wang, 2012) | 8.14 | 10.32 | 12.24 | 10.73 | 10.66 | 10.96 | 10.50 |
| CS-LBFL (Lu, Liong, & Zhou, 2015) | 5.06 | 6.53 | 7.15 | 6.32 | 6.27 | 6.94 | 6.46 |
| CS-LBMEL (Lu et al., 2015) | 4.84 | 5.85 | 5.70 | 6.10 | 4.98 | 5.50 | 5.49 |
| DEEPRANK (Yang, Lin, Chang, & Chen, 2013) | 5.99 | 7.12 | 8.15 | 6.35 | 7.77 | 6.68 | 7.01 |
| DEEPRANKER+(Yang et al., 2013) | 5.86 | 7.87 | 7.80 | 6.66 | 7.49 | 6.59 | 7.04 |
| LSDML (Liu et al., 2018) | 3.88 | 3.49 | 4.41 | 5.10 | 4.09 | 3.87 | 4.14 |
| MLSDML (Liu et al., 2018) | 3.83 | 3.11 | 4.16 | 5.01 | 3.67 | 3.16 | 3.82 |
| Structured learning (Lou et al., 2018) | 5.97 | 6.77 | 8.17 | 8.25 | 7.07 | 8.21 | 7.40 |
| Proposed method | **0.72** | **1.02** | **1.64** | **1.44** | **1.04** | **1.57** | **1.23** |

two parts. The first part corresponds to the computational complexity of the DRF Fusion part. This is $V^*F^*L^*\mathcal{O}(n^2 d\ n_{trees})$ where: $n$ is the number of face images, $d$ is the average number of features, $n_{trees}$ is the number of trees per forest, $V$ is the number of original input feature vectors, $L$ is the levels number in the DRF and $F$ is the number of forests in each level. The second part corresponds to the computational complexity of the fd-DRF part (in our work it contains one single level) is given by $F^*\mathcal{O}(n^2 d\ n_{trees})$. Finally,

**Table 19**
Comparison of our method with the results obtained using the well-known DEX-CHALEARN network. The comparison is carried out with six databases.

| Method | Database | | | | | | |
|---|---|---|---|---|---|---|---|
| | FG-NET | MORPH Caucasian | PAL | LFW+ | FACES (Average) | APPA-REAL Real Age | APPA-REAL Apparent Age |
| DEX-CHALEARN | 4.12 | 4.54 | 6.71 | 7.61 | 7.73 | 9.57 | 5.11 |
| Proposed method | **3.65** | **3.88** | **2.73** | **5.82** | **1.24** | **5.25** | **3.36** |

**Table 20**
Running time (in ms) of the different phases of the proposed approach (extraction and age prediction) for one face image. Two types of features were used FC6 and FC7. The architecture adopted one layer for DRF-Fusion.

| Phase | Pre-processing | Feature extraction | DRF-Fusion 1 layer | Prediction | Total time |
|---|---|---|---|---|---|
| Time ms | 11.6 | 370.1 | 10.6 | 0.35762 | 392.65 |

**Table 21**
Running time (in ms) of the different phases of the proposed approach for one face image. Two types of features were used FC6 and FC7. The architecture adopted two layers for DRF-Fusion.

| Phase | Pre-processing | Feature extraction | DRF-Fusion 2 layers | Prediction | Total time |
|---|---|---|---|---|---|
| Time (ms) | 11.6 | 370.1 | 30.1 | 0.3077 | 412.1 |

**Table 22**
Running time (in seconds) when the PAL dataset is used as a training set. It includes the feature extraction (using the pre-trained model DEX-Chalearn) and the learning phase of the DRF in both cases one layer and two layers.

| Phase | Feature extraction | Training (1 layer) | Training (2 layers) |
|---|---|---|---|
| Time (s) | 493.865 | 169.0506 | 186.947 |

for the training phase, the computational complexity of the total architecture is $V^*F^*L^*\mathcal{O}(n^2 d\ n_{trees}) + F * \mathcal{O}(n^2 d\ n_{trees})$.

For the test phase, the computational complexity is $V^*F^*L^*\mathcal{O}(d\ n_{trees}) + \mathcal{O}(d\ n_{trees})$.

Our experiments use a PC equipped with Intel(R) Core(TM) i7-4702MQ cpu @2.20 GHz and 8Go of RAM. Table 20 depicts the running times (in ms) associated with the extraction and age estimation for one face image. It offers a good guess for the total running time of the proposed method with any complete database. Table 20 also details the running time of every sub-process using DRF-Fusion with one layer.

As can be seen, the running time of the DRF-Fusion with one layer is 10.6 ms. We note that $V$ is equal to two corresponding to the use of the input vectors FC6 and FC7 in our experiments. Table 21 depicts the total running time when the DRF-Fusion used 2 layers. The running time associated with the DRF fusion increased due to the use of more than one layer, that influence the prediction running time. In Tables 20 and 21, feature extraction running time has the highest running time compared with other processes, especially when we compared it with the DRF-Fusion running time. The deep feature extraction influences the total time of the proposed architecture. That fact encourages to envision the use of other types of features that are much faster to extract.

This can be given by the hand-crafted features.

The main advantage of the proposed method is its training's cheap computational cost, even when deep features, like DEX-Chalearn, are used. The complexity of the training stage is lower than that of classic deep learning approaches. Table 22 shows the running time for the training phase using the overall PAL dataset which contains 1046 images. When dealing with such tasks, the computational tool of the training phase is lighter than the commonly used deep learning method.

## 7. Conclusion and future work

Throughout this work, we have proposed a new architecture for age estimation based on facial images. This architecture stands on a recently proposed classification method, currently known as Deep Random Forest. Our architecture is mainly built on a cascade of classification forests ensembles similar to those found in the DRF method and is composed of two types of DRFs. One seeking the enrichment of the feature representation of a given facial descriptor followed by a fusion of the enriched (high level) feature vectors. The other operates on the fused form of all of the enhanced representations in order to estimate the age. Experiments were conducted on different public databases: FG-NET, MORPH Caucasian, PAL, LFW+, FACES, and APPA-REAL. These experiments demonstrate the outperformance of the proposed architecture over many existing state-of-the-art methods. Some of the highlights of the work can be summed up in the following points:

- The reduction of the mean absolute error shows the efficiency of the DRF based extended feature along with the fusion representation, compared to the original feature.
- An even further reduction of the age error was obtained by using the concept of $N_{max}$ probabilities function that was a natural output of the proposed architecture. This concept has shown its superiority over the original decision process.
- The computational complexity of the training stage is cheaper than that of classic deep learning approaches.

Future work concerns the fusion enrichment phase using several input features of both (deep features and hand-crafted features). Other new proposals can investigate the decision stage. To this end, the concept of a weighted average of ages using the highest probabilities can be applied to all individual forests in the last layer.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work was funded by the Spanish Ministerio de Ciencia, Innovacion y Universidades, Programa Estatal de I+D+i Orientada a los Retos de la Sociedad, RTI2018-101045-B-C21.

## References


Abdulnabi, A. H., Wang, G., Lu, J., & Jia, K. (2015). Multi-task CNN model for attribute prediction. *IEEE Transactions on Multimedia*, *17*(11), 1949–1959.

Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., & Rothe, R. (2017). Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database. In *12th IEEE international conference and workshops on automatic face and gesture recognition*. IEEE.

Angulu, R., Tapamo, J. R., & Adewumi, A. O. (2018). Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, *2018*(1), 42.

Antipov, G., Baccouche, M., Berrani, S.-A., & Dugelay, J.-L. (2017). Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, *72*, 15–26.

Bekhouche, S., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., & Hadid, A. (2017). Pyramid multi-level features for facial demographic estimation. *Expert Systems with Applications*, *80*, 297–310.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Cao, X., Li, R., Ge, Y., Wu, B., & Jiao, L. (2018). Densely connected deep random forest for hyperspectral imagery classification. *International Journal of Remote Sensing*, 1–16.

Chang, K.-Y., Chen, C.-S., & Hung, Y.-P. (2010). A ranking approach for human ages estimation based on face images. In *2010 20th international conference on pattern recognition* (pp. 3396–3399). IEEE.

Chang, K.-Y., Chen, C.-S., & Hung, Y.-P. (2011). Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Computer vision and pattern recognition, 2011 Ieee conference on* (pp. 585–592). IEEE.

Chen, K., Gong, S., Xiang, T., & Change Loy, C. (2013). Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2467–2474).

Chen, S., Zhang, C., Dong, M., Le, J., & Rao, M. (2017). Using ranking-cnn for age estimation. In *The IEEE conference on computer vision and pattern recognition*.

Choi, S. E., Lee, Y. J., Lee, S. J., Park, K. R., & Kim, J. (2011). Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition*, *44*(6), 1262–1281.

Dornaika, F., Arganda-Carreras, I., & Belver, C. (2018). Age estimation in facial images through transfer learning. *Machine Vision and Applications*, 1–11.

Escalera, S., Fabian, J., Pardo, P., Baro, X., Gonzalez, J., Escalante, H. J., et al. (2015). Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 1–9).

Geng, X., Yin, C., & Zhou, Z.-H. (2013). Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(10), 2401–2412.

Geng, X., Zhou, Z.-H., & Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(12), 2234–2240.

Günay, A., & Nabiyev, V. V. (2016). Age estimation based on hybrid features of facial images. In *Information sciences and systems 2015* (pp. 295–304). Springer.

Guo, G., Fu, Y., Dyer, C. R., & Huang, T. S. (2008a). Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, *17*(7), 1178–1188.

Guo, G., Fu, Y., Dyer, C. R., & Huang, T. S. (2008b). A probabilistic fusion approach to human age prediction. In *Computer Vision and Pattern Recognition Workshops, IEEE Computer Society Conference on* (pp. 1–6). IEEE.

Guo, G., Mu, G., Fu, Y., & Huang, T. S. (2009). Human age estimation using bio-inspired features. In *Computer Vision and Pattern Recognition, IEEE Conference on* (pp. 112–119). IEEE.

Guo, G., & Wang, X. (2012). A study on human age estimation under facial expression changes. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2547–2553). IEEE.

Han, J., & Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2), 316–322.

Han, H., Jain, A. K., Wang, F., Shan, S., & Chen, X. (2018). Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2597–2609.

Han, H., Otto, C., Liu, X., & Jain, A. K. (2015). Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 1148–1161.

Hu, J., Lu, J., & Tan, Y.-P. (2014). Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1875–1882).

Hu, J., Lu, J., & Tan, Y.-P. (2015). Deep transfer metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 325–333).

Hu, Z., Wen, Y., Wang, J., Wang, M., Hong, R., & Yan, S. (2017). Facial age estimation with age difference. *IEEE Transactions on Image Processing*, *26*(7), 3087–3097.

Huerta, I., Fernández, C., Segura, C., Hernando, J., & Prati, A. (2015). A deep analysis on age estimation. *Pattern Recognition Letters*, *68*, 239–249.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1867–1874).

Kontschieder, P., Fiterau, M., Criminisi, A., & Rota Bulo, S. (2015). Deep neural decision forests. In *Proceedings of the IEEE international conference on computer vision* (pp. 1467–1475).

Kwon, Y. H., & da Vitoria Lobo, N. (1999). Age classification from facial images. *Computer Vision and Image Understanding*, *74*(1), 1–21.

Lanitis, A., Taylor, C. J., & Cootes, T. F. (2002). Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(4), 442–455.

Lee, T. S. (1996). Image representation using 2D gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (10), 959–971.

Liu, K.-H., & Liua, T.-J. (2019). A structure-based human facial age estimation framework under a constrained condition. *IEEE Transactions on Image Processing*.

Liu, H., Lu, J., Feng, J., & Zhou, J. (2017). Group-aware deep feature learning for facial age estimation. *Pattern Recognition*, *66*, 82–94.

Liu, H., Lu, J., Feng, J., & Zhou, J. (2018). Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security*, *13*(2), 292–305.

Liu, H., Lu, J., Feng, J., & Zhou, J. (2019). Ordinal deep learning for facial age estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(2), 486–501.

Lou, Z., Alnajar, F., Alvarez, J. M., Hu, N., & Gevers, T. (2018). Expression-invariant age estimation using structured learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(2), 365–375.

Louppe, G. (2015). *Understanding Random Forests from Theory to Practice* (Ph.D. thesis), Liège.

Lu, J., Liong, V. E., & Zhou, J. (2015). Cost-sensitive local binary feature learning for facial age estimation. *IEEE Transactions on Image Processing*, *24*(12), 5356–5368.

Lu, J., & Tan, Y.-P. (2010). Gait-based human age estimation. *IEEE Transactions on Information Forensics and Security*, *5*(4), 761–770.

Luu, K., Seshadri, K., Savvides, M., Bui, T. D., & Suen, C. Y. (2011). Contourlet appearance model for facial age estimation. In *2011 international joint conference on biometrics* (pp. 1–8). IEEE.

Nguyen, D. T., Cho, S. R., Shin, K. Y., Bang, J. W., & Park, K. R. (2014). Comparative study of human age estimation with or without preclassification of gender and facial expression. *The Scientific World Journal*, *2014*.

Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4920–4928.

Pontes, J. K., Britto, A. S., Fookes, C., & Koerich, A. L. (2016). A flexible hierarchical approach for facial age estimation based on multiple features. *Pattern Recognition*, *54*, 34–51.

Rothe, R., Timofte, R., & Van Gool, L. (2016). Some like it hot-visual guidance for preference prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5553–5561).

Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, *126*(2–4), 144–157.

Shakeel, M. S., & Lam, K.-M. (2019). Deep-feature encoding-based discriminative model for age-invariant face recognition. *Pattern Recognition*, *93*, 442–457.

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., & Yuille, A. (2018). Deep regression forests for age estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2304–2313). IEEE.

Xing, J., Li, K., Hu, W., Yuan, C., & Ling, H. (2017). Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, *66*, 106–116.

Yang, H.-F., Lin, B.-Y., Chang, K.-Y., & Chen, C.-S. (2013). Automatic age estimation from face images via deep ranking. *Networks*, *35*(8), 1872–1886.

Zeng, X., Ding, C., Wen, Y., & Tao, D. (2019). Soft-ranking label encoding for robust facial age estimation. arXiv preprint arXiv:1906.03625.

Zhang, Y.-L., Zhou, J., Zheng, W., Feng, J., Li, L., Liu, Z., et al. (2018). Distributed deep forest and its application to automatic detection of cash-out fraud. arXiv preprint arXiv:1805.04234.

Zhou, Z.-H., & Feng, J. (2017). Deep forest: Towards an alternative to deep neural networks. arXiv preprint arXiv:1702.08835.

Zhou, S. K., Georgescu, B., Zhou, X. S., & Comaniciu, D. (2005). Image based regression using boosting method. In *Computer Vision, Tenth IEEE International Conference on (vol. 1)* (pp. 541–548). IEEE.