Caroppo A, Leone A, Siciliano P. Comparison between deep learning models and traditional machine learning approaches for facial expression recognition in ageing adults. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 35(5): 1127–1146 Sept. 2020. DOI 10.1007/s11390-020-9665-4

Comparison Between Deep Learning Models and Traditional Machine Learning Approaches for Facial Expression Recognition in Ageing Adults

Andrea Caroppo, Alessandro Leone, and Pietro Siciliano

Institute for Microelectronics and Microsystems, National Research Council of Italy, Lecce 73100, Italy

E-mail: {andrea.caroppo, alessandro.leone}@cnr.it; pietro.siciliano@le.imm.cnr.it

Received April 23, 2019; revised March 6, 2020.

Abstract Facial expression recognition is one of the most active areas of research in computer vision since one of the non-verbal communication methods by which one understands the mood/mental state of a person is the expression of face. Thus, it has been used in various fields such as human-robot interaction, security, computer graphics animation, and ambient assistance. Nevertheless, it remains a challenging task since existing approaches lack generalizability and almost all studies ignore the effects of facial attributes, such as age, on expression recognition even though the research indicates that facial expression manifestation varies with age. Recently, a lot of progress has been made in this topic and great improvements in classification task were achieved with the emergence of deep learning methods. Such approaches have shown how hierarchies of features can be directly learned from original data, thus avoiding classical hand designed feature extraction methods that generally rely on manual operations with labelled data. However, research papers systematically exploring the performance of existing deep architectures for the task of classifying expression of ageing adults are absent in the literature. In the present work a tentative to try this gap is done considering the performance of three recent deep convolutional neural networks models (VGG-16, AlexNet and GoogLeNet/Inception V1) and evaluating it on four different benchmark datasets (FACES, Lifespan, CIFE, and FER2013) which also contain facial expressions performed by elderly subjects. As the baseline, and with the aim of making a comparison, two traditional machine learning approaches based on handcrafted features extraction process are evaluated on the same datasets. Carrying out an exhaustive and rigorous experimentation focused on the concept of "transfer learning", which consists of replacing the output level of the deep architectures considered with new output levels appropriate to the number of classes (facial expressions), and training three different classifiers (i.e., Random Forest, Support Vector Machine and Linear Regression), VGG-16 deep architecture in combination with Random Forest classifier was found to be the best in terms of accuracy for each dataset and for each considered age-group. Moreover, the experimentation stage showed that the deep learning approach significantly improves the baseline approaches considered, and the most noticeable improvement was obtained when considering facial expressions of ageing adults.

Keywords computer vision, deep learning, facial expression, machine learning, ageing adult

1 Introduction

The constant increase of the life expectancy and the consequent ageing phenomenon will inevitably produce in the next 20 years deep social changes that lead to the need of innovative services for elderly people, focused on maintaining independence, autonomy and, in general, improving the well-being and the quality of life of ageing adults[1]. It is obvious how in this context many potential applications, such as robotics, communications, security, medical and assistive technology, would benefit from the ability of automatically recognizing facial expression[1–3], because different facial expressions can reflect the mood, the emotions, and also the mental activities of an observed subject.

Facial expression recognition (FER) is related to

Regular Paper

[1] United Nations Programme on Ageing. The ageing of the world's population, December 2013. http://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2013.pdf, July 2018.

systems that aim to automatically analyze the facial movements and facial feature changes of visual information to recognize a facial expression. It is important to mention that FER is different from emotion recognition. The emotion recognition requires a higher level of knowledge. Despite the facial expression could indicate an emotion, the analysis of the emotion information like context, body gesture, voice, and cultural factors is also necessary [4].

A classical automatic facial expression analysis usually employs three main stages: face acquisition, facial data extraction and representation (feature extraction), and classification. Ekman's initial research [5] determined that there were six basic classes in FER: anger, disgust, fear, happiness, sadness, and surprise.

Proposed solutions for the classification of aforementioned facial expressions can be divided into two main categories: the first category includes the solutions that perform the classification by processing a set of consecutive images while the second one includes the approaches which carry out FER on each single image.

By working on image sequences much more information is available for the analysis. Usually, the neutral expression is used as a reference and some characteristics of facial traits are tracked over time in order to recognize the evolving expression. The major drawback of these approaches is the inherent assumption that the sequence content evolves from the neutral expression to another one that has to be recognized. This constraint strongly limits their use in real-world applications where the evolution of facial expressions is completely unpredictable. For this reason, the most attractive solutions are those performing facial expression recognition on a single image.

For static images various types of features might be used for the design of an FER system. Generally, they are divided into the following categories: geometric-based, appearance-based, and hybrid-based approaches.

More specifically, geometric-based features are able to depict the shape and locations of facial components such as mouth, nose, eyes, and brows using the geometric relationships between facial points to extract facial features. Three typical geometric feature based extraction methods are active shape models (ASM) [6], active appearance models (AAM) [7] and scale-invariant feature transform (SIFT) [8]. Appearance-based descriptors aim to use the whole-face or specific regions in a face image to reflect the underlying information in a face image. There are mainly three representative appearance-based feature extraction methods, i.e., Gabor Wavelet representation [9], Local Binary Patterns (LBP) [10] and Histogram of Oriented Gradient (HOG) [11]. Hybrid-based approaches combine the two previous types of features in order to enhance the system's performance and it might be achieved in either features extraction or classification level.

Geometric-based, appearance-based, and hybrid-based approaches have been widely used for the classification of facial expressions even if it is important to emphasize how all the aforementioned methodologies require a process of feature definition and extraction very daunting. Extracting geometric or appearance-based features usually requests an accurate feature point detection technique and generally this is difficult to implement in real-world complex background. In addition, this category of methodologies easily ignores the changes in skin texture such as wrinkles and furrows that are usually accentuated by the age of the subject. Moreover, the task often expects the development and subsequent analysis of complex models with a further process of fine-tuning of several parameters, which nonetheless can show large variances depending on individual characteristics of the subject that performs facial expressions. Last but not least recent studies have pointed out that classical approaches used for the classification of facial expression are not performing well when used in real contexts where face pose and lighting conditions are broadly different from the ideal ones used to capture the face images within the benchmark datasets.

Among the factors that make FER very difficult, the most discriminating one is the age [12, 13]. In particular, expressions of older individuals appeared harder to decode, owing to age-related structural changes in the face, which supports the notion that the wrinkles and folds in older faces actually resemble emotions. Consequently, state-of-the-art approaches based on handcrafted features extraction may be inadequate for the classification of FER performed by ageing adults.

It seems therefore very important to analyze automatic systems that make the recognition of facial expressions of the ageing adults more efficient, considering that facial expressions of elderly, as highlighted above, are broadly different from those of young or middle-aged for a number of reasons. For example, in [14] researchers found that the expressions of ageing adults (women in this case) were more telegraphic in the sense that their expressive behaviours tended to involve fewer regions of the face, and yet more complex in that they used blended or mixed expressions when recounting emotional events. These changes, in part, account for why the facial expressions of ageing adults are more difficult to read. Another study showed that when emotional memories were prompted and subjects asked to relate their experiences, ageing adults were more facially expressive in terms of the frequency of emotional expressions than younger individuals across a range of emotions, as detected by an objective facial affect coding system[15]. One of the other changes that comes with age, making an ageing facial expression difficult to recognize, involves the wrinkling of the facial skin and the sag of facial musculature. Of course, part of this is due to biological aspects of ageing, but individual differences also appear linked to personality process, as demonstrated in [16].

To the best of our knowledge, only few studies in literature address the problem of FER in ageing adults. In [12] the authors performed a computational study within and across different age groups and compared the FER accuracies, founding that the recognition rate is influenced significantly by human ageing. The major issue of this work is related to the feature extraction step, in fact they manually labelled the facial fiducial points, and given these points, Gabor filters are used to extract features for subsequent FER. Consequently, this process is inapplicable in the application context under consideration, where the objective is to provide new technologies able to function automatically and without human intervention.

On the other hand, the application described in [17] recognizes emotions of ageing adults using an active shape model[6] for feature extraction. To train the model the authors of [17] employed three benchmark datasets that do not contain adult faces getting an average accuracy of 82.7% on the same datasets. Tests performed on old faces acquired with the webcam reached an average accuracy of 79.2%, without any verification of how the approach works for example on a benchmark dataset with older faces.

Analyzing the results achieved it seems appropriate to investigate new methodologies which must make the feature extraction process less difficult, while at the same time strengthening the classification of facial expressions.

Recently, a viable alternative to the traditional feature design approaches has been represented by deep learning (DL) algorithms which leads directly to automated feature learning[18]. Research using DL techniques could make better representations and create innovative models to learn these representations from unlabelled data. These approaches became computationally feasible thanks to the availability of powerful GPU processors, allowing high-performance numerical computation in graphics cards. Some of the DL techniques like convolutional neural networks (CNNs), deep Boltzmann machine, deep belief networks and stacked autoencoders are applied to practical applications like pattern analysis, audio recognition, computer vision and image recognition where they produce challenging results on various tasks[19]. Recently, Li *et al.*[20] presented a complete survey of aforementioned DL techniques specifically for the FER topic. Ginne *et al.*[21] also presented a survey on CNN-based FER techniques. Much of the research on FER based on deep convolution networks has focused on improving the accuracy in expression recognition. It is important to note how at the same accuracy level, a smaller CNN architecture can provide more efficient distributed training, a smaller parameter model, and more suitability for deployment on memory-constrained devices, limiting costs and consequently favouring a wider distribution.

As evidenced by their use in a number of state-of-the-art algorithms, CNNs have worked very well for FER. In particular, a kind of CNN architecture with few layers was the winners of FER competitions[22], particularly previous years' EmotiW challenge[23, 24].

In this paper, three state-of-the-art deep models based on CNN, i.e., VGG-16[25], AlexNet[26], and GoogLeNet/Inception V1[27], have been introduced with the aim of verifying the influence of age in the recognition of facial expressions. The deep learning approaches have been also compared with two traditional approaches selected among the most promising ones and effective present in the literature. Moreover, in order to tackle the problem of the overfitting, which occurs in the case of datasets containing few images, this work proposes also in the pre-processing step, standard methods for data generation in a synthetic way (techniques indicated in the literature as "data augmentation") to cope with the limitation inherent the amount of data into two of four datasets employed.

The structure of the paper is as follows. Subsection 2.1 reports some details about the implemented pipeline for FER in ageing adults, emphasizing theoretical details for pre-processing steps. Subsection 2.2 describes also the deep architectures and the algorithmic procedures used in this work in order to adapt the aforementioned models to the problem of FER in ageing adults. Subsection 2.3 reports, on the other hand, details about two traditional machine learning approaches used for comparison. Section 3 presents datasets, experimental procedures, and results obtained, while discussion and conclusion are summarized in Section 4.

2 Methods

Fig.1 shows the structure of our FER system. First, the implemented pipeline performs a pre-processing task on the input images (data augmentation, face detection, cropping & resizing, normalization). Once the images are pre-processed, they can be used either to train the implemented deep networks or to extract handcrafted features (both geometric and appearance-based).

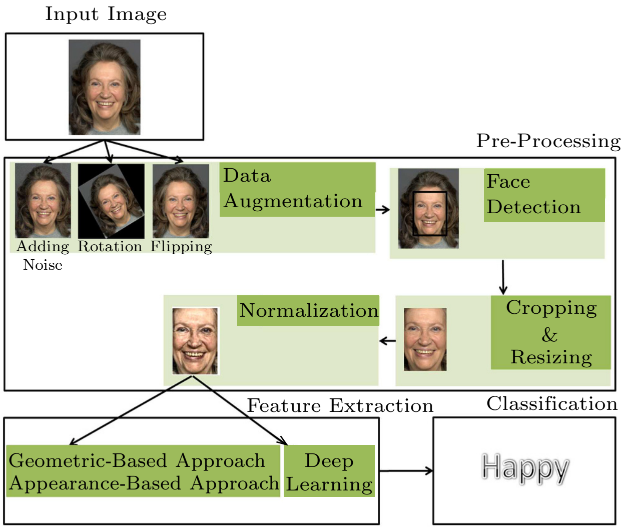


Fig.1. Pipeline of the proposed system. First a pre-processing task on the input images was performed. The obtained normalized face image is used to train the deep neural network architectures. Moreover, both geometrical and appearance-based features are extracted from the normalized image. Finally, each image is classified by associating it with a label of most probably facial expression.

2.1 Pre-Processing

Here are some details about the blocks that perform the pre-processing algorithmic procedure, whereas Subsection 2.2 and Subsection 2.3 illustrate the theoretical details of the deep architectures and the two classical machine learning approaches used for comparison.

It is well known that one of the main problems of deep learning methods is that they need a lot of data in the training phase to perform this task properly. Therefore before training the CNN-based deep learning models, it was considered appropriate to augment the data with various transformations for generating various small changes in appearances and poses. The number of available images has been increased with three data augmentation strategies. The first strategy is to use flip augmentation, mirroring images about the y-axis producing two samples from each image. The second strategy is to change the lighting condition of the images. In this work lighting condition is varied by adding Gaussian noise in the available face images. The last strategy consists in rotating the images of a specific angle. Consequently each facial image has been rotated through 7 angles randomly generated in the range $[-30°, +30°]$ with respect to the y-axis. Summarily, starting from each image present in the datasets, and through the combination of the previously described data augmentation techniques, 32 facial images have been generated.

The next step consists in the automatic detection of the facial region. Here, the facial region is automatically identified on the original image by means of the Viola-Jones face detector[28]. Once the face has been detected by the algorithm introduced in [28], a simple routine was written in order to crop the face image. This is achieved by detecting the coordinates of the top-left corner, the height and the width of the face enclosing rectangle, removing in this way all background information and image patches that are not related to the expression. Since the facial region could be of different sizes after cropping, in order to remove the variation in face size and keep the facial parts in the same pixel space, the algorithmic pipeline provides both a down-sampling step and an increasing resolution step that generate face images with the specific size required by various proposed architectures. For the down-sampling a simple linear interpolation was used, whereas a nearest-neighbour interpolation was implemented in order to increase the size of the facial images. Finally, since the image brightness and contrast could vary even in images that represent the same facial expression performed by the same subject, an intensity normalization procedure was applied in order to reduce these issues. Generally histogram equalization is applied to enhance the contrast of the image by transforming the image intensity values since images which have been contrast enhanced are easier to recognize and classify. However, the noise can also be amplified by the histogram equalization when enhancing the contrast of the image through a transformation of its intensity value since a number of pixels fall inside the same grey-level range. Therefore, instead of applying the histogram equalization, in this work the method introduced in [29] called “contrast limited adaptive histogram equalization” (CLAHE) was used. This algorithm is an improvement of the histogram equalization algorithm and essentially consists in the division of the original image into contextual regions, where histogram equalization was made on each of these sub-regions. These sub-regions are called tiles. The neighbouring tiles are combined by using a bilinear interpolation to eliminate artificially induced boundaries. This could give much better contrast and provide accurate results.

2.2 Description of CNN-Based Deep Learning Approaches

CNN is a type of deep learning model for processing data that has a grid pattern, such as images, which is inspired by the organization of animal visual cortex[30] and designed to automatically and adaptively learn spatial hierarchies of features, from low-level to high-level patterns. CNN is a mathematical construct that is typically composed of three types of layers (or building blocks): convolution, pooling, and fully connected layers.

The first two, convolution and pooling layers, perform feature extraction, whereas the third, a fully connected layer, maps the extracted features into final output, such as classification. A typical implementation of CNN for FER encloses three learning stages in just one framework. The learning stages are: 1) feature learning, 2) feature selection, and 3) classifier construction. Moreover, two main phases are provided: training and test. During training, the network acquires facial images (the normalized image output of the pre-processing step), together with the respective expression labels, and learns a set of weights.

The process of optimizing parameters (i.e., training) is performed with the purpose to minimize the difference between outputs and ground truth labels through an optimization algorithm. Generally the order of presentation of the facial images can influence the classification performance. Consequently to avoid this problem, usually a group of images are selected and separated for a validation procedure, useful to choose the final best set of weights out of a set of trainings performed with samples presented in different orders. After, in the test step, the architecture receives a face image and outputs the predicted expression by using the final network weights learned during training.

Creating a CNN from scratch is not an easy task. Therefore, in order to save ourselves from this overhead, in the present work the concept of “transfer learning” was adopted[31]. Transfer learning is a common and recently strategy to train a network also on a small dataset, (which is one of the main problems in the case of recognition of facial expressions of the elderly) where a network is pre-trained on an extremely large dataset, such as ImageNet[26], which contains 1.4 million images with 1 000 classes, and then reused and applied to the given task of interest. The underlying assumption of transfer learning is that generic features learned on a large enough dataset can be shared among seemingly disparate datasets. This portability of learned generic features is a unique advantage of deep learning that makes itself useful in various domain tasks with small datasets.

As written above, given the complexity of the task of recognizing facial expressions (especially those performed by ageing adults), the solution adopted in the present study is to use pre-trained deep learning models avoiding unnecessary efforts to define network tuning parameters, since the architectures introduced below have been already trained with lots of images. Consequently, they can be directly used for our FER task.

2.2.1 VGG-16

VGG-16 model (developed by the Visual Geometry Group (VGG) at the University of Oxford) can be considered as a milestone among the numerous deep CNN models developed in the last years[25]. It was pre-trained on the ImageNet database[26] in order to extract features from images that can distinguish one image class from another. Numerous recent publications have shown that VGG-16 achieves excellent performances even when applied to image recognition and classification datasets in other domains.

The aforementioned deep architecture, involving 144 million parameters, contains 13 convolutional layers with very small receptive fields 3×3 and five max-pooling layers of size 2×2 for carrying out spatial pooling, followed by three fully-connected layers, with the final layer as the soft-max layer. Rectification nonlinearity (ReLu) activation is applied to all hidden layers. The model also uses dropout regularization in the fully-connected layers. A schematic of the VGG-16 architecture trained is shown in Fig.2. When the fully-connected classifier (or the “bottleneck” layer) is removed from the pre-trained VGG16 network, it can be used as a deep feature generator for producing semantic image vectors for the images containing facial expressions. These semantic image vectors can then be

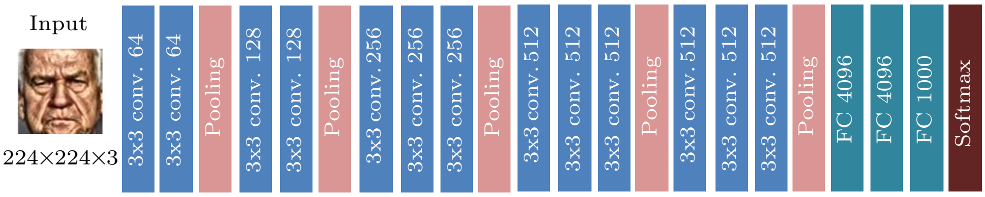


Fig.2. Schematic of the VGG-16 deep convolutional neural network (DCNN) architecture trained on the ImageNet database. The network is 16-layer deep and can classify images into 1 000 object categories. As a result, the network has learned rich feature representations for a wide range of images. The network has an image input size of 224 × 224 pixels.

trained and tested using another classifier for predicting the labels that identify expressions.

2.2.2 AlexNet

AlexNet is a popular deep network that was known for its successful demonstration in the ImageNet challenge[32]. It was designed by Alex Krichevsky, which at the time was not attempted for such challenges. The results achieved in 2012 had surprised the research area after it achieved 16% error, which was at 25% in 2011. It has produced a significant increase in performance compared with other hand-crafted techniques, and the error decrease trend was then followed yearly in the ImageNet challenges by other deep networks. AlexNet consists of five convolutional layers and three fully connected layers (a schematic representation of the architecture is depicted in Fig.3). Multiple convolutional kernels (filters) extract interesting features in an image. In a single convolutional layer, there are usually many kernels of the same size. For example, the first convolutional layer of AlexNet contains 96 kernels of size 11 × 11 × 3.

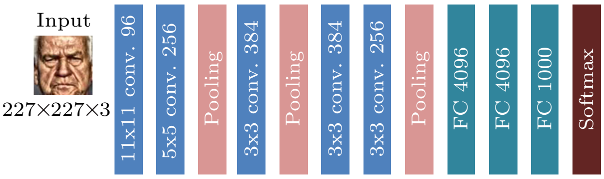


Fig.3. Architecture of AlexNet. It consists of five convolutional layers, some of which are followed by maximum pooling layers and then three fully-connected layers and finally a 1 000-way softmax classifier.

The width and the height of the kernel are usually the same and the depth is the same with the number of channels. Beyond the convolutional layers, the original architecture has three layers of fully connected (FC) neurons, which perform the classification. Associated with the last layer there is a soft-max structure. In the architecture implemented in the present work, the five convolutional layers extract characteristics from pre-processed facial images and the outputs of the last convolutional layer are used as inputs for a generic classifier.

Also, in AlexNet implementation proposed in the present work, the last layers (fully connected layers and softmax) are not used.

2.2.3 GoogLeNet/Inception V1

GoogleNet/Inception V1 is the earliest version of GoogleNet, appearing in 2014[27]. It is known that the performance of the "deep" networks improves with the growth of the depth and the width of the network. The disadvantage in this case is to be found in the high number of parameters. However, so many parameters will not only cause overfitting but also increase the computation. In [27] the authors demonstrated that the fundamental way to solve these two drawbacks is to convert the connections, even the convolutions, to a sparse set of connections. For non-uniform sparse data, the computational efficiency of computer software and hardware is very poor, thereby determining an approach that not only keeps the sparsity of the network, but also permits the high computational performance associated with dense matrices, is a key issue. A large number of papers show that the computing performance can be improved by clustering the sparse matrix into dense submatrices. Inspired by these methods, the inception module was designed to realize the above ideas.

GoogLeNet starts with a sequential chain of convolution, pooling, and local response normalization operations, in a similar fashion to previous CNN models, such as AlexNet. Later papers on the inception architectures refer to this initial segment as the "stem". The stem stands in contrast to the rest of the GoogLeNet architecture, which is primarily made up of what are referred to as "inception" modules. The basic building block of GoogLeNet, the inception module, is a set of convolutions and poolings at different scales, each done in parallel and concatenated together.

Fig.4(a) shows the initial version of the inception module. Different sizes of convolutions mean different

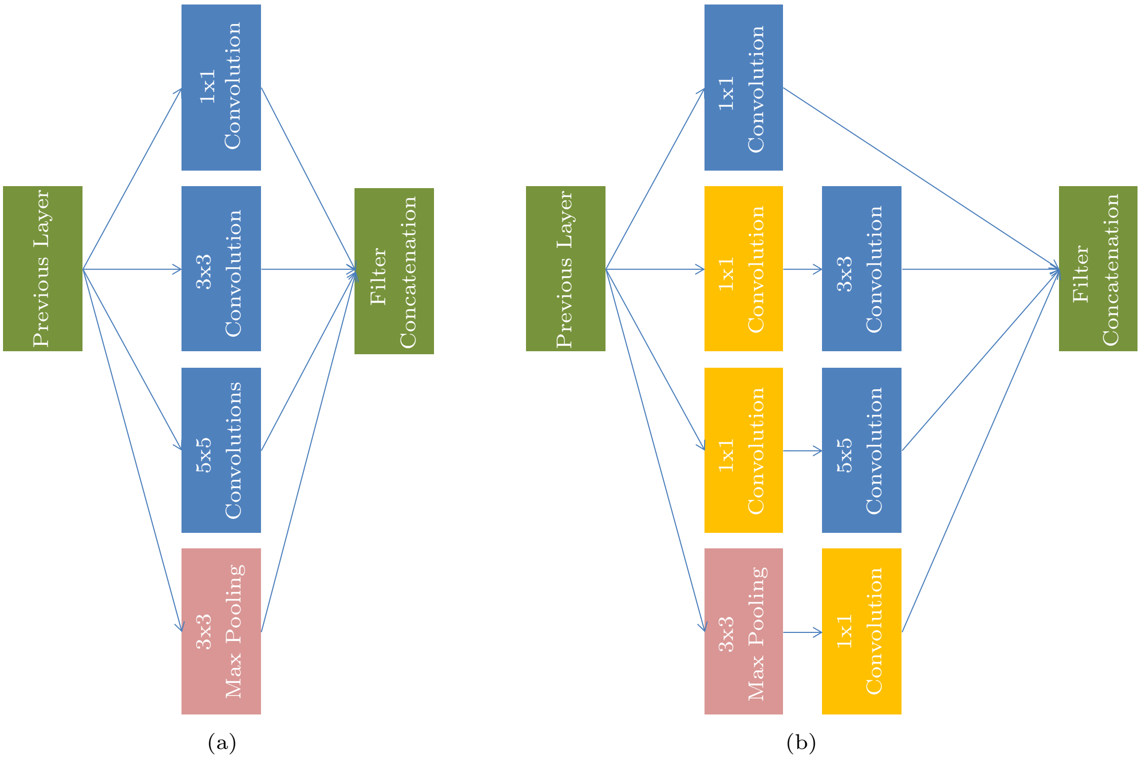


Fig. 4. (a) Schematic representation of the inception module architecture (original version). (b) Schematic representation of the inception module architecture (with dimension reductions).

sizes of receptive fields; filter concatenation fuses diverse scale features. As the network deepens, the features tend to become more abstract, and the receptive field of each feature involved is also increased. Thus, with an increasing number of layers, the proportion of 3×3 and 5×5 convolutions also increases, resulting in a huge computational load. Inspired the work proposed in [33], a 1×1 convolutional kernel is applied to dimensionality reduction. The dimension-reduction form of the inception module is shown in Fig.4(b).

2.3 FER Approaches Based on Handcrafted Features

In contrast to deep learning approaches, FER approaches based on handcrafted features do not provide a feature learning stage but a manual feature extraction process. The commonality of various types of conventional approaches is detecting the face region and extracting geometric features or appearance-based features. Even in this category of approaches, the behaviour and relative performance of algorithms is poorly analysed by scientific literature with images of expressions performed by ageing adults. Consequently, in this work, two of the best performing handcrafted features extraction methodologies have been implemented and tested on benchmark datasets.

Generally, geometric features methods are focused on the extraction from the shape or salient point locations of specific facial components (e.g., eyes, mouth, nose, eyebrows). From an evaluation of the recent research activity in this field, Active Shape Model (ASM) [6] turns out to be a performing method for FER. Here, the face of an ageing subject was processed with a facial landmarks extractor exploiting the Stacked Active Shape Model (STASM) approach. STASM uses the Active Shape Model for locating 76 facial landmarks with a simplified form of scale-invariant feature transform (SIFT) descriptors and it operates with multivariate adaptive regression splines (MARS) for descriptor matching [34]. After using the obtained landmarks, a set of 32 features, useful to recognize facial expressions, has been defined.

The 32 geometric features extracted are divided into the following three categories: linear features (18), elliptical features (4), and polygonal features (10). The last step provides a classification module that uses a support vector machine (SVM) for the analysis of the obtained features vector in order to get a prediction in terms of facial expression (Fig.5).

Regarding the use of appearance-based features, lo-

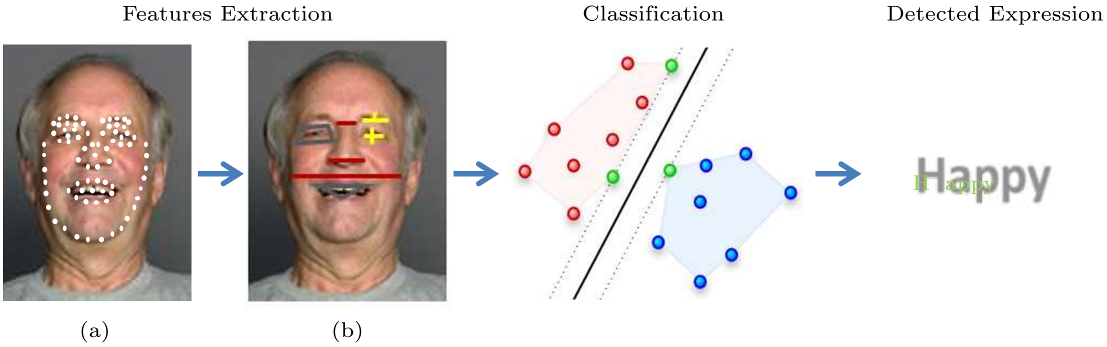


Fig.5. FER based on the geometric features extraction methodology. (a) Facial landmark localization. (b) Extraction of 32 geometric features (linear, elliptical and polygonal) using the obtained landmarks. The classification module is based on an SVM classifier able to output the most probable expression.

cal binary pattern (LBP)[35] is an effective texture description operator, which can be used to measure and extract the adjacent texture information in an image. The LBP feature extraction method used in the present work contains three crucial steps. At first, the facial image is divided into several non-overlapping blocks (set to 8 × 8 after experimenting with different block sizes). Then, LBP histograms are calculated for each block. Finally, the block LBP histograms are concatenated into a single vector. The resulting vector encodes both the appearance and the spatial relations of facial regions. In this spatially enhanced histogram, we effectively have a description of the facial image on three different levels of locality: the labels for the histogram contain the information about the patterns on a pixel level, the labels are summed over a small region to produce the information on a regional level, and the regional histograms are concatenated to build a global description of the face image. Finally, also in this case, an SVM classifier is used for the recognition of facial expression (Fig.6).

3 Experimental Setup and Results

3.1 Datasets

To validate our methodology a series of experiments were conducted using the age-expression datasets FACES[36] and Lifespan[37], the only ones in the literature that contain images of facial expressions divided by age group. Moreover, in order to have a further validation of the pipeline, experiments were conducted with the Candid Image Facial Expression (CIFE) dataset[38] and FER2013 dataset[39]. The first two datasets contain images acquired in the laboratory, with frontal view and controlled lighting conditions. Instead the other two datasets contain facial expressions more difficult to recognize as they are extracted from the web and with a non-frontal face pose.

The FACES dataset is comprised of 58 young (age range: 19–31), 56 middle-aged (age range: 39–55), and 57 old (age range: 69–80) Caucasian women and men (in total 171 subjects). The faces are frontal with fixed illumination mounted in front of and above the faces. The age distribution is not uniform and in total

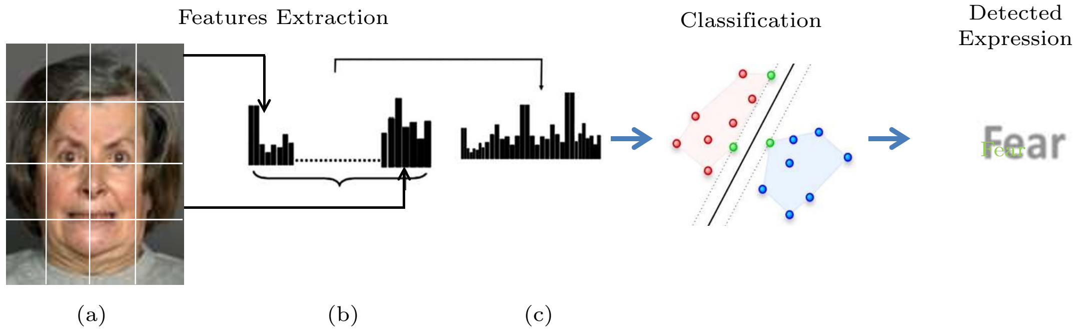


Fig.6. Appearance-based approach used for FER in ageing adults. (a) Facial image is divided into non-overlapping blocks of 8 × 8 pixels, and (b) for each block the LBP histogram is computed and then concatenated into (c) a single vector. The classification module is based on an SVM classifier able to output the most probable expression.

there are 37 different ages. Each model in the FACES dataset is represented by two sets of six facial expressions (anger, disgust, fear, happy, sad, and neutral) totaling $171 \times 2 \times 6 = 2\,052$ frontal images.

Table 1 presents the total number of persons in the final FACES dataset, broken down by age group and gender, whereas in Fig.7 some examples of expressions performed by ageing adults are represented (one for each class of facial expression).

Table 1. Total Number of Subjects Contained in FACES Dataset

Gender	Age (Years)			
	19–31	39–55	69–80	Total (19–80)
Male	29	27	29	85
Female	29	29	28	86
Total	58	56	57	171

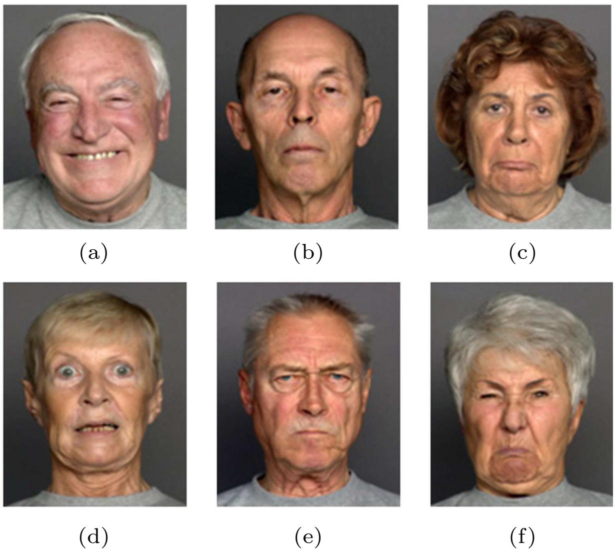


Fig.7. Some examples of expressions performed by aging adults from the FACES database. (a) Happiness. (b) Neutrality. (c) Sadness. (d) Fear. (e) Anger. (f) Disgust.

The Lifespan dataset is a collection of faces of subjects from different ethnicities showing different expressions. The ages of the subjects range from 18 to 93 years and in total there are 74 different ages. The dataset has no labeling for the subject identities. The expression subsets have the following sizes: 580, 258, 78, 64, 40, 10, 9, and 7 for neutrality, happiness, surprise, sadness, annoyed, anger, grumpiness, and disgust, respectively. Although both datasets cover a wide range of facial expressions, the FACES dataset is more challenging for FER as it contains all the facial expressions to test the methodology. Instead, only four facial expressions (neutrality, happiness, surprise, and sadness) can be considered for the Lifespan dataset due to the limited number of images in the other categories of facial expression.

Table 2 presents the total number of persons in the Lifespan dataset, divided into four different age groups and further distinguished by gender, whereas in Fig.8 some examples of expressions performed by ageing adults are represented (only for "happiness", "neutrality", "surprise" and "sadness").

Table 2. Total Number of Subjects Contained in Lifespan Dataset

Gender	Age (Years)				
	18–29	30–49	50–69	70–93	Total (18–93)
Male	114	29	28	48	219
Female	105	47	95	110	357
Total	219	76	123	158	576

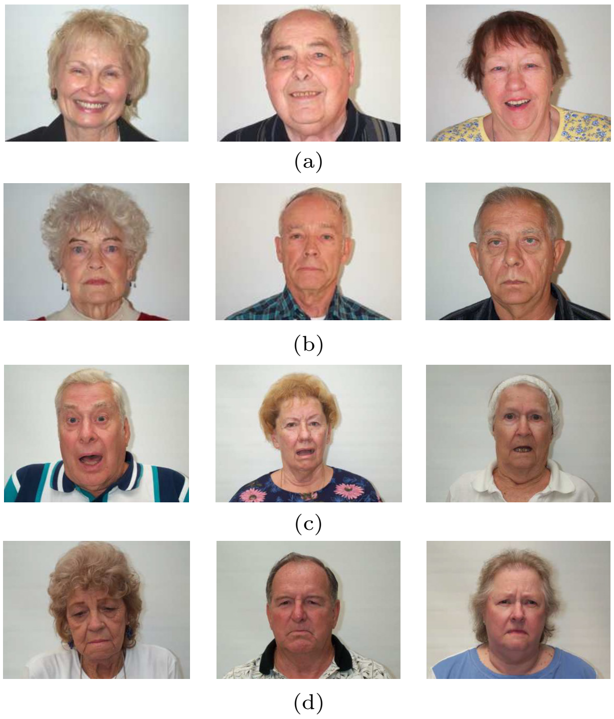


Fig.8. Some examples of expressions performed by ageing adults from the Lifespan database. (a) Happiness. (b) Neutrality. (c) Surprise. (d) Sadness.

The CIFE dataset is composed of facial images with seven different types of expressions. The expression subsets have the following sizes: 3 636, 1 905, 975, 2 485, 1 994, 1 381 and 2 381 for happiness, anger, disgust, sadness, surprise, fear, and neutrality respectively. The images are extracted from the web with gathering techniques permitted to collect in total 14 757 images containing candid expression images that are randomly posed. This last detail redeems the process of expression recognition more difficult than the two previous datasets, which are made up of facial frontal expressions acquired in controlled environments.

Since the CIFE dataset contains images with only the label of facial expression and without any indication about the age of the subject, it was necessary to perform an age estimation technique consolidated in the literature. The approach used in the present work is inspired from the algorithm described in [40] permitted to group the images into four different subgroups. Table 3 presents the total number of images in the CIFE dataset, divided according to the estimated age, whereas in Fig.9 some examples of expressions performed by ageing adults are represented.

Table 3. Total Number of Images Contained in CIFE Dataset

Age (Years)	Number of Images
< 35	5 587
35–55	4 828
56–68	2 263
> 68	2 079
Total	14 757

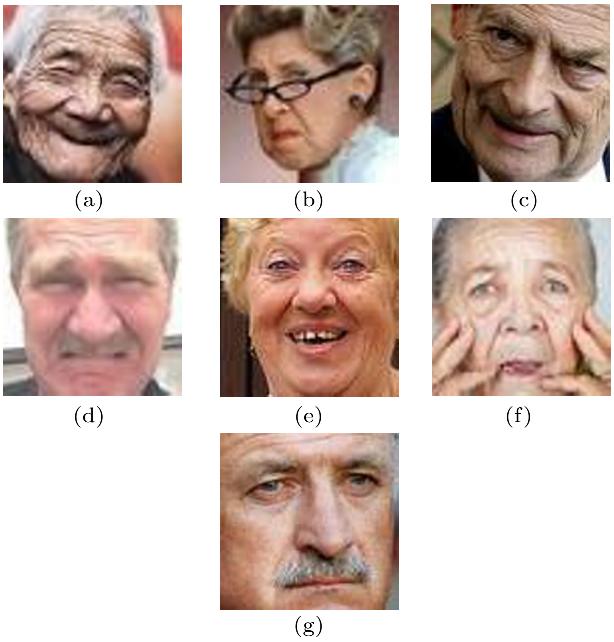


Fig.9. Some examples of expressions performed by ageing adults from the CIFE dataset. (a) Happiness. (b) Anger. (c) Disgust. (d) Sadness. (e) Surprise. (f) Fear. (g) Neutrality.

FER-2013 is a large-scale FER dataset used in the ICML 2013 workshop's facial expression recognition challenge[40]. The dataset has seven expressions including anger, disgust, fear, happiness, sadness, surprise, and neutrality. It is comprised of 48 × 48 pixel grey-scale images of human faces. The training set consists of 28 709 examples, while both the test and the validation sets are composed of 3 589 examples. The images of FER-2013 were collected from the Internet and the faces greatly vary in age, pose and occlusion, thus resulting in that the accuracy of human recognition is only approximately 65 ± 5%[41]. As a powerful machine learning tool, the CNN can now surpass human beings on the FER-2013 task, and the state-of-the-art accuracy on FER-2013 is 75.42% by combining CNN extracted features and handcrafted features for training[42].

Also the FER2013 dataset does not contain the subdivision of the images into age groups; therefore the technique proposed in [40] was used again. Table 4 presents the total number of images in the FER2013 dataset, divided according to the estimated age, whereas in Fig.10 some examples of expressions performed by ageing adults are represented.

Table 4. Total Number of Images Contained in FER2013 Dataset

Age (Years)	Number of Images
< 35	13 560
35–55	7 432
56–68	6 128
> 68	5 178
Total (training+test/validation)	32 298

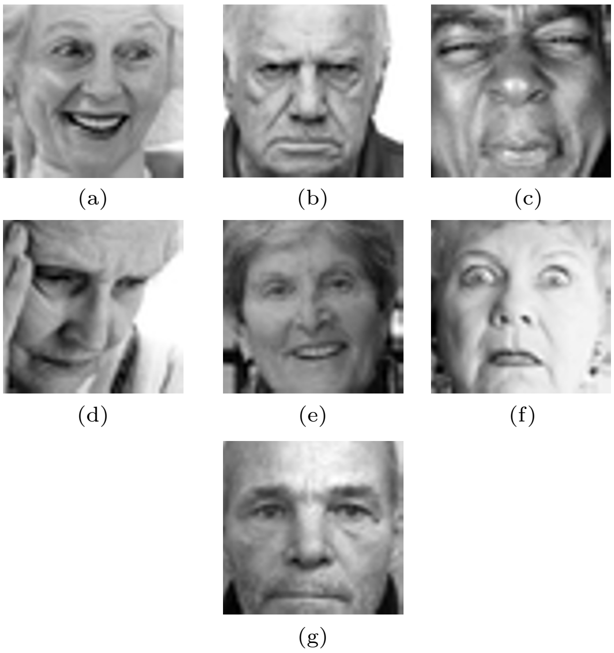


Fig.10. Some examples of expressions performed by ageing adults from the FER2013 dataset. (a) Happiness. (b) Anger. (c) Disgust. (d) Sadness. (e) Surprise. (f) Fear. (g) Neutrality.

From the examples we see how all of the images are preprocessed as they are mostly centered and adjusted so that the face occupies about the same amount of space in each image.

3.2 Performance Evaluation

The training and the testing phase were performed on Intel i7 3.5 GHz workstation with 16 GB DDR3 and equipped with GPU NVidia Titan X using Keras that is TensorFlow's high-level API developed for implementing, training, testing, and deploying deep learning models [43].

For the performance evaluation of the methodologies all the images of the FACES dataset were preprocessed, whereas only the facial images of Lifespan with the four facial expressions considered in the present work were pre-processed. Consequently, applying the data augmentation techniques previously described (see Section 2), in total 65 664 facial images of FACES (equally distributed among the facial expression classes) and 31 360 facial images of Lifespan were used, a sufficient number for using a deep learning technique. Moreover, evaluating the number and the diversity in pose and dimension of images contained in the CIFE and FER2013 datasets, it was not considered appropriate to implement data augmentation techniques for this last dataset.

Various experiments were conducted to assess the FER performance of the pre-trained VGG-16, AlexNet and GoogleNet/Inception V1 deep networks with transfer learning and of traditional machine learning methods. Regarding the usage of the deep pre-trained model, the experiments described in this section follow a standard process:

1) pre-process all facial images with the algorithmic steps described in Subsection 2.1 (only for FACES and Lifespan);

2) run the labelled facial images dataset through different deep learning networks to generate image vectors;

3) train a classifier using the training set image vectors to predict the labels;

4) predict test set labels using the trained classifier in step 3 and using the test set image vectors generated in step 2;

5) evaluate FER detection performance metrics with varying the age.

In relation to step 4 described above, standard fine-tuning procedure was implemented which consists in removing the last (classification) layer for all the deep architectures.

For the final classifier layer, specific machine learning classifiers that have shown promising results in previous FER studies were compared, such as random forest (RF), support vector machine (SVM) and logistic regression (LR).

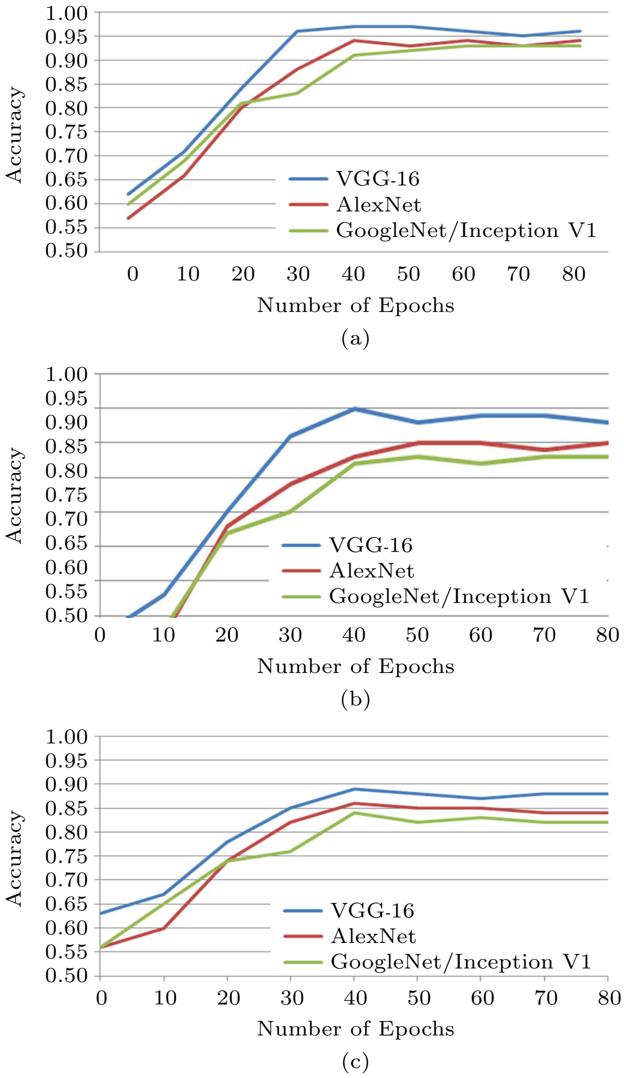
The metric used in this work for evaluating the methodologies is the accuracy, whose value is calculated using the average of n-class classifier accuracy for each expression (i.e., number of hits of an expression per total number of images with the same expression):

$$Acc = \frac{\sum_1^n Acc_{\text{expr}}}{n},$$
$$Acc_{\text{expr}} = \frac{Hit_{\text{expr}}}{Total_{\text{expr}}},$$

where Hit_{expr} is the number of hits in the expression expr, $Total_{\text{expr}}$ represents the total number of samples of that expression and n is the number of expressions to be considered.

Figs.11–14 report the average accuracy for FER on images of the FACES, Lifespan, CIFE and FER2013 datasets respectively. For each dataset the accuracy obtained with the three considered deep-learning approaches is reported, depending on the classifiers used. The obtained accuracy is evaluated with varying a very important hyper-parameter (i.e., epoch) that it is necessary to set well when the performance of a deep neural network must be assessed.

Obviously, different parameters have also been tested for the used classifiers, selecting the combination that allows to obtain the best performances in terms of accuracy without slowing down the classification process considerably. In the case of the RF classifier, it is sensitive to two parameters namely the number of trees in the forest *nb_tree* and the number of features chosen for a split *mtry*. Here, we set $nb\_tree = 100$ and $mtry = 50$. For the SVM classifier the most important parameter to set is the kernel function and in the present work, after different tests with polynomial, Gaussian, sigmoid and Radial Basis Function (RBF) kernel, the latter was selected. Finally, for LR we set only the parameter C (that is the inverse of regularization strength λ) to 0.01.



(a)

(b)

(c)

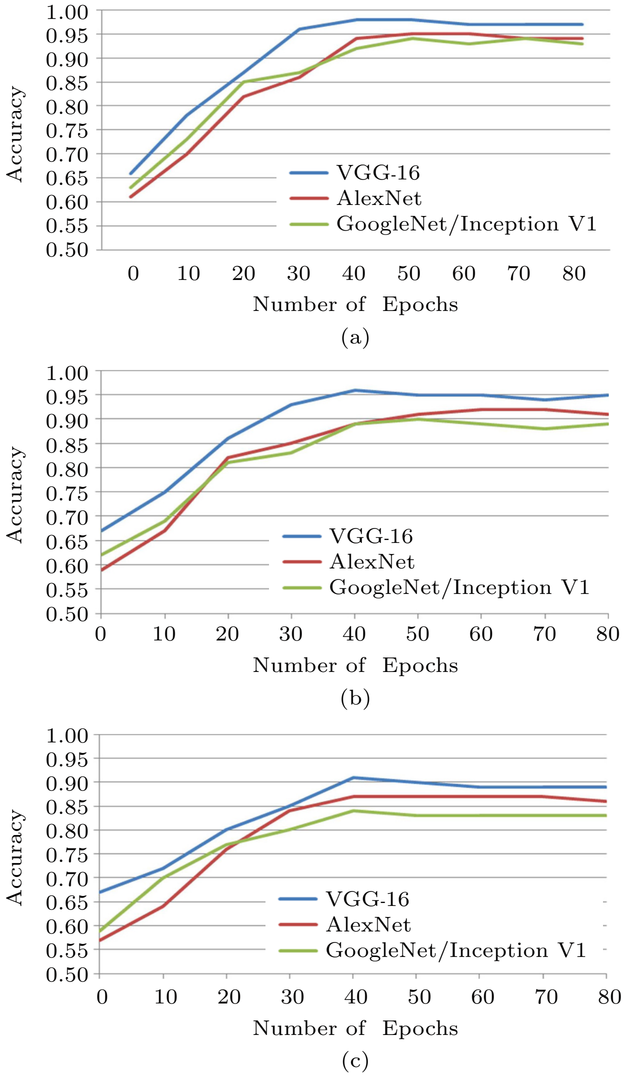
Fig.11. Average accuracy of deep learning approaches on the FACES dataset with (a) RF classifier, (b) SVM classifier, and (c) LR classifier.

From the results reported in Figs.11–14, it is evident that the recognition performances of facial expressions vary significantly as the dataset changes and the number of facial expression classes contained within them. The deep learning approaches analyzed obtain higher expression recognition rates for a number of epochs between 40 and 50. Moreover, the VGG-16 architecture is the one that performs best for all the considered datasets, with the RF classifier which tends to provide a greater accuracy in the results (about 2.5% with respect to SVM and 6% with respect to LR).

A first comparison to verify the quality of the results obtained can be made with the classification performances reported in [44], where the authors evaluated the FER accuracy only in the FACES dataset (limiting the performance evaluation only to frontal images of the face) using a CNN architecture inspired at the classical LeNet-5 architecture. In this work, an average accuracy of 92.81% was achieved, whereas in this work an average accuracy of 97.21% was obtained with the VGG-16 deep learning architecture associated with the RF classifier, obtaining an overall improvement of 4.4 % in the classification of facial expressions.

Table 5. Performance Comparison on CIFE Dataset

Method	Accuracy (%)
CNN with 7 layers [40]	81.5
VGG [45]	76.3
AlexNet [45]	73.5
Our best method (VGG-16+RF)	**84.0**

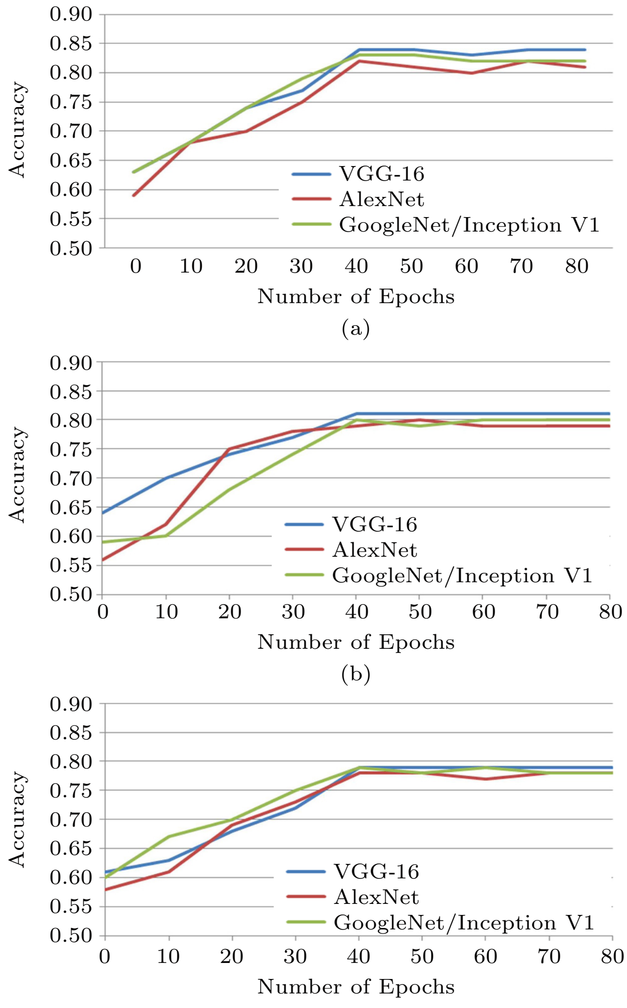


(a)

(b)

(c)

Fig.12. Average accuracy of deep learning approaches on the Lifespan dataset with (a) RF classifier, (b) SVM classifier, and (c) LR classifier.



(a)

(b)

(c)

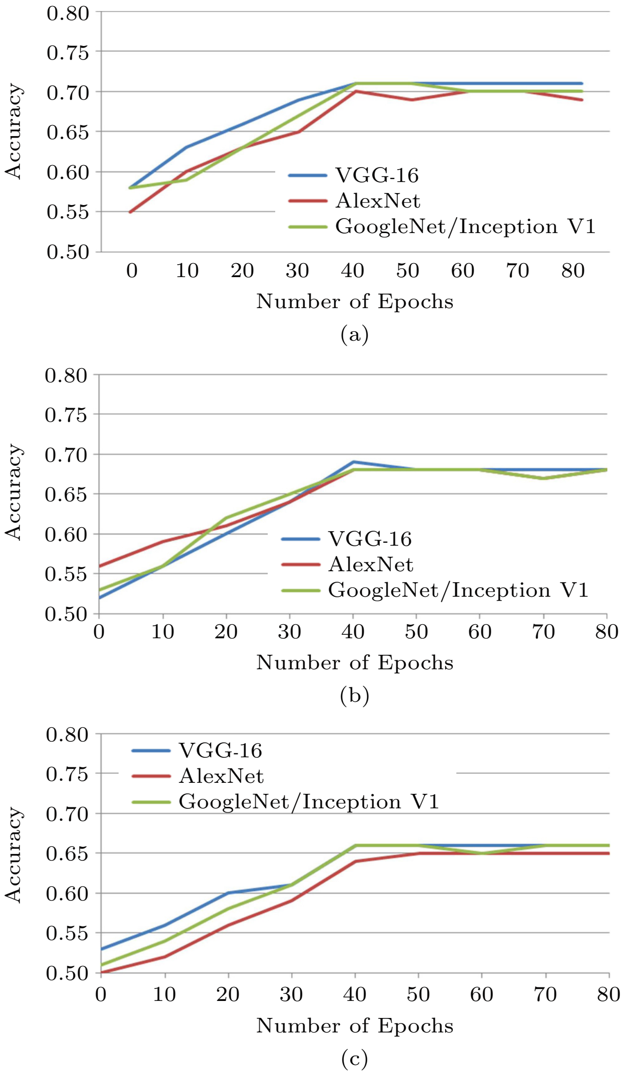
Fig. 13. Average accuracy of deep learning approaches on the CIFE dataset with (a) RF classifier, (b) SVM classifier, and (c) LR classifier.

Table 6. Performance Comparison on FER2013 Dataset

Methodolog	Accuracy (%)
CNN with 5 layers+L2-SVM loss function [23]	**71.6**
VGG-11 [46]	59.6
AlexNet [46]	54.4
Multiple Kernel Learning+SIFT [47]	67.4
Our best method (VGG-16+RF)	71.5

Table 7. FER Accuracy on FACES Dataset

Age Group	VGG-16+ RF (%)	ASM+ SVM (%)	LBP+ SVM (%)
Young (19–31 years)	97.84	86.42	87.22
Middle-aged (39–55 years)	97.06	86.81	87.47
Old (69–80 years)	96.73	84.98	85.61
Overall accuracy	97.21	86.07	86.77



(a)

(b)

(c)

Fig. 14. Average accuracy of deep learning approaches on the FER2013 dataset with (a) RF classifier, (b) SVM classifier, and (c) LR classifier.

Since FACES and Lifespan contain facial expressions with a frontal face acquired under controlled conditions, the quality of the results can be appreciated in an ample manner by analyzing the classification performances obtained on the CIFE and FER2013 datasets, which are more challenging as the depicted faces vary significantly in terms of face pose and other factors, reflecting realistic conditions. Consequently our experiment results were compared with other published recognition accuracy results. Table 9 shows that our method outperforms most of the published results considering the CIFE dataset whereas Table 10 shows that the winner of the FER challenge (that was held as part of the ICML workshop in the year 2013) achieves a higher accuracy (71.6%) than our method (71.5%) considering FER2013, but only slightly.

Table 8. FER Accuracy on Dataset Lifespan

Age Group	VGG-16+ RF (%)	ASM+ SVM (%)	LBP+ SVM (%)
Young (18–29 years)	99.34	90.16	90.54
Middle-aged (30–49 years)	98.63	89.24	90.01
Old (50–69 years)	97.44	86.12	86.32
Very old (70–93 years)	96.91	85.28	86.01
Overall accuracy	98.08	87.70	88.22

Table 9. FER Accuracy on Dataset CIFE

Age Group	VGG-16+ RF (%)	ASM+ SVM (%)	LBP+ SVM (%)
Young (< 35 years)	86.13	68.05	71.98
Middle-aged (35–55 years)	84.26	66.20	65.97
Old (56–68 years)	83.09	62.64	61.18
Very old (> 68 years)	82.52	58.23	57.35
Overall accuracy	84.00	63.78	64.12

Table 10. FER Accuracy on Dataset FER2013

Age Group	VGG-16+ RF (%)	ASM+ SVM (%)	LBP+ SVM (%)
Young (< 35 years)	75.33	60.68	63.11
Middle-aged (35–55 years)	71.98	57.87	59.51
Old (56–68 years)	70.22	55.99	56.24
Very old (> 68 years)	68.47	51.22	53.76
Overall accuracy	71.50	56.44	58.08

It is important to underline how the previous results were obtained by considering all age groups in which facial expressions were divided. Since the main objective of this work lies in assessing the impact of ages on the recognition of facial expressions, experiments have been carried out with the aim of measuring the performance of the methodologies by grouping the images of the datasets into different age groups. Given the results obtained previously, the approach that combines the VGG-16 architecture with the RF classifier was tested for the classification of facial expressions using deep learning.

Table 7–Table 10 report the final accuracy obtained by the proposed deep architecture for each age group detected in FACES, Lifespan, CIFE and FER2013 respectively. Moreover, in order to make a comparison, the same tables show the accuracy values obtained using traditional machine learning techniques described in Subsection 2.3 (ASM+SVM and LBP+ SVM).

The results reported in Tables 7–9 confirm that our proposed deep architecture combined with random forest classifier is superior to traditional approaches based on handcrafted features and this is true for any age group in which the datasets are partitioned. Analyzing in more detail the performance, it is clear that the proposed architecture obtains a better improvement in the case of recognition of facial expressions performed by ageing adults. Moreover, the hypotheses concerning the difficulties of traditional algorithms in extracting features from an ageing face were confirmed from the fact that ASM and LBP get a greater accuracy with faces of the young and the middle-aged for each analyzed dataset.

Often, in real-life applications, the expression performed by an observed subject could be very different from the training samples used, in terms of uncontrolled variations such as illumination, pose, age and gender. Therefore, it is important for an FER system to have a good generalization power. As a result it turns out to be essential to design and implement a methodology for feature extraction and classification that is still able to achieve a good performance when the training and test sets are from different datasets. In this paper, we also conduct experiments to test the robustness and accuracy of the compared approaches in the scenario of cross-dataset FER.

Table 11 shows the results when the training and the testing sets are two different datasets (FACES and Lifespan) within which there are subjects of different ethnicity and of different ages. Furthermore, image resolution and acquisition conditions are also significantly different. From the results obtained it is evident that the recognition rates for the three basic emotions in common between the two datasets ("happiness", "neutrality" and "sadness") decrease significantly, because cross-dataset FER is a challenging task. Moreover, this difficulty in classification is greater in the case of facial expressions of young subjects who express emotions more strongly than the ageing adults.

On the other hand, in Table 12 cross-dataset FER performances are evaluated considering the other two datasets proposed in the present work (CIFE and FER2013), each of which contains all six facial expressions considered in the Ekman classification plus the neutral expression.

Here we note how the accuracy is lower than that obtained by considering the FACES and Lifespan datasets, but this is mainly due to two causes: the number of facial expression classes considered (seven

Table 11. Comparison of Recognition Rate (%) of Methods on Cross-Dataset FER Containing Images with Only Frontal View and Acquired Under Controlled Conditions

Training Set	Testing Set	Method	Classifier	Group		
				Young (%)	Middle-Aged (%)	Old/Very Old(%)
FACES	Lifespan	VGG-16	RF	51.38	57.34	59.64
		ASM	SVM	42.44	46.89	51.68
		LBP	SVM	44.56	50.13	52.78
Lifespan	FACES	VGG-16	RF	53.47	55.98	60.07
		ASM	SVM	41.87	45.12	49.89
		LBP	SVM	41.13	47.76	51.81

Table 12. Comparison of Recognition Rate (%) of Methods on Cross-Dataset FER Containing "in the Wild" Images

Training Set	Testing Set	Method	Classifier	Group		
				Young (%)	Middle-Aged (%)	Old/Very Old (%)
CIFE	FER2013	VGG-16	RF	55.56	55.34	56.30
		ASM	SVM	44.21	45.12	46.38
		LBP	SVM	46.46	46.33	47.17
FER2013	CIFE	VGG-16	RF	55.12	56.24	56.39
		ASM	SVM	45.99	46.47	47.13
		LBP	SVM	44.38	44.85	45.24

vs three) and the type of facial images contained within the CIFE and FER2013 datasets ("in the wild" images). Finally, it should be noted that in this experiment the performances do not vary significantly with ages, which is more noticeable by analyzing the experimental results contained in Table 12.

In a multi-class recognition problem, as the FER one, the use of an average recognition rate (i.e., accuracy) among all the classes could not be exhaustive since there is no possibility to inspect what is the separation level, in terms of correct classifications, among classes (in our case, different facial expressions). To overcome this limitation, for each dataset the confusion matrices are then reported in Tables 13–16 (only the facial images of ageing adults were considered). The numerical results obtained make possible a more detailed analysis of the misclassification and the interpretation of their possible causes. First of all, from the confusion matrices it is possible to observe that the pipeline based on the proposed VGG-16 architecture (in combination with RF classifier) decreases its FER performance when the number of classes increases or the considered dataset contains images with non-frontal pose.

Going into a more detailed analysis on the results reported in Table 13 and related to the FACES dataset, "anger" and "fear" are the facial expression better recognized, whereas "sadness" and "neutrality" are the facial expression confused the most. Finally, "sadness" is the facial expression with the lowest accuracy. Instead, the confusion matrix reported in Table 14 and related to facial expression classes of the Lifespan dataset highlights that "happiness" is the facial expression with the best accuracy, whereas the expression "surprise" is the worst expression recognized. "Surprise" and "happiness" are the facial expression confused the most.

Table 13. Confusion Matrix on Dataset FACES (Performed by Old Adults) Using the Proposed VGG-16+RF Architecture

Actual (%)	Estimated (%)					
	An	Di	Fe	Ha	Sa	Ne
An	**98.9**	0.0	0.2	0.0	0.9	0.0
Di	1.2	**96.2**	0.7	0.0	1.9	0.0
Fe	0.7	0.2	**98.1**	0.0	0.6	0.4
Ha	0.0	0.0	1.9	**97.6**	0.0	0.5
Sa	0.3	0.0	2.5	0.0	**93.5**	3.7
Ne	2.1	0.1	1.8	0.1	0.0	**95.9**

Note: An = Anger, Di = Disgust, Fe = Fear, Ha = Happiness, Sa = Sadness, and Ne = Neutrality.

Table 14. Confusion Matrix on Lifespan Dataset (Performed by the Old and Very Old Adults) Using the Proposed VGG-16+RF Architecture

Actual (%)	Estimated (%)			
	Ha	Ne	Su	Sa
Ha	**98.8**	0.0	1.2	0.0
Ne	1.1	**98.1**	0.6	0.2
Su	3.0	0.7	**96.2**	0.1
Sa	0.3	2.6	1.4	**95.7**

Note: Ha = Happiness, Ne = Neutrality, Su = Surprise, and Sa = Sadness.

Table 15. Confusion Matrix on CIFE Dataset (Performed by Old and Very Old Adults) Using the Proposed VGG-16+RF Architecture

Actual (%)	Estimated (%)						
	Ha	An	Di	Sa	Su	Fe	Ne
Ha	**91.4**	1.1	0.4	1.5	3.7	0.6	1.3
An	0.7	**86.4**	4.0	6.9	1.1	0.7	1.0
Di	0.9	4.2	**74.9**	0.9	3.9	1.5	13.7
Sa	0.5	0.5	2.0	**87.5**	0.4	3.3	5.8
Su	2.0	1.6	1.5	2.1	**78.9**	9.9	4.0
Fe	1.9	3.3	2.5	2.7	8.4	**76.7**	4.5
Ne	0.8	1.1	1.6	5.8	5.7	1.2	**83.8**

Note: Ha = Happiness, An = Anger, Di = Disgust, Sa = Sadness, Su = Surprise, Fe = Fear, and Ne = Neutrality.

Table 16. Confusion Matrix on FER2013 Dataset (Performed by Old and Very Old Adults) Using the Proposed VGG-16+RF Architecture

Actual (%)	Estimated (%)						
	Ha	An	Di	Sa	Su	Fe	Ne
Ha	**79.4**	1.5	1.5	2.0	6.9	3.9	4.8
An	3.9	**58.5**	16.8	11.9	3.6	3.4	1.9
Di	1.1	11.0	**77.2**	0.6	6.8	1.8	1.5
Sa	2.4	15.6	2.6	**52.9**	3.1	1.9	21.5
Su	7.9	0.2	0.1	6.1	**81.1**	1.6	3.0
Fe	2.1	14.5	2.4	13.8	3.4	**61.6**	2.2
Ne	10.2	1.0	2.4	9.9	0.9	1.2	**74.4**

Note: Ha = Happiness, An = Anger, Di = Disgust, Sa = Sadness, Su = Surprise, Fe = Fear, and Ne = Neutrality.

The results reported in Table 15 and Table 16, on the other hand, first of all highlight the greater level of difficulty in classifying facial expressions of the elderly in unchecked conditions. These results also lead to a very important observation analyzing the most confused expressions, that is, the elderly are less expressive and tend to group expressions into positive, negative and neutral.

In fact "anger" is more confused with "disgust" and "sadness" which are expressions categorized in reality as negative, and "happiness" is confused most with "surprise" (and vice versa). A last important consideration extracted from the results must be made on the "neutrality" expression, which confirms the confusion of this expression with the expression of "sadness", that is, a typical attitude of an ageing adult.

Even in this case, in order to make a complete and exhaustive comparison, the confusion matrices obtained with traditional machine learning methods are reported in Tables 17–20 (ASM+SVM) and Tables 21–24 (LBP+SVM).

Table 17. Confusion Matrix on FACES Dataset (Performed by Old Adults) Using ASM with SVM Classifier

Actual (%)	Estimated (%)					
	An	Di	Fe	Ha	Sa	Ne
An	**86.6**	1.6	4.3	1.6	5.1	0.8
Di	4.8	**84.1**	2.4	0.9	6.3	1.5
Fe	5.0	2.1	**85.9**	0.7	4.7	1.6
Ha	1.2	2.0	6.3	**85.3**	1.8	3.4
Sa	1.9	1.6	2.4	7.5	**83.0**	3.6
Ne	3.4	2.3	4.4	1.6	3.2	**85.1**

Note: An = Anger, Di = Disgust, Fe = Fear, Ha = Happiness, Sa = Sadness, and Ne = Neutrality.

Table 18. Confusion Matrix on Lifespan Dataset (Performed by Old and Very Old Adults) Using ASM with SVM Classifier

Actual (%)	Estimated (%)			
	Ha	Ne	Su	Sa
Ha	**87.0**	2.8	8.1	2.1
Ne	7.5	**86.1**	4.2	2.2
Su	9.4	3.8	**85.3**	1.5
Sa	2.3	7.9	5.4	**84.4**

Note: Ha = Happiness, Ne = Neutral, Su = Surprise, and Sa = Sadness.

Table 19. Confusion Matrix on CIFE Dataset (Performed by Old and Very Old Adults) Using ASM with SVM Classifier

Actual (%)	Estimated (%)						
	Ha	An	Di	Sa	Su	Fe	Ne
Ha	**67.6**	3.1	2.2	7.8	10.1	2.3	6.9
An	3.3	**62.8**	11.1	10.5	3.6	4.5	4.2
Di	2.2	3.0	**57.7**	2.1	12.7	3.9	18.4
Sa	2.8	2.7	4.5	**63.7**	3.1	10.4	12.8
Su	3.9	4.2	2.8	3.5	**55.5**	15.4	14.7
Fe	2.4	6.7	2.9	4.0	17.7	**53.9**	12.4
Ne	3.0	3.5	4.9	11.7	11.6	3.7	**61.6**

Note: Ha = Happiness, An = Anger, Di = Disgust, Sa = Sadness, Su = Surprise, Fe = Fear, and Ne = Neutrality.

Table 20. Confusion Matrix on FER2013 Dataset (Performed by Old and Very Old Adults) Using ASM with SVM Classifier

Actual (%)	Estimated (%)						
	Ha	An	Di	Sa	Su	Fe	Ne
Ha	**56.8**	2.9	1.8	2.2	16.9	4.1	15.3
An	3.7	**44.4**	21.5	19.4	4.2	4.7	2.1
Di	2.6	15.0	**65.3**	8.9	3.3	2.2	2.7
Sa	4.0	16.8	5.0	**43.9**	3.5	6.7	20.1
Su	10.5	3.4	4.4	8.9	**62.4**	3.1	7.3
Fe	2.6	18.9	4.7	19.5	5.6	**45.5**	3.2
Ne	12.4	3.7	5.1	14.3	4.5	3.1	**56.9**

Note: Ha = Happiness, An = Anger, Di = Disgust, Sa = Sadness, Su = Surprise, Fe = Fear, and Ne = Neutrality.

Table 21. Confusion Matrix on FACES Dataset (Performed by Old Adults) Using LBP with SVM Classifier

Actual (%)	Estimated (%)					
	An	Di	Fe	Ha	Sa	Ne
An	**88.8**	1.7	2.7	1.5	3.5	1.8
Di	4.3	**84.4**	2.1	1.6	5.6	2.0
Fe	4.3	1.6	**87.4**	0.9	4.4	2.4
Ha	1.7	1.5	4.8	**86.0**	2.2	3.8
Sa	2.0	1.4	5.4	2.5	**82.9**	5.8
Ne	5.3	2.6	4.9	1.6	1.5	**84.1**

Note: An = Anger, Di = Disgust, Fe = Fear, Ha = Happiness, Sa = Sadness, and Ne = Neutrality.

Table 22. Confusion Matrix on Lifespan Dataset (Performed By Old and Very Old Adults) Using LBP with SVM Classifier

Actual (%)	Estimated (%)			
	Ha	Ne	Su	Sa
Ha	**87.8**	2.4	7.9	1.9
Ne	8.3	**86.9**	3.2	1.6
Su	8.4	3.3	**85.5**	2.8
Sa	2.2	8.7	4.5	**84.6**

Note: Ha = Happiness, Ne = Neutrality, Su = Surprise, and Sa = Sadness.

Table 23. Confusion Matrix on CIFE Dataset (Performed By Old and Very Old Adults) Using LBP with SVM Classifier

Actual (%)	Estimated (%)						
	Ha	An	Di	Sa	Su	Fe	Ne
Ha	**68.8**	3.1	2.3	7.5	9.1	2.8	6.4
An	3.1	**65.2**	10.6	10.1	3.1	4.1	3.8
Di	3.2	4.9	**52.2**	2.8	13.0	4.3	19.6
Sa	2.2	2.4	4.5	**65.2**	3.1	10.2	12.4
Su	3.4	5.3	3.1	3.7	**54.4**	15.9	14.2
Fe	2.9	5.7	2.9	4.5	16.6	**53.0**	14.4
Ne	2.8	2.5	2.1	17.4	16.7	2.2	**56.3**

Note: Ha = Happiness, An = Anger, Di = Disgust, Sa = Sadness, Su = Surprise, Fe = Fear, and Ne = Neutrality.

Table 24. Confusion Matrix on FER2013 Dataset (Performed By Old and Very Old Adults) Using LBP with SVM Classifier

Actual (%)	Estimated (%)						
	Ha	An	Di	Sa	Su	Fe	Ne
Ha	**61.4**	1.9	1.3	2.5	14.8	4.1	14.0
An	3.1	**50.8**	19.4	18.0	3.2	3.7	1.8
Di	3.1	16.2	**64.6**	7.8	3.5	2.5	2.3
Sa	3.2	15.7	4.8	**47.9**	3.9	6.1	18.4
Su	11.0	3.8	5.4	9.2	**58.5**	3.6	8.5
Fe	2.2	18.3	4.5	19.0	5.5	**47.5**	3.0
Ne	13.3	4.0	4.8	14.3	5.7	3.6	**54.3**

Note: Ha = Happiness, An = Anger, Di = Disgust, Sa = Sadness, Su = Surprise, Fe = Fear, and Ne = Neutrality.

4 Discussion and Conclusions

Deep learning has absolutely dominated computer vision over the last few years, achieving top scores on many tasks such as automatic speech recognition, image recognition, natural language processing (NLP), drug and materials discovery, and so on. However, the large number of hidden neurons and layers used in deep architectures results in computationally-intensive matrix and vector computations involving millions of parameters, requiring the use of high-performance computing systems. Another issue in this context lies in the impossibility to have “big data” with labelled samples in many domains to be able to train an entire deep architecture from scratch. Consequently, the use of a pre-trained deep learning model with few data available has shown to be a winning solution across domains.

The purpose of this paper was to explore and evaluate the deep transfer learning approach for the FER task in ageing adults considering that the majority of the studies in the literature that address FER topic are based on benchmark datasets that contain facial images with a small span of lifetime (generally young and middle-aged subjects). The evaluation was done using three different deep learning architectures trained on “big data” image datasets and “transfer” their learning ability to automatically recognize the facial expression belonging to four different datasets. Moreover, after testing which deep learning architecture achieves the best performance, the latter has been compared with traditional machine learning methods. This last step was done to highlight how the age of the face affects the recognition of facial expressions.

The following are some significant findings: the overall approach of using pre-trained deep convolutional neural network model for FER was shown to be successful. In particular, the Keras library[43] provides a nearly ready to use a platform for easy implementation of deep transfer learning approach for facial expression classification. Among the three different deep architectures tested, the pre-trained VGG-16 deep convolutional neural network in combination with an RF classifier yielded the best performance for each considered dataset and for each age group in which the dataset has been divided. Since the facial images contained in datasets CIFE and FER2013 used in this study are significantly more complex than those used in FACES and Lifespan, a significantly higher order of complexity was also introduced. With regard to the CIFE dataset, the implemented approach improves the current state of the art, while for the FER2013 dataset the performances are in line with those published in other research papers.

Another important conclusion that has been reached in the present work is that the proposed deep architecture (VGG-16+RF) is more effective in the classification of facial expressions with respect to the two considered methodologies of machine learning, and the greatest progress in terms of accuracy was found in correspondence with the recognition of facial expressions of elderly subjects. Probably, these results are related to the deformations (wrinkles, folds, etc.) that are more present on the face of the elderly, which greatly affect the use of handcrafted features for classification purposes.

A further added value of this work lies in the implementation of pre-processing blocks. In fact, it was necessary to implement "data augmentation" methodologies as the facial images available in the FACES and Lifespan datasets were not sufficient for a correct use of a deep learning methodology. The implemented pipeline also provides a series of algorithmic steps which produce normalized facial images, which represent the input for the implemented FER methodologies.

The classification accuracy in cross dataset evaluation showed that FER in ageing adults is still a topic to be investigated in depth, and even the difficulty in classification has been accentuated more in the case of facial expressions of young and middle-aged subjects, but that is probably due to the fact that these subjects express emotions more strongly than the ageing adults.

Future studies will deal with four main aspects. First of all, the proposed deep architecture will be tested in the field of assistive technologies, 1) validating it in a smart home setup and 2) testing the pipeline in a real ambient assisted living environment, which is the old person's home. In particular, the idea is to develop an application that uses the webcam integrated in TV, smartphone or tablet with the purpose to recognize the facial expression of ageing adults in real time and through various cost-effective commercially available devices that are generally present in the living environments of the elderly. The application to be implemented will have to be the starting point to evaluate and eventually modify the mood of the old people living alone at their homes, for example by subjecting it to external sensory stimuli, such as music and images. Secondly, a more wide analysis of how a non-frontal view of the face can affect the facial expression detection rate using the proposed deep architecture will be done, as it may be necessary to monitor the mood of the elderly by using for example a camera installed in the "smart" home for other purposes (e.g., activity recognition or fall detection), and the position of these cameras almost never allows to have a frontal face image of the monitored subject.

Another development of this work might be to perform the pre-training of deep architectures on datasets different from ImageNet and more specific for the topic considered, such as EmotioNet[48] and AffectNet[49].

Last but not least, it will be necessary to extend the number of compared deep learning approaches since a limitation of the present work was the evaluation of only three pre-trained deep architectures which, analyzing the current state of the art, have already been overcome in terms of image classification, from more deeper architectures like ResNet[50], Inception-v3[51], Inception-v4[52], Inception-ResNet-V2[52] and Xception[53].

References

[1] Zeng Z, Pantic M, Roisman G I, Huang T S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(1): 39-58.

[2] Pantic M, Rothkrantz L J M. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22(12): 1424-1445.

[3] Fasel B, Luettin J. Automatic facial expression analysis: A survey. *Pattern Recognition*, 2003, 36(1): 259-275.

[4] Carroll J M, Russell J A. Do facial expressions signal specific emotions? Judging emotion from the face in context. *Journal of Personality and Social Psychology*, 1996, 70(2): 205-218.

[5] Rolls E T, Ekman P, Perrett D I *et al.* Facial expressions of emotion: An old controversy and new findings: Discussion. *RSPTB*, 335(1273): 69.

[6] Shbib R, Zhou S. Facial expression analysis using active shape model. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2015, 8(1): 9-22.

[7] Cheon Y, Kim D. Natural facial expression recognition using differential-AAM and manifold learning. *Pattern Recognition*, 2009, 42(7): 1340-1350.

[8] Soyel H, Demirel H. Facial expression recognition based on discriminative scale invariant feature transform. *Electronics Letters*, 2010, 46(5): 343-345.

[9] Gu W, Xiang C, Venkatesh Y V, Huang D, Lin H. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, 2012, 45(1): 80-91.

[10] Shan C, Gong S, McOwan P W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 2009, 27(6): 803-816.

[11] Chen J, Chen Z, Chi Z, Fu H. Facial expression recognition based on facial components detection and HOG features. In *Proc. the Scientific Cooperations International Workshops on Electrical and Computer Engineering Subfields*, Aug. 2014, pp.884-888.

[12] Guo G, Guo R, Li X. Facial expression recognition influenced by human ageing. *IEEE Transactions on Affective Computing*, 2013, 4(3): 291-298.

[13] Wang S, Wu S, Gao Z, Ji Q. Facial expression recognition through modeling age-related spatial patterns. *Multimedia Tools and Applications*, 2016, 75(7): 3937-3954.

[14] Malatesta C Z, Izard C E. The facial expression of emotion: Young, middle-aged, and older adult expressions. In *Emotion in Adult Development*, Malatesta C Z, Izard C E (eds.), Sage Publications, 1984, pp.253-273.

[15] Malatesta-Magai C, Jonas R, Shepard B, Culver L C. Type A behavior pattern and emotion expression in younger and older adults. *Psychology and Aging*, 1992, 7(4): 551-561.

[16] Malatesta C Z, Fiore M J, Messina J J. Affect, personality, and facial expressive characteristics of older people. *Psychology and Aging*, 1987, 2(1): 64-69.

[17] Lozano-Monasor E, López M T, Vigo-Bustos F, Fernández-Caballero A. Facial expression recognition in ageing adults: From lab to ambient assisted living. *Journal of Ambient Intelligence and Humanized Computing*, 2017, 8(4): 567-578.

[18] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436-444.

[19] Yu D, Deng L. Deep learning and its applications to signal and information processing [Exploratory DSP]. *IEEE Signal Processing Magazine*, 2011, 28(1): 145-154.

[20] Li S, Deng W. Deep facial expression recognition: A survey. arXiv:1804.08348, 2018. https://arxiv.org/abs/1804.08348, Dec. 2019.

[21] Ginne R, Jariwala K. Facial expression recognition using CNN: A survey. *International Journal of Advances in Electronics and Computer Science*, 2018, 5(3): 13-16.

[22] Goodfellow I J, Erhan D, Carrier P L *et al.* Challenges in representation learning: A report on three machine learning contests. In *Proc. the 20th International Conference on Neural Information Processing*, Nov. 2013, pp.117-124.

[23] Kahou S E, Pal C, Bouthillier X *et al.* Combining modality specific deep neural networks for emotion recognition in video. In *Proc. the 15th ACM on International Conference on Multimodal Interaction*, Dec. 2013, pp.543-550.

[24] Liu M, Wang R, Li S, Shan S, Huang Z, Chen X. Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *Proc. the 16th International Conference on Multimodal Interaction*, Nov. 2014, pp.494-501.

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014. https://arxiv.org/abs/1409.1556, Dec. 2019.

[26] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In *Proc. the 26th Annual Conference on Neural Information Processing Systems*, Dec. 2012, pp.1106-1114.

[27] Szegedy C, Liu W, Jia Y *et al.* Going deeper with convolutions. In *Proc. the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp.1-9.

[28] Viola P, Jones M J. Robust real-time face detection. *International Journal of Computer Vision*, 2004, 57(2): 137-154.

[29] Zuiderveld K. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*, Heckbert P S (ed.), Academic Press Professional, 1994, pp.474-485.

[30] Hubel D H, Wiesel T N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 1968, 195(1): 215-243.

[31] Pan S J, Yang Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1345-1359.

[32] Russakovsky O, Deng J, Su H *et al.* ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, 115(3): 211-252.

[33] Lin M, Chen Q, Yan S. Network in network. arXiv:1312.4400, 2013. https://arxiv.org/abs/1312.4400, Dec. 2019.

[34] Milborrow S, Nicolls F. Active shape models with SIFT descriptors and MARS. In *Proc. the 9th International Conference on Computer Vision Theory and Applications*, Jan. 2014, pp.380-387.

[35] Shan C, Gong S, McOwan P W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 2009, 27(6): 803-816.

[36] EbnerN C, Riediger M, Lindenberger U. FACES — A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 2010, 42(1): 351-362.

[37] Minear M, Park D C. A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 2004, 36(4): 630-633.

[38] Li W, Li M, Su Z, Zhu Z. A deep-learning approach to facial expression recognition with candid images. In *Proc. the 14th IAPR International Conference on Machine Vision Applications*, May 2015, pp.279-282.

[39] Goodfellow I J, Erhan D, Carrier P L *et al.* Challenges in representation learning: A report on three machine learning contests. In *Proc. the 20th International Conference on Neural Information Processing*, Nov. 2013, pp.117-124.

[40] Wu T, Turaga P, Chellappa R. Age estimation and face verification across ageing using landmarks. *IEEE Transactions on Information Forensics and Security*, 2012, 7(6): 1780-1788.

[41] Giannopoulos P, Perikos I, Hatzilygeroudis I. Deep learning approaches for facial emotion recognition: A case study on FER-2013. In *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications*, Hatzilygeroudis I, Palade V (eds.), Springer, 2018, pp.1-16.

[42] Georgescu M I, Ionescu R T, Popescu M. Local learning with deep and handcrafted features for facial expression recognition. arXiv:1804.10892, 2018. https://arxiv.org/pdf/1804.10892.pdf, Dec. 2019.

[43] Abadi M, Barham P, Chen J *et al.* TensorFlow: A system for large-scale machine learning. In *Proc. the 12th USENIX Symposium on Operating Systems Design and Implementation*, Nov. 2016, pp.265-283.

[44] Caroppo A, Leone A, Siciliano P. Facial expression recognition in ageing adults: A comparative study. In *Ambient Assisted Living*, Leone A, Caroppo A, Rescio G *et al.* (eds.), Springer, 2018, pp.349-359.

[45] Li W, Tsangouri C, Abtahi F, Zhu Z. A recursive framework for expression recognition: From web images to deep models to game dataset. *Machine Vision and Applications*, 2018, 29(3): 489-502.

[46] Wang X, Wang X, Ni Y. Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational Intelligence and Neuroscience*, 2018, Article No. 7208794.

[47] Ionescu R T, Popescu M, Grozea C. Local learning to improve bag of visual words model for facial expression recognition. In *Proc. the 2013 ICML Workshop on Challenges in Representation Learning*, June 2013.

[48] Benitez-Quiroz C F, Srinivasan R, Feng Q, Wang Y, Martinez A M. EmotioNet challenge: Recognition of facial expressions of emotion in the wild. arXiv:1703.01210, 2017. https://arxiv.org/abs/1703.01210, Dec. 2019.

[49] Mollahosseini A, Hasani B, Mahoor M H. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2019, 10(1): 18-31

[50] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.770-778.

[51] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.2818-2826.

[52] Szegedy C, Ioffe S, Vanhoucke V, Alemi A A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proc. the 31st AAAI Conference on Artificial Intelligence*, February 2017, pp.4278-4284.

[53] Chollet F. Xception: Deep learning with depthwise separable convolutions. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.1800-1827.

Andrea Caroppo received his Master's degree in computer science engineering in 2004 from the University of Lecce, Lecce. From 2004 to 2006 he was a researcher fellow at the Italian National Research Council (CNR), Institute of Study of Intelligent Systems for Automation (ISSIA) in Bari (Italy). His research interests were in the area of image and video processing/coding, neural networks, motion estimation in video sequences and multidimensional signal processing. From 2012 to 2017 he was a researcher fellow at the National Research Council (CNR) of Italy, Institute for Microelectronics and Microsystems (IMM) in Lecce (Italy). Since 2018 he has been a researcher in the same institute. He is interested in signal and image processing, pattern recognition, computer vision and development of enabling technologies for healthcare with particular focus on the new Ambient Assisted Living (AAL) technologies. He is the author of more than 30 papers in national and international journals and conference proceedings.

Alessandro Leone received his Master's degree in computer science engineering in 2003 from the University of Lecce, Lecce. Since 2003 he has been a researcher at the National Research Council of Italy, Institute for Microelectronics and Microsystems in Lecce. He is interested in signal and image processing, pattern recognition, computer vision and smart multi-sensorial systems with particular focus on the new ambient assisted living technologies. Eng. Leone is the technical coordinator of the Signal&Image Processing Laboratory and he is mainly involved in development of enabling technologies for healthcare: fall detection & prevention, neurodegenerative cognitive rehabilitation, interoperability platforms and smart wearable sensors for vital signs monitoring. He is the author of more than 80 papers in national and international journals and conference proceedings. He is the technical coordinator of InnovAALab (the Apulian Living Lab on "Healthy, Active & Assisted Living") which is hosted by InnovAAL — PPP for research, development and testing of new technologies and services on AAL. Moreover, since 2014 Dr. Leone has been a member of the Europen AAL Forum Program Committee.

Pietro Siciliano received his Master's degree in physics in 1985 from the University of Lecce, Lecce. He took his Ph.D. degree in physics in 1989 at the University of Bari. He is currently a director of research at the Institute for Microelectronics and Microsystems (IMM-CNR) at National Research Council of Italy in Lecce, where he has been working for many years in the field of sensors, MEMS, microsystems, being in charge of the sensors and microsystems group. He is the author of about 350 scientific papers. Dr. Siciliano is a referee and a member of the advisory board of international journals. He has been responsible for several national (FISR, FIRB, PON) and international (V, VI and VII EU Framework) projects at IMM-CNR. He is a member of the Steering Committee of AISEM, the Italian Association on Sensors and Microsystems. He is the director of the section in Lecce of IMM-CNR. He is the president of the Italian Association on "Ambient Assisted Living" (AitAAL), responsible of "INNOVAAL", the Public-Private Partnership on Active & Assisted Living and President of the National Technological Cluster for "Smart Living Technologies".