Facial Expression Recognition in Older Adults using Deep Machine Learning

Andrea Caroppo, Alessandro Leone and Pietro Siciliano

National Research Council of Italy, Institute for Microelectronics and Microsystems, Lecce, Italy
{andrea.caroppo,alessandro.leone,pietro.siciliano}@le.imm.cnr.it

Abstract. Facial Expression Recognition is still one of the challenging fields in pattern recognition and machine learning science. Despite efforts made in developing various methods for this topic, existing approaches lack generalizability and almost all studies focus on more traditional hand-crafted features extraction to characterize facial expressions. Moreover, effective classifiers to model the spatial and temporary patterns embedded in facial expressions ignore the effects of facial attributes, such as age, on expression recognition even though research indicates that facial expression manifestation varies with ages. Although there are large amount of benchmark datasets available for the recognition of facial expressions, only few datasets contains faces of older adults. Consequently the current scientific literature has not exhausted this topic. Recently, deep learning methods have been attracting more and more researchers due to their great success in various computer vision tasks, mainly because they avoid a process of feature definition and extraction which is often very difficult due to the wide variability of the facial expressions. Based on the deep learning theory, a neural network for facial expression recognition in older adults is constructed by combining a Stacked Denoising Auto-Encoder method to pre-train the network and a supervised training that provides a fine-tuning adjustment of the network. For the supervised classification layer, the M-class softmax classifier was implemented, where M is the number of expressions to be recognized. The performance are evaluated on two benchmark datasets (FACES and Lifespan), that are the only ones that contain facial expressions of the elderly. The achieved results show the superiority of the proposed deep learning approach compared to the conventional non-deep learning based facial expression recognition methods used in this context.

Keywords: Ambient Assisted Living, Facial Expression Recognition, Mood, Deep Machine Learning, Stacked Denoising Auto-Encoder, Graphical Processing Units (GPU) computing.

1 Introduction

Ambient Assisted Living (AAL) addresses the needs of the ageing population to reduce innovation barriers of forthcoming promising markets, but also to lower future

social security costs. AAL aims, by the use of intelligent products and the provision of remote services including care services, at extending the time older people can live in their home environment by increasing their autonomy and assisting them in carrying out activities of daily living (ADLs). Consequently, in the current context, it is a challenge to provide new technologies for automatic recognition of emotion or moods, with the purpose to improve the quality of life of older adults [1].

Facial expression recognition (FER) has been attracting considerable attention due to its wide variety of applications, such as robotics, communications, security, medical and assistive technology. Moreover, different facial expressions can reflect the emotions and also mental activities of the observed subject. Consequently, it is crucial to investigate new methodologies for the automatic recognition of facial expressions (mainly performed by the older adults) for the implementation of intelligent systems able to customize, for example, the response of the environment.

FER is effected by many factors among which one of the most discriminating is the age [2,3,4]; 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 hand-crafted features extraction may be inadequate for the classification of FER performed by older adults.

Recently, a viable alternative to such traditional feature design is represented by deep learning (DL) algorithms which straightforwardly leads to automated feature learning [5]. Research using DL techniques could make better representations and create innovative models to learn these representations from unlabelled data. Some of the DL techniques like Convolutional Neural Networks, Deep Boltzmann Machine, Deep Belief Networks and Stacked Auto-Encoders are applied to practical applications like pattern analysis, audio recognition, computer vision and image recognition where they produce challenging results on various tasks [6].

Although there has been much work on automatic FER using DL, the algorithms have been experimentally validated primarily on young faces. The facial expressions on older faces has been totally excluded or they have been taken into consideration jointly with faces representatives of different ages.

In this paper, we focus on the Stacked Denoising Auto-Encoder (SDAE) method [7] for FER in older adults, since Denoising Auto-Encoder (DAE) is very robust to noise which is present in real contexts under different declinations, and SDAE can obtain higher level features, through which we are able to distinguish facial expressions of elderly. Moreover, since sparsity of features might improve the separation capability, we utilized an activation function in SDAE to extract high level and sparse features which, from the analysis of the achieved results, allows a significant improvement in FER of older adults, thus confirming the goodness of the approach.

The remainder of this paper is organized as follows: Section 2 describes related work, Section 3 reports some details about the implemented SDAE approach, Section 4 discussed the experimental results and, finally, conclusions are summarized in Section 5.

2 Related Work

Ekman's initial research [8] determined that there were six basic classes in FER: anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA) and surprise (SU).

Proposed solutions for the classification of the aforementioned facial expressions can be divided into two main categories: the first category includes solutions that classify facial expressions by processing a set of consecutive images while, the second one, includes approaches which perform 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 constrain 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, there are two types of facial feature extraction methods: geometric feature-based methods and appearance-based methods.

Geometric features are able to depict the shape and locations of facial components such as mouth, nose, eyes and brows. The main purpose of geometric feature-based methods is to use the geometric relationships between facial feature points to extract facial features. Three typical geometric feature-based extraction methods are active shape models (ASM) [9], active appearance models (AAM) [10] and scale-invariant feature transform (SIFT) [11]. Extracting geometric features usually requests an accurate feature point detection technique. This is difficult to implement in real-world complex background. In addition, geometric feature-based methods easily ignore the changes in skin texture such as wrinkles and furrows that are usually accentuated by the age of the subject.

Appearance-based methods 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 [12], Local Binary Patterns (LBP) [13] and Histogram of Oriented Gradient (HOG) [14].

However, all the above mentioned methodologies require a process of feature definition and extraction very daunting; 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. As a consequence such approaches may not achieve the same recognition performance, in the considered application context, as they have been validated almost always through datasets containing only young faces. It seems therefore very important to analyze approaches that can make the recognition of facial expressions of the older adults more efficient, since many research studies in literature have shown that facial expressions of elderly are broadly different from those of young or middle-aged for a number of reasons. For example, in [15] researchers found that the expressions of older adults (women in this

case) were more telegraphic in the sense that their expressive behaviors tended to involve fewer regions of the face, and yet more complex in that they used more blended or mixed expressions when recounting emotional events. These changes, in part, account for why the facial expressions of older adults are more difficult to read. Another study showed that when emotional memories were prompted and subjects asked to relate their experiences, older 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 (FACS) [16].

One of the other changes that comes with age, making facial expression of older adults more difficult to recognize, involves the wrinkling of the facial skin and the sag of facial musculature. Of course, part of this is due to biologically based aspects of aging, but individual differences also appear linked to personality process, as demonstrated in [17].

To the best of our knowledge, only few works in literature address the problem of FER in older adults. In [18] the authors perform a computational study within and across different age groups and compare the FER accuracies, founding that the recognition rate is influenced significantly by human aging. 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 [12] 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 [19] recognizes emotions of ageing adults using an Active Shape Model [9] for feature extraction. To train the model the authors employ three benchmark datasets that do not contain adult faces getting an average accuracy of 82.7% on the same datasets. Tests performed on older 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.

3 Methodology

In this work, a deep learning method for FER in older adults was implemented, based on stacking layers of DAE. Before the application of the methodology, the implemented pipeline performs a pre-processing task on the input images. Once the images are pre-processed they can be either used to train the network or to test it (i.e. recognition step). In the training step, a set of pre-processed images are given to the network so that the best set of network weights for classification can be found. In the testing step, the network is configured with the weight set found during the training and the recognitions are performed.

The first step of the pre-processing procedure is a cropping of the input image. This step aims to keep the methodology focused only on specific regions, removing all background information and image patches that are not related to the expression. The cropping region is automatically delimited based on the original Viola-Jones face detector [20]. The second step of the pre-processing procedure is a down-sampling of

Fig. 1. Pipeline of pre-processing task applied to each image: (a) original image, (b) automatic cropping of the face region using Viola-Jones algorithm, (c) down-sampling of the facial region (96x96 pixel) and conversion of the RGB image into a grayscale intensity image

the input image. In fact, after the cropping step, the images will be of different sizes. Therefore, the images are down-sampled, using a linear interpolation, to 96×96 pixels in order to remove the variation in face size and keep the facial parts in the same pixel space. Finally, the last step convert the pre-processed image into a grayscale image (Figure 1).

3.1 Overview of the proposed deep learning approach

A generic neural network (NN) that uses auto-encoders (AE) trains the network by constraining the output values to be equal to the input values, using the error generated in the reconstruction of the input for the adjustment of the weights of each layer of the NN. The input data are represented in a good way through the features learned by AE whose training is performed in unsupervised way, since the label information is not required. DAE is an extension of AE but is more robust. A general DAE contains three layers: input layer, hidden layer, and output layer, where the hidden layer and output layer are also called encoding layer and decoding layer, respectively. More specifically, an AE takes an input $\boldsymbol{x} \in \mathbb{R}^p$ where $\boldsymbol{p}$ represents the dimension of the input data. DAE is an AE with noise corruptions that produces a corrupted version $\tilde{\boldsymbol{x}}$ of the original input. A typical way of corruption is randomly masking elements of $\boldsymbol{x}$ as zeros or adding Gaussian noise to $\boldsymbol{x}$.

The latent representation (encoding) of DAE is obtained by a nonlinear transformation: $\boldsymbol{y} = s(\boldsymbol{W}\tilde{\boldsymbol{x}} + \boldsymbol{b})$ where $\boldsymbol{y} \in \mathbb{R}^q$, $\boldsymbol{q}$ is the number of units in the hidden layer and $\boldsymbol{y}$ denotes the output of the hidden layer.

The matrix $\boldsymbol{W} \in \mathbb{R}^{qxp}$ is the input-to-hidden weights, $\boldsymbol{b}$ denotes the bias and $s()$ is the activation function of the hidden layer. In the present work the rectified linear unit *ReLU* is used as activation function [21].

The latent representation $\boldsymbol{y}$ is then mapped back (with a decoder) into a reconstruction $\boldsymbol{z}$ of the same shape as $\boldsymbol{x}$. The mapping happens through a similar transformation, e.g.: $\boldsymbol{z} = s(\widetilde{\boldsymbol{W}}\boldsymbol{y} + \widetilde{\boldsymbol{b}})$ where $\boldsymbol{z} \in \mathbb{R}^p$ is the output of DAE and should be seen as a prediction of $\boldsymbol{x}$, given $\boldsymbol{y}$. Optionally, the weight matrix $\widetilde{\boldsymbol{W}}$ of the reverse mapping may be constrained to be the transpose of the forward mapping: $\widetilde{\boldsymbol{W}} = \boldsymbol{W}^{\boldsymbol{T}}$. This is referred to as tied weights. The biases $\boldsymbol{b}$ and $\widetilde{\boldsymbol{b}}$ are still different even when the weights are tied. DAE is trained by minimizing the reconstruction error, consequently the reconstruction error is used as the cost function or objective function.

Moreover, DAE can be stacked to obtain high level features, resulting in SDAE approach. Each DAE with one hidden layer is trained independently, and for this reason the training of SDAE is layer-wise. In the presented methodology, the step after SDAE training consists in removing decoding layers with the purpose to retain the encoding layers that produce features. The M-class softmax classifier is added to the output layer for the classification task and a fine-tuning adjustment of the network is obtained via gradient descent optimization method like backpropagation [22] where the initial weights of the output layer are randomly initialized while the weights of the hidden layers are the ones obtained in the pre-training phase.

3.2 SDAE for the Recognition of Facial Expression in Older Adults

The theoretical description given in the previous section can be reported to the problem of FER in older adults. Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m\} \subset \mathbb{R}^p$ be a set of m unlabelled training examples (i.e., facial expression images), the SAE aims to train the network by requiring the output data $\boldsymbol{z}$ to reconstruct the input data $\boldsymbol{x}$, which is also called reconstruction-oriented training. Such task is accomplished by minimizing with respect to $\boldsymbol{W}$ and $\boldsymbol{b}$ the following cost function:

$$\mathrm{J_E}(\boldsymbol{W}, \boldsymbol{b}) = \frac{1}{2m}\sum_{i=1}^{m}\|\boldsymbol{z}_i - \boldsymbol{x}_i\|^2 + \frac{\lambda}{2}\sum_{k=1}^{l-1}\sum_{i=1}^{n_k}\sum_{j=1}^{n_{k+1}}\left(w_{ji}^{(k)}\right)^2 + \beta\sum_{k=2}^{l-1}\sum_{j=1}^{n_k}\mathrm{KL}\left(\rho||\hat{\rho}_j^{(k)}\right) \tag{1}$$

where λ is the weight decay parameter (typically expressed as regularization term and fixed at 0.003 in the present work), β is a constant value that manages the sparsity penalty term, $\hat{\rho}_j^{(k)} = \frac{1}{m}\sum_{i=1}^{m} y_j^{(k)}(\boldsymbol{x}_i)$ with $y_j^{(k)}(\boldsymbol{x}_i)$ denoting the activation of the corresponding unit when the input $\boldsymbol{x}_i$ is given to the network, ρ is a sparsity parameter typically near to zero and the term $\mathrm{KL}(\rho||\hat{\rho}) = \rho\log\frac{\rho}{\hat{\rho}} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}}$ is the Kullback-Leibler (KL) divergence between two Bernoulli distributions with mean ρ and $\hat{\rho}$, respectively [23].

The unsupervised feature learning is followed by a supervised classification layer, namely the M-class softmax classifier. Let $T = \{(\boldsymbol{x}_1, o_1), (\boldsymbol{x}_2, o_2), \dots, (\boldsymbol{x}_n, o_n)\}$ be a training set with $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^p$ images with facial expressions taken as examples and $o_1, o_2, \dots, o_n \in \{A_1, \dots, A_M\}$ be the corresponding labels indicating the different

expressions which we intend to classify ($A_1 = "Anger"$, $A_2 = "Disgust"$, $A_3 = Fear"$, $A_4 = "Happy"$, $A_5 = "Sad"$, $A_6 = "Neutral"$). The softmax classification is done by minimizing the following cost function with respect to parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2 \dots\ \boldsymbol{\theta}_M] \in \mathbb{R}^{p,m}$:

$$\mathrm{J_F}(\boldsymbol{\theta}) = -\sum_{i=1}^{m}\sum_{j=1}^{M} \mathbf{1}\{o_i = A_j\} P\{o_i = A_j | \boldsymbol{x}_i; \boldsymbol{\theta}\} \quad (2)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function ($\mathbf{1}(C) = 1$ if condition C is true, $\mathbf{1}(C) = 0$ if condition C is false), and the conditional probability $P\{o_i = A_j | \boldsymbol{x}_i; \boldsymbol{\theta}\} = \log\left(e^{\boldsymbol{\theta}_j^T \boldsymbol{x}_i} / \sum_{k=1}^{M} e^{\boldsymbol{\theta}_k^T \boldsymbol{x}_i}\right)$ should be large when $\boldsymbol{x}_i$ belongs to the class A_j and small otherwise.

4 Results

In this section the evaluation of the DL approach described is reported. To validate our model a series of experiments were conducted using the age-expression datasets FACES [24] and Lifespan [25].

The FACES dataset involves 171 people showing six different expression (anger, disgust, fear, happy, sad and neutral). The subjects are divided into three main groups according to their age (young: 19-31 years old, middle-aged: 39-55 years old, older: 69-80 years old). For each subject 2 examples of each expression are saved, so in total the dataset consists of 171*2*6=2052 frontal images.

The Lifespan dataset is a collection of faces of subjects from different ethnicities showing different expressions. The expression subsets have the following sizes: 580, 258, 78, 64, 40, 10, 9, and 7 for neutrality, happiness, surprise, sadness, annoyed, anger, grumpy, and disgust, respectively.

For the performance evaluation of the methodology only facial expression of older adults were considered and pre-processed. Consequently, the images that belongs to FACES used for training and testing are 684 (57 older adults that perform twice the six expressions), whereas only 223 neutral faces and 69 happy faces from Lifespan dataset were pre-processed.

	# of images						**Total**
	anger	**disgust**	**fear**	**happy**	**sad**	**neutral**	
FACES	114	114	114	114	114	114	**684**
Lifespan				69		223	**292**

Table 1. Two aging datasets (FACES and Lifespan) with the corresponding number of facial expressions used for the evaluation of the proposed methodology

Fig. 2. Some examples of expressions performed by older adults from the FACES database (line up) and Lifespan database (bottom line)

4.1 Performance Evaluation

The training and testing phase were performed on Intel i7 3.5GHz workstation with 16GB DDR3 and equipped with GPU NVidia Titan X using the Python library for machine learning Tensorflow, developed for implementing, training, testing and deploying deep learning models [26].

Network configuration contains four parameters, which are the number of hidden layers, the number of units in hidden layer, the sparsity parameter value (ρ) and the standard deviation of Gaussian noise (used for the production of the corrupted version of the original input image). The number of hidden layers (HL) is selected in the range from 1 to 3, the number of units is chosen in order to obtain different compression factors of the input image, ρ has been tuned in the range 0.05-0.3 and the standard deviation of Gaussian noise is selected from [0.2, 0.4, 0.6, 0.8]. The optimal selection of these parameters is obtained according to the optimal classification results on the testing data. As the pre-processing procedure returned images of the same size for the two datasets, the same parameter configuration was used for tests on both datasets, in particular considering two hidden layers and a Gaussian noise of 0.6 the classification of the facial expressions reaches the highest accuracy in both datasets.

Several experiments have been conducted with the aim of evaluating the optimum number of nodes in each hidden layer. Table 2 reports the most significant configuration settings (CS) taken into account (the compression factor with respect to the input data size is reported in brackets).

Figure 3 and 4 report, for each dataset, the average detection rate (accuracy) obtained at varying of the sparsity parameter ρ, which is the parameter involved in KL divergence formula reported in section 3.4. In this formula ρ and β control sparseness. In particular, ρ is the expected activation of a hidden unit (averaged across the training set). In other words, the representation will become sparser and sparser as it becomes smaller. This sparseness is imposed by adjusting the bias term, and β controls the size of its updates. In the performed test, the value of β was set to 3.

	# of nodes HL1	*# of nodes HL2*
configuration setting(CS) 1	2304 (4)	2304 (4)
configuration setting (CS) 2	2304 (4)	1152 (8)
configuration setting (CS) 3	1152 (8)	576 (16)
configuration setting (CS) 4	1152 (8)	1152 (8)
configuration setting (CS) 5	576 (16)	576 (16)

Table 2. Number of nodes for each hidden layer at varying of the five considered configuration settings of the network

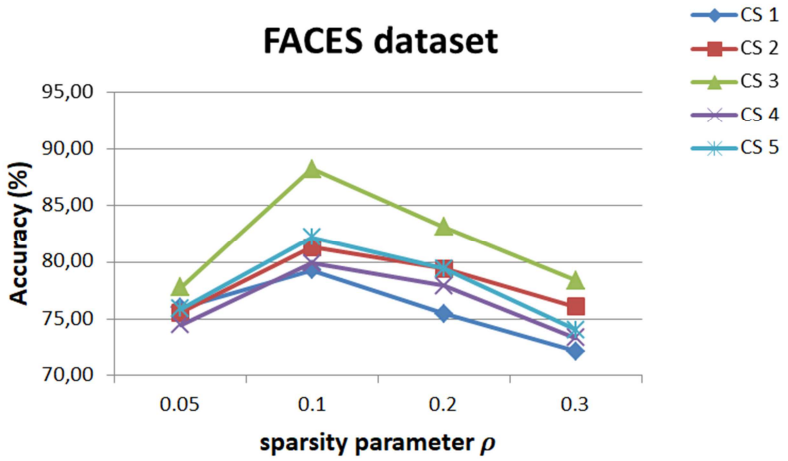

Fig. 3. Average accuracy obtained at varying of the sparsity parameter ρ for FACES dataset

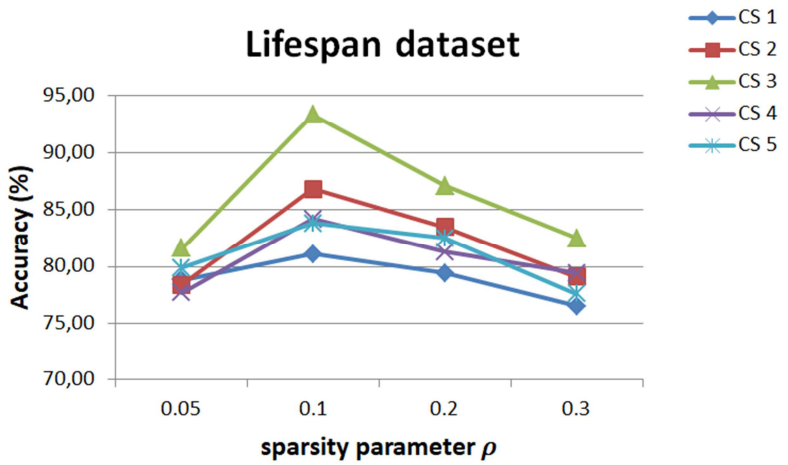

Fig. 4. Average accuracy obtained at varying of the sparsity parameter ρ for Lifespan dataset

The average accuracy measured allows to set the optimum number of nodes for each hidden layer which is equal to 1152, with a compression factor of about 8 times for HL1 and 576 (compression factor of 16 times) for HL2. In addition, the trend of the accuracy value demonstrates that an increase in value of ρ worsens the system's overall performance, consequently it was considered advisable to not carry out experiments with values greater than 0.3.

In a multi-class recognition problem, as the FER one, the use of an average performance value among all the classes could be not 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 3 and 4.

The results are based on three distinct percentages (65%, 70%, and 75%) of sample dataset for training purpose. Analyzing the trend of recognition rate, this has led to the conclusion that for both datasets training samples do not play significant role in increasing and decreasing the recognition rate.The numerical results obtained in terms of recognition rate of each class of facial expression makes 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 proposed pipeline achieved an average detection rate value over 90.7 % for all the tested datasets and that, as expected, its FER performance decreased when the number of classes, and consequently the problem complexity, increased. In fact, in the case of the FACES dataset with 6 expressions, the obtained average accuracy was of 88.2 % whereas the average accuracy obtained on Lifespan dataset was 93.3%.

		Estimated (%)					
		Anger	*Disgust*	*Fear*	*Happy*	*Sad*	*Neutral*
Actual (%)	*Anger*	***91,3***	0	0	0	5,8	2,9
	Disgust	6,4	***87,2***	0	1,6	3,2	1,6
	Fear	0	0	***91,6***	2,8	5,6	0
	Happy	1,7	5,1	1,7	***89,8***	0	1,7
	Sad	1,6	0	6,5	0	***84,1***	7,8
	Neutral	5,5	3,7	5,6	0	0	***85,2***

Table 3. Confusion Matrix of six basic expression on FACES dataset

		Estimated (%)	
		Neutral	*Happy*
Actual (%)	*Neutral*	***92,9***	7,1
	Happy	6,3	***93,7***

Table 4. Confusion Matrix of two basic expression on Lifespan dataset

Going into a more detailed analysis on the results reported in Table 2, anger and fear are the facial expression better recognized, whereas sad and neutral are the facial expression confused the most. Finally, sad is the facial expression with the lowest accuracy.

4.2 Comparison with Non-Deep Learning Approaches

In this section the achieved results are compared with those of the leading state-of-the-art FER solutions. Differently from other research fields, in the FER one there is not a shared dataset to be used as benchmark for a fair evaluation of different algorithms.

The most used datasets for comparing a new FER methodology are Japanese Female Facial Expression (JAFFE) [27] and the Extended Cohn-Kanade (CK+) [28], but unfortunately these two datasets do not contain images of facial expressions performed by older adults. In order to accomplish this crucial task, in this work two popular FER methods, selected among the most powerful ones in the literature, have been implemented from scratch.

The first is a geometric feature-based method in which the feature extraction step is performed by an active shape model (ASM) able to extract the landmark points from each face. Then, FER is performed based on these geometric features using Support Vector Machine (SVM) as classifier.

The second is an appearance-based method that uses Local Binary Pattern (LBP) for feature extraction step and SVM as classifier. Table 5 reports the comparison results demonstrating that the proposed approach gave the best average recognition rate for both datasets used. In particular, the deep learning approach improves the overall performance when more facial expressions to distinguish are considered. In fact the table shows less differences in the obtained average accuracy on Lifespan dataset, for which only two different facial expressions were considered.

Approach	*Dataset*	*Avg Accuracy (%)*
ASM+SVM	FACES	85,3
	Lifespan	92,1
LBP+SVM	FACES	84,5
	Lifespan	91,4
Proposed	FACES	**88,2**
	Lifespan	**93,3**

Table 5. Performance comparison of our approach versus different state-of-the-art approaches (bold value indicates the best results)

5 Conclusions

A deep learning approach based on SDAE for the automatic recognition of facial expression in older adults has been presented and validated through experiments performed on two benchmark datasets (the only ones that contain facial expressions performed by older people). The testing phase of the implemented methodology has allowed to outline the correct parameters for the definition of the best model for the used datasets. After tuning of these parameters, numerical results obtained for FER on both datasets demonstrate the goodness of the implemented approach. Moreover, two non-deep learning approaches were implemented and tested on the same datasets and the results obtained have demonstrated the superiority of the presented methodology with respect classical non-deep learning approaches.

Future works will deal with two main aspects. On the one hand the methodology will be tested in the field of assistive technologies, first validating it in a smart home setup and after testing the pipeline in a real AAL environment, which is the older person's home. In particular, the idea is to develop an application that uses the webcam integrated in TV or smartphone/tablet camera with the purpose to recognize the facial expression of older 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 older people living alone at their homes, for example by subjecting it to external sensory stimuli, such as music and images. On the other hand, a more wide analysis of how a non-frontal view of the face can affect the facial expression detection rate using the implemented approach 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.

References

1. Castillo, J.C., Fernández-Caballero, A., Castro-González, Á., Salichs, M.A. and López, M.T.: A framework for recognizing and regulating emotions in the elderly. In International Workshop on Ambient Assisted Living (pp. 320-327). Springer International Publishing. (2014)
2. Guo, G., Guo, R., & Li, X.: Facial expression recognition influenced by human aging. IEEE Transactions on Affective Computing, 4(3), 291-298. (2013)
3. Algaraawi, N., & Morris, T.: Study on Aging Effect on Facial Expression Recognition. In Proceedings of the World Congress on Engineering (Vol. 1). (2016)
4. Wang, S., Wu, S., Gao, Z., & Ji, Q.: Facial expression recognition through modeling age-related spatial patterns. Multimedia Tools and Applications, 75(7), 3937-3954. (2016)
5. LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning. Nature, 521(7553), pp.436-444. (2015)
6. Yu, D., & Deng, L.: Deep learning and its applications to signal and information processing [exploratory dsp]. IEEE Signal Processing Magazine, 28(1), 145-154. (2011)

7. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11(Dec), 3371-3408. (2010)
8. Ekman, P., Rolls, E. T., Perrett, D. I., & Ellis, H. D.: Facial expressions of emotion: An old controversy and new findings [and discussion]. Philosophical Transactions of the Royal Society B: Biological Sciences, 335(1273), 63-6.9 (1992)
9. Shbib, R., & Zhou, S.: Facial expression analysis using active shape model. International Journal of Signal Processing, Image Processing and Pattern Recognition, 8(1), 9-22. (2015)
10. Cheon, Y. and Kim, D.: Natural facial expression recognition using differential-AAM and manifold learning. Pattern Recognition, 42(7), pp.1340-1350. (2009)
11. Soyel, H. and Demirel, H.: Facial expression recognition based on discriminative scale invariant feature transform. Electronics letters, 46(5), pp.343-345. (2010)
12. Gu, W., Xiang, C., Venkatesh, Y.V., Huang, D. and Lin, H.: Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Pattern Recognition, 45(1), pp.80-91. (2012)
13. Shan, C., Gong, S. and McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6), pp.803-816. (2009)
14. Chen, J., Chen, Z., Chi, Z. and Fu, H.: Facial expression recognition based on facial components detection and hog features. In International Workshops on Electrical and Computer Engineering Subfields (pp. 884-888). (2014)
15. Malatesta C. Z. & Izard C. E.: The facial expression of emotion: young, middle-aged, and older adult expressions, in Emotion in Adult Development, eds Malatesta C. Z., Izard C. E., editors. (London: Sage Publications;), 253–273. (1984)
16. 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, 7(4), 551. (1992)
17. Malatesta, C. Z., Fiore, M. J., & Messina, J. J.: Affect, personality, and facial expressive characteristics of older people. Psychology and aging, 2(1), 64. (1987)
18. Guo, G., Guo, R., & Li, X.: Facial expression recognition influenced by human aging. IEEE Transactions on Affective Computing, 4(3), 291-298. (2013)
19. 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, 1-12. (2017)
20. Viola, P., & Jones, M. J.: Robust real-time face detection. International journal of computer vision, 57(2), 137-154. (2004)
21. Nair, V., & Hinton, G. E.: Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814). (2010)
22. Rumelhart, D. E., Hinton, G. E., & Williams, R. J.: Learning representations by back-propagating errors. Cognitive modeling, 5(3), 1. (1988)
23. Kullback, S., & Leibler, R. A.: On information and sufficiency. The annals of mathematical statistics, 22(1), 79-86. (1951)
24. Ebner, N. 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, 42(1), 351-362. (2010)
25. Minear, M., & Park, D. C.: A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, & Computers, 36(4), 630-633. (2004)

26. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M.: TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA. (2016)
27. Lyons, M., Akamatsu, S., Kamachi, M., & Gyoba, J.: Coding facial expressions with gabor wavelets. In Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on (pp. 200-205). IEEE. (1998)
28. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 94-101). IEEE. (2010)