Article

Enhancing Mouth-based Emotion Recognition using Transfer Learning

Valentina Franzoni [1,†,‡,*]0000-0002-2972-7188, Giulio Biondi [2,‡]0000-0002-1854-2196, Damiano Perri [2,‡]0000-0001-6815-6659 and Osvaldo Gervasi [1,‡,*]0000-0003-4327-520X

1 University of Perugia; valentina.franzoni@dmi.unipg.it, osvaldo.gervasi@unipg.it
2 University of Florence; giulio.biondi@unifi.it, damiano.perri@unifi.it
* Correspondence: osvaldo.gervasi@unipg.it, valentina.franzoni@dmi.unipg.it [OG][VF]
† Via Vanvitelli 1, 06123 Perugia, Italy
‡ These authors contributed equally to this work.

Abstract: The paper concludes the first research on mouth-based Emotion Recognition (ER), adopting a Transfer Learning (TL) approach. Transfer Learning results paramount for mouth-based emotion ER, because a few data sets are available, and most of them include emotional expressions simulated by actors, instead of adopting a real-world categorization. Using TL we can use fewer training data than training a whole network from scratch, thus more efficiently fine-tuning the network with emotional data and improving the convolutional neural network accuracy in the desired domain. The proposed approach aims at improving the Emotion Recognition dynamically, taking into account not only new scenarios but also modified situations with respect to the initial training phase, because the image of the mouth can be available even when the whole face is visible only in an unfavourable perspective. Typical applications include automated supervision of bedridden critical patients in an healthcare management environment, or portable applications supporting disabled users having difficulties in seeing or recognizing facial emotions. This work takes advantage from previous preliminary works on mouth-based emotion recognition using CNN deep-learning, and has the further benefit of testing and comparing a set of networks on large data sets for face-based emotion recognition well known in literature. The final result is not directly comparable with works on full-face ER, but valorizes the significance of mouth in emotion recognition, obtaining consistent performances on the visual emotion recognition domain.

Keywords: Transfer Learning; Convolutional Neural Networks; Emotion Recognition

1. Introduction

Visual emotion recognition has been widely studied, as one of the first affective computing techniques, mainly based on visual features of the face expression combining features about eyes, mouth and various facial elements at the same time. Several different approaches at visual recognition obtained different grades of classifications for different visual recognition techniques [1–3]. Recently, works using only the mouth for facial emotion recognition obtained promising results, still not gaining the proper attention in the state-of-the-art. Such works used light convolutional neural networks to detect basic emotions from smartphone or computer camera devices, to produce a textual, audio or graphic feedback for humans, or digital output to support other services, mainly for healthcare systems [4]. Being focused on a personal training on a single individual, their neural network is able to obtain a very good result with a relatively small data set of images, thought for example to detect particular states needing an immediate medical intervention, or changes over the time underlying a degenerative health condition.

Our analysis focuses on emotion classification from the mouth only, to study how much the mouth is expressive in emotion expression, and which accuracy can be obtained by the mouth with respect to a full-face recognition. We analyse the position and curve of lips through a neural network. Recently,

some of our preliminary works [1,4,5] obtained very promising results regarding emotion analysis using mouth, and this is our first attempt to recap and complete our full study on mouth-based emotion recognition on extensive data sets. All our previous works used convolutional neural networks (CNN) to reach their goals on the topic, using a self-collected data set. In-deep work has been primarly made on three emotions (i.e., joy, disgust, neutral) [5], where disgust is the less studied in literature.

After proving that the mouth can be itself a promising element for emotion recognition, in this work we focus on the mouth as a unique element for facial-expression-based emotion recognition using deep learning advanced techniques. The scope of this approach is to enhance the mouth-based approach to emotion recognition in order to generalize the previous experimentation on a multiple-users data set. To this aim, advanced deep learning techniques have been implemented, such as knowledge transfer with elements of continuous learning. Such techniques are particularly suitable to be applied in the healthcare sector. For instance, connecting this architecture to appropriate services can help users to effectively convey emotions in an automated way, e.g. providing augmented emotional stimuli to users affected from autism or other conditions involving social relationship abilities, when users experiment difficulties in recognizing emotions expressed by other people.

Another example is a system able to recognize severe conditions and call for intervention a human assistant, e.g., for hospitalized patients feeling intense pain or needing help for a depression crisis, which may then add a proper feedback to enhance continuous learning.

The main problem in facial-based emotion recognition is the lack of proper data sets of images for training. Most of the available data sets take in consideration only the Ekman model [6] or its subsets [5], making actually nearly impossible to use more complex and complete models, such as Plutchik [7], or models based on emotional affordance [8]. Moreover, the main problem of the various image data set is that they contain non-genuine expressions (e.g. made by professional actors) rather than spontaneous and natural forms of facial expression [9]. Being deep learning able to extract features not even recognizable by a human agent, we consider all these items related to non-real emotions absolutely unable to train effectively a neural network. For these reasons, we focused for our first steps on a small subset of the Ekman model, including the neutral expression as a state of control for recognition results on emotions [4], where we invested in selecting good images from the available data sets, both from internet sources and from self-produced material. This is feasible for emotions which are easily triggered, e.g. joy and disgust, while is quite difficult to obtain for other emotions where ethical questions are involved in the stimulus, e.g., anger and fear. In order to improve results with relatively small set of images for each emotion, we decided to use transfer learning.

We tested the most promising neural networks [1] for face recognition and mouth recognition, e.g., lip reading [10], with a previous training on widely used large data sets of images including human figures [11]. This knowledge transfer allowed our neural networks to be pre-trained regarding low-level features, for instance: edges, corners and color distribution.

In the last layers we carried out an ad-hoc training dedicated to face recognition on emotional labels. This work concludes the experiments testing and comparing several CNNs on a widely-used data set [12] of human faces labeled with emotions from the Ekman model. As in the preliminary works, also for the final data set we provide a filtered version, where photos showing clearly simulated emotions, i.e. non spontaneous expressions, have been removed.

A straightforward application of the method is emotion/pain recognition for healthcare management and automated supervision of critical patients, e.g., in hospitals. The system can enhance the positive experience of patients, using emotion recognition during night or when a direct human assistance is not available, for an advanced detection of an initial pain or discomfort state of the patient, raising a signal and letting the sanitary staff be informed to react promptly, avoiding the patient suffering. We can plan supportive systems using emotion recognition to assist patients after car accidents, still in the emergency phase, in order to understand their pain level, or post-surgery, or for early recognition of depression states. If the system recognizes a critical situation such as the ones listed, it can report the case to nurses or send an intervention/checkout request for the patient's room.

Our emotion recognition system is planned to give a feedback to the user (e.g., text, emoticon, sound) and set a channel for information transfer to software using emotion recognition.

Another use case may be to support people with difficulties to see or interpret emotions, such as blind users, or people with autism spectrum disorders, or semi-paralyzed or terminally ill patients, not able to ask for help in case of need.

Mouth is an extremely important element of human face, almost symmetric and usually visible in any point of view, thus the perfect element to focus for face recognition in all those cases where the user cannot follow capturing rules such as posing in a particular perspective. Besides showing on an extensive data set how the mouth can provide sufficient information to classify the user emotions using advanced techniques of deep learning, focusing on the mouth as a sole source for emotion recognition (eventually added to others when available, for enhancing accuracy in practical applications) can provide interesting and useful information to understand the value of this element in an eventually multi-modal acquisition system.

Previous works were focused on the acquisition of data from a single user, i.e., where researchers train the network on a specific user's face. This approach lets a user train the network with precision on his/her face, reaching advanced recognition performance [1,4]. This work focuses on a more general approach using a multi-user data set, including face images from very different kind of people, regarding age, cultural background, gender, and appearance. To this aim, we use advanced methods of deep learning (e.g., transfer learning and continuous learning).

Transfer learning uses general-purpose neural networks pre-trained on very large data sets including different shapes, later fine-tuned on the classification domain of mouth-based emotion recognition. Our application can be easily adapted for continuous learning, given a domain where such method is useful and appropriate data are available, to track the weights of the neural network and prosecute the training later on. Such method should allow to enhance the precision of emotion recognition in real-time situations, where the final weights of a previous training can be used in a subsequent time. The collateral effect of such a technique is the requirement of greater computational capabilities and a semi-supervised system in order to stop and rollback in the case of evident errors or over fitting on a specific environment, e.g., a particular camera resolution or light situation in a specific place or timing.

Continuous learning can continuously adapt and enhance the network accuracy, leading on the go to a more reliable system, when used for a prolonged time. At the actual state of the art, one of the most relevant issues for our goal is the total lack of a sufficient number of images of real emotions. Most of the available data sets include a mixed set of real and fake emotions, performed by actors or pretended by the researchers themselves, which cannot be considered usable for a technique like deep learning, able to base the classification on very detailed features, which are different between a real and a fake emotion, even at a micro-expression level, and may vary in different cultures. It is mandatory to report that most of the available data sets provide images depicting actors. In fact, people who are taking a pose for a photograph (for example, they are smiling voluntarily in front of the camera), assume an artificial, distorted expression that is not generated autonomously by the human brain, even if it can be properly identified and recognized by the human brain, thanks to mirror neurons. It should be considered that there are emotions such as anger and fear that are not easily replicated in a laboratory, while emotions such as joy or the neutral state, on the other hand, can be easily triggered.

Another relevant limitation is that many data sets include only images of a particular light state, perspective, image resolution, which will lead the system over fitting that environment, and fail on others. Furthermore, most of the data sets include only emotions from the Ekman model, and never include information on more complex and precise models, such as the Plutchik model.Until a proper large data set will be ready for the test, any result has to be trivially considered preliminary. In this work, we discarded any image which we can recognize as a fake emotion. Keeping in mind all these considerations, we focused on a subset of the Ekman model, investing efforts in polishing the image data set from any fake emotional expression.

1.1. Convolutional Neural Networks

Convolutional Neural Networks are among the most used methods for affective image classification thanks to their flexibility for transfer learning, and easy tools available on the Web; they constitute a class of deep neural networks which prove particularly efficient for tasks on data organized in a grid topology, e.g. time series and visual inputs. Common applications include image recognition, segmentation, detection and retrieval[13–15], on which CNN have immediately achieved state-of-the-art performance. This major breakthrough can be attributed to three key factors boasted by CNN, i.e. sparse interactions, parameter sharing and equivariant representation [16]; massive amounts of training samples can be processed with improved efficiency and greatly reduced training times, using deeper networks with millions of parameters to learn more complex and descriptive image features. An additional advantage is the possibility of having a variable input size, in contrast with traditional, fully-connected networks which require fixed size input. In Convolutional Neural Networks, some of the traditional fully-connected layers are replaced with convolution layers, which scan through the ordered, grid-like structured data to process it in subsets, multiplying each subset by a matrix named *kernel* (or, equivalently in this context, *filter*), to produce a *feature map* as output. The process resembles how individual neurons respond to visual inputs; each neuron is responsible for a small portion of the input, called its *receptive field*, and ignores additional information. Training a fully-connected NN with the same feature recognition and, consequently, classification capabilities would require a much higher effort, e.g. for an image of size 1000x1000, training 1000000 weights for each neuron of a layer, whereas a CNN would have a number of trainable parameters dependent on the size of the applied kernel, but still much lower.

Tipically, CNN are interlaced sequences of three different types of layers: the previously described convolution layers, conventional fully-connected layers and Pooling Layers, which are used to aggregate multiple features into a single one, effectively down-sampling the feature maps according to different strategies. The CNN proposed in literature for reference tasks such as ImageNet classification, although varying in the number of layers, filters size and quantity, are composed essentially by the building blocks cited above.

1.2. Affective Computing and Emotion Recognition

Affective computing is a novel topic in Artificial Intelligence, defined for the first time in 2003 by Rosalind Picard [17] however, soon becoming one of the most trending mainstream multidisciplinary studies. Such an approach involving the collaboration of several disciplines, e.g., psychology, physiology, neurology, liberal studies, together with computer science and robotics, recently stressed out the importance of the extraction and recognition of affective mental states, e.g., emotions, moods, sentiments, and personality traits. Notwithstanding the interest on such new researches, still most of the available research focuses on pure theoretical models (e.g., in psychology and social sciences) and on product-based applications [18], mainly for marketing purposes. On the other hand, we prefer to focus on self-aid [19], health management, and communication in understanding and supporting humans in real-life problems, for any demanding task (e.g., due to disabilities [20], psychological states in emergencies, particular environments) with automated detectors and artificial assistants with machine emotional intelligence capabilities [21].

In our socially interconnected world, individuals already use and produce daily an overwhelming amount of heterogeneous data in a manageable and personalized subset of classified items. Data can be directly usable for emotion recognition (e.g., photos and text shared on social networks), or can contain elements indirectly inferrable for emotional intelligence (e.g., physiological data collected by wearables, including sleep patterns, continuous heart-rate, movement tracking, or emotion expressed by art [22] or experiencing exciting games [23]). If sentiment analysis relates only to recognizing the positiveness, negativeness or neutrality of sentiments, moods, and emotions, the process of emotion recognition is still less studied, implying the recognition of specific emotions of an emotional model. Since scientists do not agree on all the available models, research on ER in any applicative domain

starts from the widely recognized model of Ekman, which we chose for our study. Recent research underlines that Ekman's primary emotional states, including happiness, sadness, anger, disgust, or neutral state [24] can be recognized based on text [25], physiological clues such as heart rate, skin conductance, gestures, facial expression, and sound, that can be managed with a multidimensional approach [8].

Among all these approaches, facial recognition is still predominant. Under this point of view, we decided to focus on mouth as a particular facial element, almost always visible in any facial expression, anyhow considering that some cultures underestimate the mouth expressiveness than others. Some tribes in south America, for instance, as Ekman himself underlined in his first experiments on facial emotions labelling [6], rely more on the upper part of the face. The interesting issue is that, in general, the neutral emotion seems from such studies the most misunderstood both by humans, who tend to label it as anger or sadness. Our preliminary investigations, which are also confirmed by this final work, show that the same emotion is also the most missed by automatic mouth-based recognition, in the same classes of errors.

1.3. Artificial Intelligence assisting Health Care

Multidisciplinary studies at the basis of our work in Artificial Intelligence stressed out the importance of computer science for automatizing real-life tasks for assistive technologies [26,27]. Among them, labial detection and lip-reading [28] constitute our main background starting point [29].

One of the most promising advances of recent years for AI-assisted health care is the opportunity to develop Mobile Apps on widely-spread devices [4], e.g., smartphones, tablets, smartwatches, that can be easily exploited to be used in to support disabled users.

2. Problem description and proposed solution

Our study exploits the high precision of CNN processing mouth images to recognize emotional states. In this work, we test the technique on generic faces data sets, in order to find solutions to the following research questions:

- *With which precision it is possible to recognize facial emotions solely from the mouth?*
- *Is the proposed technique capable of recognizing emotions if trained on a generalized set of facial images?*

In a user-centred implementation, the software supports personalized emotional feedback for each particular user: personal traits, such as scars or flaws, or individual variations in emotional feeling and expression, help the training to precise recognition. The system can recognize different users because trained on a general data set including images varying in ethnicity, age and gender.

In order to obtain optimized results, the ambient light setting should not require a particular setup. A consistent implementation should meet the following requirements:

- **Robustness**: The algorithm must be able to operate even in the presence of low-quality data (e.g., low resolution, bad light conditions);
- **Scalability**: The user position should not be necessarily fixed in front of the camera, in order to avoid constraining the person. Therefore, the software should be able to recognize the user despite her/his position;
- **Luminosity**: luminosity is an important issue because the variation of light hardly influences the recognition capabilities of a CNN-based automated system. Available data sets usually provide photos shot under a precise lighting setup. On the contrary, a sufficient amount of training samples of each considered lighting (e.g., natural light, artificial bulbs, low-light conditions) should be provided for an efficient categorization.

3. The Framework Implementation

The proposed implementation has been developed using the *Python* programming language and the *Keras* framework [30]. The tests were carried out using the Google *Colab* platform, in which the Python code was developed and executed, exploiting the power fullness of the *Keras* libraries. The

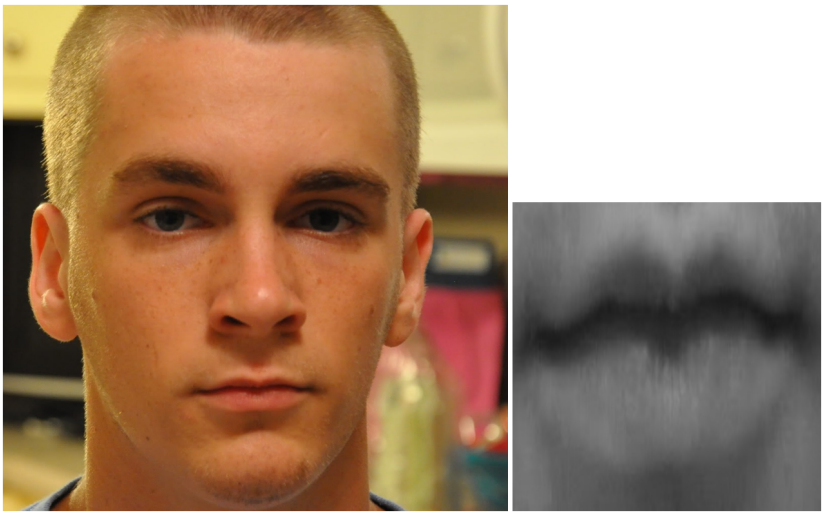


Figure 1. Mouth detection, cropping and resizing (source image from AffectNet database)

process of data analysis and processing has been developed with a chain of operations which output provides the input to the next step. The implemented procedure can be described with the following steps:

1. raw data set import: in the first phase the data set (which at the beginning is considered a raw data set) is composed of RGB images, and is imported into the system.
2. raw data set cleaning: all the faces contained in the photos of the data set are manually scanned and catalogued according to the correct emotion;
3. data set generation: from each image of the raw data set the portion representing the mouth of the subject has been automatically extracted;
4. training with data augmentation: at every step of the training, we read a portion of images from the data set. At each reading, we perform data augmentation [31] using random values of rotation degree within a given range. The model used for image analysis is created using *transfer learning*;
5. model generation: at the end of the training we save the structure and the weights of the neurons of the network that achieved the best performance in training;
6. evaluation of the model: to evaluate the quality of the model, we analyzed the percentage of recognition of emotions on the images of the validation set.

The mouth extraction from the images of the raw data set is carried out using a pre-trained neural network, the `shape_predictor_68_face_landmarks.dat` [12][32][2], which produces in output 68 landmarks, which are detected for each image. The shape predictor is pre-trained on the `ibug 300-W data set`. The landmarks are expressed as a series of coordinates identifying specific elements of a face, e.g., the position of mouth, eyes, cheekbones. Once we have obtained the landmarks of a face, we use those identifying the area of the mouth, cropping the image (i.e., cutting out the original image to obtain only the part related to the mouth). All images of the mouth are transformed so that they all have the same size of 299x299 pixels.

In Figure 1, we can see an example of mouth detection, cropping and resizing. On the left side, an image taken from the data set is shown. On the right, the corresponding extracted mouth is visible. On the completion of this process, the images of mouths are correctly stored in a directory structure associated with the considered emotion. We used data augmentation techniques [31] to increase the robustness of the neural network. Data augmentation randomly transforms the images, e.g., using a rotation with a value in the range [-4,+4] degrees, a flip or other transformation.

Peer-reviewed version available at Sensors 2020, 20, 5222; doi:10.3390/s20185222

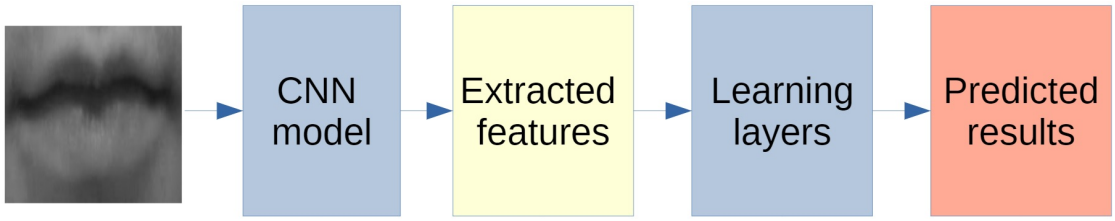


Figure 2. General scheme of our CNN network

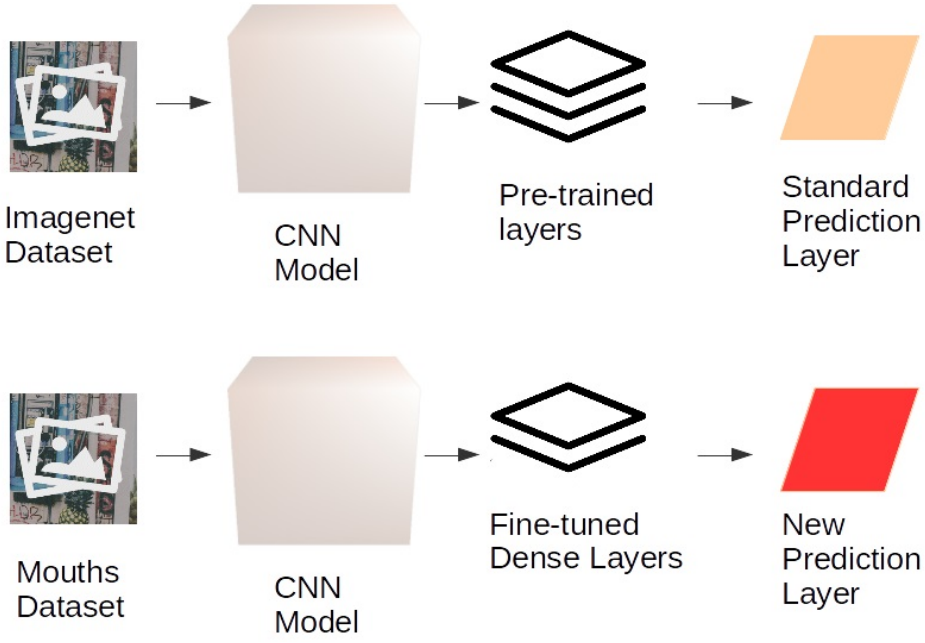


Figure 3. General scheme of the adapted transfer learning techniques on the considered CNNs

3.1. Neural Networks

The experiments have been performed on four Convolutional Neural Networks: VGG16 [33], InceptionResNetV2 [34], InceptionV3 [35] and Xception [36]. Among CNNs, these networks outperform AlexNet [13] on the widely used ImageNet [11] data set, which is one of the largest image data set used for the Transfer Learning pre-training phase.

All neural networks, whose general behaviour is described in Figure 3, have been tested using Transfer Learning as described in Figure 2, showing the process for a general CNN.

Adopting Transfer Learning, we used the pre-trained neural networks, which weights have been computed on the *ImageNet* [11] data set. The first layers of the networks have been frozen for the training phase, i.e. the convolutional layers weights have been fixed and have not been altered during the fine-tuning phase on the mouth emotion data set. This choice is due to the high capability of the CNN networks to recognize low-level features, e.g., points, lines, edges, corners, color distribution. We replaced the final layers for fine-tuning the network on the mouth emotions, using two dense layers with 64 neurons each, and a *SoftMax layer*. This final layer ranks the likelihood of the most appropriate class, thus returning the emotion classification. Only the final layers of the CNN networks, i.e. the *fully-connected* and the final *Softmax* level therefore changed, and can be re-trained in the fine-tuning phase. The model is setup saving the weights only when they have improved the emotion classification accuracy concerning the previous epoch, thus resulting in a final best neural network training configuration.

We tested two optimisers: Adam and SGD. Adam performed better with InceptionV3 and VGG16 with a learning rate equal to 0.001, with Xception with a learning rate of 0.01. SGD performed better

with InceptionResNetV2 with a learning rate equal to 0.001, momentum=0.9 and nesterov equal True. The remaining parameters used were the following: batch size equal to 25, maximum number of epochs equal to 100, however we used the early stopping technique: if there are not improved results, the training is stopped.

4. Image collection and Training phase

4.1. Data Set

The data set *AffectNet* provided by the University of Denver [37], USA. has been used for the emotional training (i.e., fine-tuning of the networks) and classification test. The raw data set is composed by the images of a large number of faces, and a spreadsheet file used to assign to each image an emotion label. We analyze a subset of the Ekman emotions, i.e., *Neutral*, *Happy*, *Surprise*, and *Anger*. A sample image for each class of emotion is shown in Figure 4. The data set has been cleaned analyzing images and removing duplicates. The automated mouth recognition removed also all the items where the mouth was not visible. If grainy or unsatisfactory resolution photographs can be removed, it is also important to maintain a broad difference in resolution, point of view, and light conditions. This variety helps the network to train on general images and not on only a particular light condition, resolution, and so on. The main filtering improvement of the data set has been exploited on the content of images, avoiding all the photographs showing fake emotions, i.e. facial expressions apparently simulated.

In Table 1 the number of images per type of emotion in the cleaned data set includes is shown.

	neutral	*happy*	*surprise*	*anger*
# of images	1239	562	463	478

Table 1. Number of images per type of emotion considered in our study.

5. Results and discussion

The experimental work has been divided in two phases. A preliminary phase included experiments carried out freezing the convolutional layers and training only the learning layers, as explained in section 3. The results suggested to discard the CNN networks with the lowest accuracy percentage, e.g., *VGG16*. On the contrary, the network immediately achieving the best performance is *Inception Resnet V2*. In the second phase the experiments focused on the networks achieving the best performances in phase one, and included the fine-tuning phase, i.e., removing the freezing of the convolutional layers. The best model obtained in the training phase has been saved for each network, rolling back the weights to keep the highest accuracy. Iterating this process, we obtained the final accuracy of 79.5% on the Validation Set of the best network, as shown in Table 2. As shown, the network that performed worse is *VGG16*, while the best network is *Inception Resnet V2*. In the table we included the last results for all the networks, where only the best network is fine-tuned in all its layers.

In Figure 5 the plot of the loss and accuracy of the model is shown as a function of the epochs (i.e., periods) of the simulation. The Confusion Matrix of the training set is reported in Figure 6. The Confusion Matrix of the validation set is reported in Figure 7.

A final conceptual remark can be reported about how the *Neutral* class has been misclassified. In Table 3 the absolute values of misclassified images is reported for each class on the *Neutral* evaluation. It is noticeable that the *Neutral* class, as introduced in subsection 1.2, is the most ambiguous class, in both the classification directions, i.e., *Neutral* is often misclassified, and the other classes are often misclassified as *Neutral*. Such behaviour is particularly relevant because Ekman pointed out in his experiments that humans tend to have the same bias. We can say that our evidence suggests that our

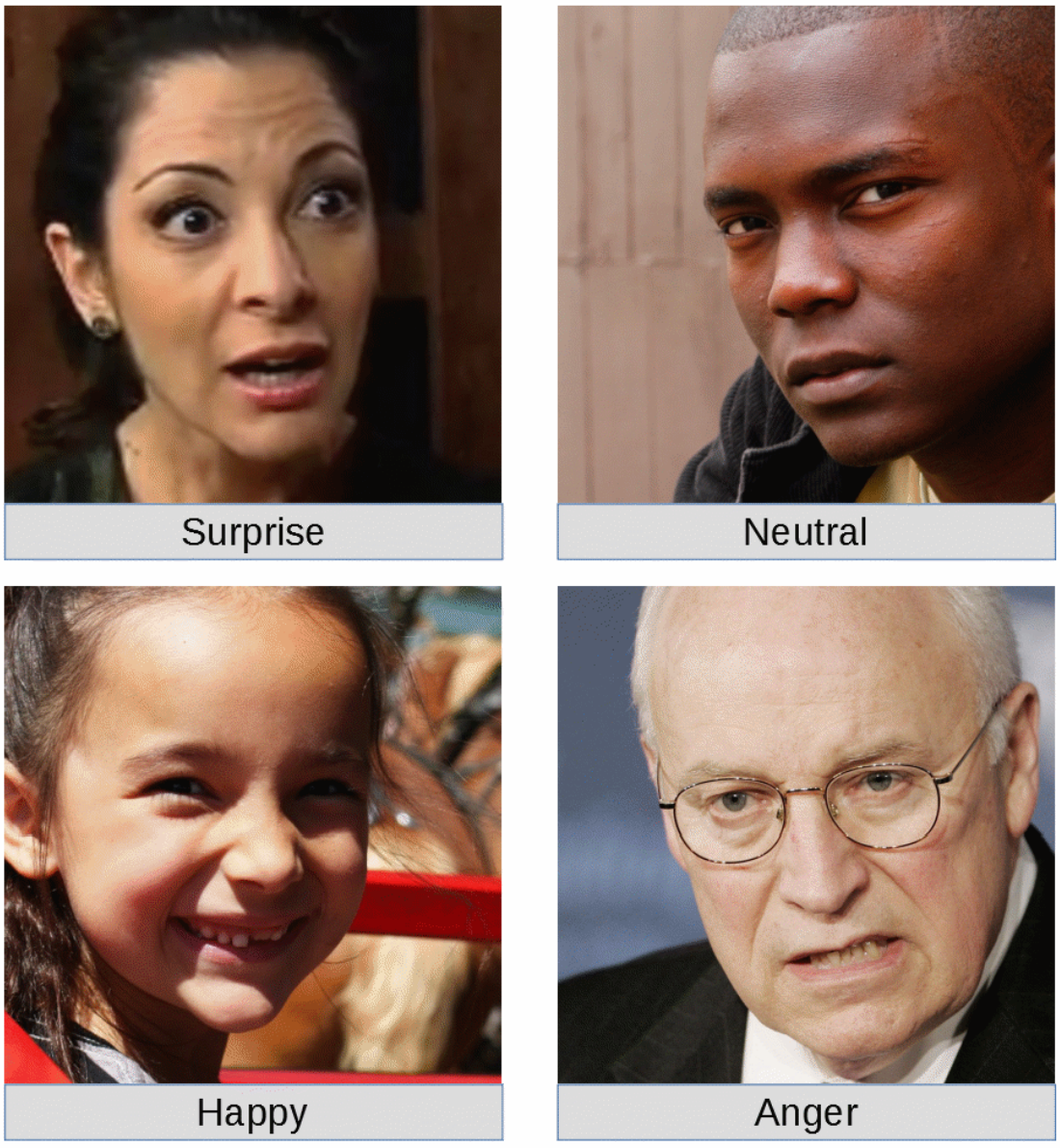


Figure 4. Sample images from the filter dataset

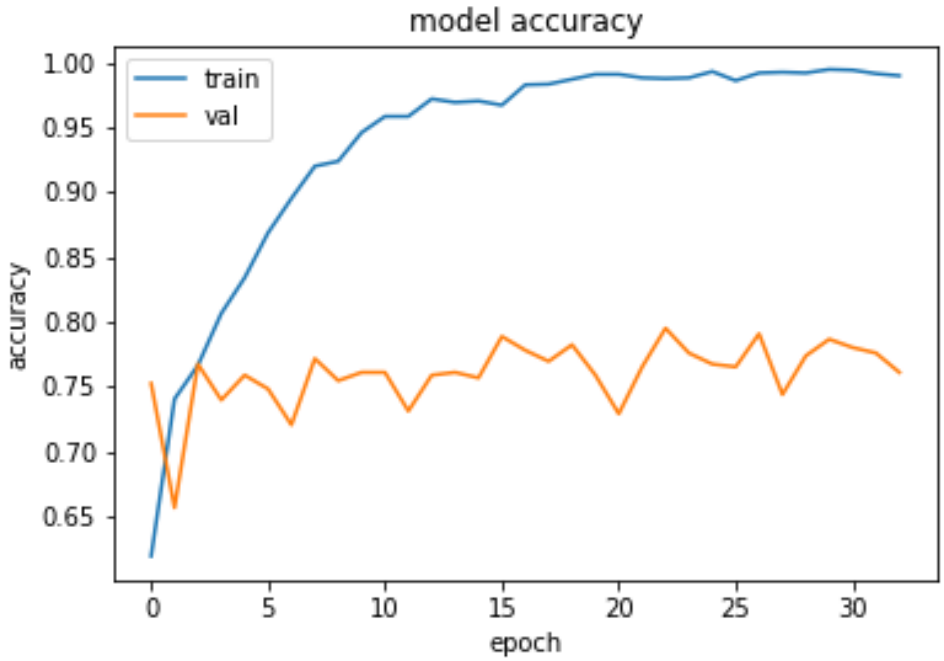


Figure 5. Training and validation accuracy of Inception Resnet V2

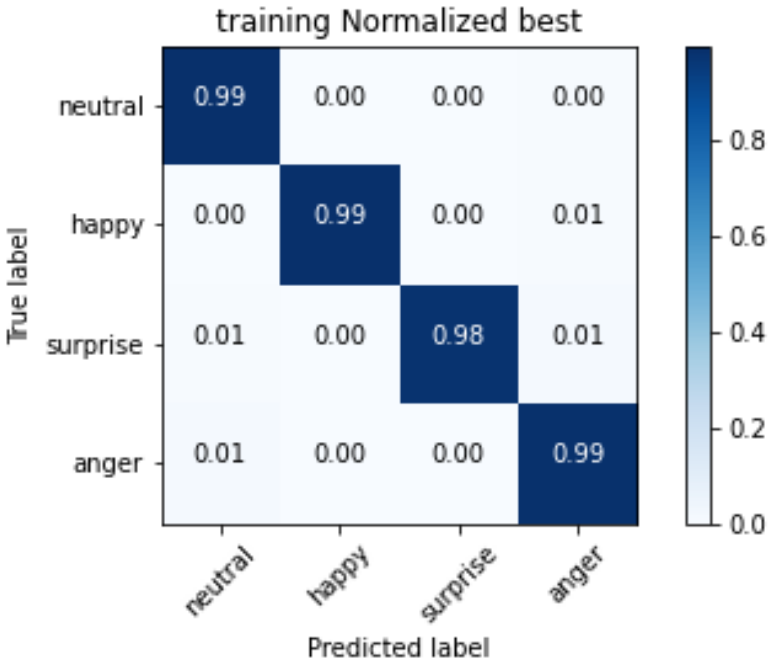


Figure 6. Confusion Matrix - Training Set - Inception Resnet V2

Network	*Accuracy*
Vgg-16	71.8 %
Inception Resnet v2	79.5 %
Inception V3	77.0 %
Xception	75.5 %

Table 2. Final results related to the considered CNNs

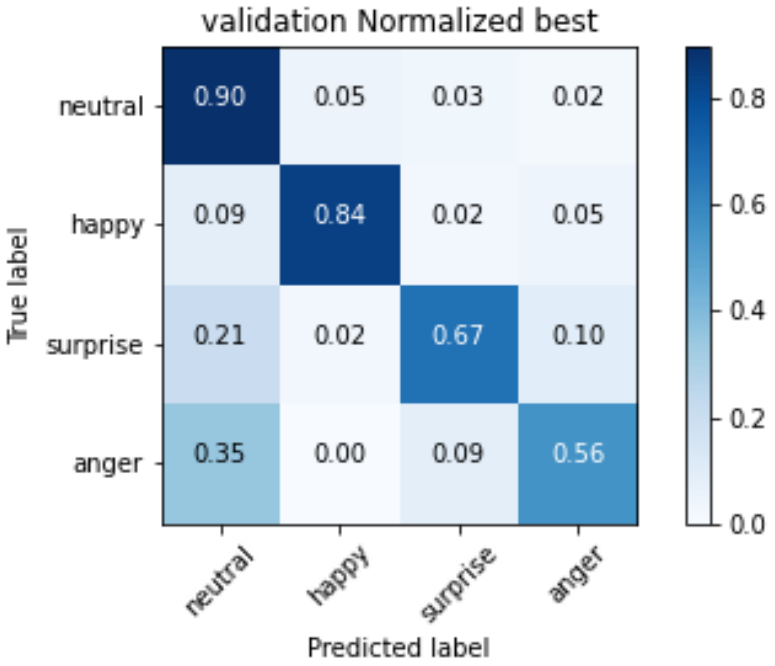


Figure 7. Confusion Matrix - Validation Set - Inception Resnet V2

networks behave similarly to the human brain, possibly learning the features associated to the human bias present in the considered data set.

	happy	*surprise*	*anger*
# of images	11	7	4

Table 3. Number of misclassified images of neutral faces.

Funding: "This research received no external funding"

Acknowledgments: We acknowledge Google Inc. for the usage of the Google Colab infrastructure for carrying out our calculations. We acknowledge the University of Denver, CO, USA, for having provided the access to the database of images we used for the present study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

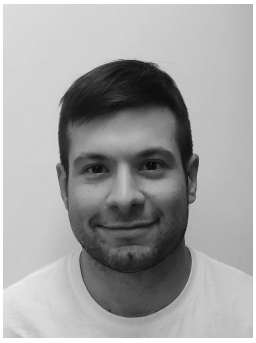
The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism
CNN	Convolutional Neural Network
AI	Artificial Intelligence
NN	Neural Network
GPU	Graphic Processing Unit
ER	Emotion Recognition
RGB	Red Green Blue

References

1. Gervasi, O.; Franzoni, V.; Riganelli, M.; Tasso, S. Automating facial emotion recognition. *Web Intelligence* **2019**. doi:10.3233/WEB-190397.
2. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. A Semi-automatic Methodology for Facial Landmark Annotation. 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 896–903.
3. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.
4. Riganelli, M.; Franzoni, V.; Gervasi, O.; Tasso, S. *EmEx, a Tool for Automated Emotive Face Recognition Using Convolutional Neural Networks*; Springer International Publishing: Cham, 2017; pp. 692–704.
5. Biondi, G.; Franzoni, V.; Gervasi, O.; Perri, D. An Approach for Improving Automatic Mouth Emotion Recognition. Computational Science and Its Applications – ICCSA 2019; Misra, S.; Gervasi, O.; Murgante, B.; Stankova, E.; Korkhov, V.; Torre, C.; Rocha, A.M.A.C.; Taniar, D.; Apduhan, B.O.; Tarantino, E., Eds.; Springer International Publishing: Cham, 2019; pp. 649–664.
6. Ekman, P. An Argument for Basic Emotions. *Cognition and Emotion* **1992**. doi:10.1080/02699939208411068.
7. Plutchik, R. A psychoevolutionary theory of emotions. *Social Science Information* **1982**, *21*, 529–553. doi:10.1177/053901882021004003.
8. Franzoni, V.; Milani, A.; Vallverdú, J. Emotional Affordances in Human-Machine Interactive Planning and Negotiation. Proceedings of the International Conference on Web Intelligence; Association for Computing Machinery: New York, NY, USA, 2017; WI '17, p. 924–930. doi:10.1145/3106426.3109421.
9. Franzoni, V.; Milani, A.; Vallverdú, J. Errors, Biases, and Overconfidence in Artificial Emotional Modeling. Proceedings of the International Conference on Web Intelligence; ACM: New York, NY, USA, 2019; WI '19.
10. Gervasi, O.; Magni, R.; Ferri, M. A method for predicting words by interpreting labial movements. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, Vol. 9787, pp. 450–464. doi:10.1007/978-3-319-42108-7_34.
11. Deng, J.; Dong, W.; Socher, R.; Li, L.; Kai Li.; Li Fei-Fei. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
12. Sagonas, C.; Antonakos, E.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 Faces In-The-Wild Challenge: database and results. *Image and Vision Computing* **2016**, *47*, 3–18. doi:https://doi.org/10.1016/j.imavis.2016.01.002.
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. doi:10.1145/3065386.
14. Perri, D.; Sylos Labini, P.; Gervasi, O.; Tasso, S.; Vella, F. Towards a Learning-Based Performance Modeling for Accelerating Deep Neural Networks. Computational Science and Its Applications – ICCSA 2019; Misra, S.; Gervasi, O.; Murgante, B.; Stankova, E.; Korkhov, V.; Torre, C.; Rocha, A.M.A.; Taniar, D.; Apduhan, B.O.; Tarantino, E., Eds.; Springer International Publishing: Cham, 2019; pp. 665–676.
15. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2013**, *35*, 1915–1929.
16. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press, 2016.

17. Picard, R.W. Affective Computing: Challenges. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 55–64. doi:10.1016/S1071-5819(03)00052-1.
18. Cieliebak, M.; Dürr, O.; Uzdilli, F. Potential and limitations of commercial sentiment detection tools. CEUR Workshop Proceedings, 2013, Vol. 1096, pp. 47–58.
19. Franzoni, V.; Milani, A. Emotion Recognition for Self-aid in Addiction Treatment, Psychotherapy, and Nonviolent Communication. Computational Science and Its Applications – ICCSA 2019; Misra, S.; Gervasi, O.; Murgante, B.; Stankova, E.; Korkhov, V.; Torre, C.; Rocha, A.M.A.C.; Taniar, D.; Apduhan, B.O.; Tarantino, E., Eds.; Springer International Publishing: Cham, 2019; pp. 391–404.
20. Hayes, G.R.; Hirano, S.; Marcu, G.; Monibi, M.; Nguyen, D.H.; Yeganyan, M. Interactive visual supports for children with autism. *Personal and Ubiquitous Computing* **2010**, *14*, 663–680. doi:10.1007/s00779-010-0294-8.
21. Picard, R.W.; Vyzas, E.; Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2001**, *23*, 1175–1191. doi:10.1109/34.954607.
22. Bertola, F.; Patti, V. Emotional responses to artworks in online collections. CEUR Workshop Proceedings, 2013, Vol. 997.
23. Canossa, A.; Badler, J.B.; El-Nasr, M.S.; Anderson, E. Eliciting Emotions in Design of Games - a Theory Driven Approach. EMPIRE@RecSys, 2016.
24. Angelov, P.; Gu, X.; Iglesias, J.A.; Ledezma, A.; Sanchis, A.; Sipele, O.; Ramezani, R. Cybernetics of the Mind: Learning Individual's Perceptions Autonomously. *IEEE Systems, Man, and Cybernetics Magazine* **2017**, *3*, 6–17.
25. Biondi, G.; Franzoni, V.; Li, Y.; Milani, A. Web-Based Similarity for Emotion Recognition in Web Objects. Proceedings of the 9th International Conference on Utility and Cloud Computing; Association for Computing Machinery: New York, NY, USA, 2016; UCC '16, p. 327–332. doi:10.1145/2996890.3007883.
26. Franzoni, V.; Milani, A.; Nardi, D.; Vallverdú, J. Emotional machines: The next revolution. *Web Intell.* **2019**, *17*, 1–7.
27. Gervasi, O.; Magni, R.; Macellari, S. A Brain Computer Interface for Enhancing the Communication of People with Severe Impairment. Computational Science and Its Applications – ICCSA 2014; Murgante, B.; Misra, S.; Rocha, A.M.A.C.; Torre, C.; Rocha, J.G.; Falcão, M.I.; Taniar, D.; Apduhan, B.O.; Gervasi, O., Eds.; Springer International Publishing: Cham, 2014; pp. 709–721.
28. Gervasi, O.; Magni, R.; Ferri, M. A method for predicting words by interpreting labial movements. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016, Vol. 9787, pp. 450–464. doi:10.1007/978-3-319-42108-7_34.
29. Bastianelli, E.; Nardi, D.; Aiello, L.C.; Giacomelli, F.; Manes, N. Speaky for robots: the development of vocal interfaces for robotic applications. *Applied Intelligence* **2016**, *44*, 43–66. doi:10.1007/s10489-015-0695-5.
30. Chollet, F.; others. Keras. https://github.com/fchollet/keras, 2015.
31. Antoniou, A.; Storkey, A.; Edwards, H. Data Augmentation Generative Adversarial Networks, 2017, [1711.04340].
32. Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. 2013 IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.
33. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014, [1409.1556].
34. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 2016, [1602.07261].
35. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision, 2015, [1512.00567].
36. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800–1807.
37. 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*, 18–31. doi:10.1109/taffc.2017.2740923.



Giulio Biondi Ph.D. candidate in Computer Science in the joint Ph.D. Program in Mathematics, Statistics and Computer Science between the Departments of Mathematics and Computer Science of the Universities of Florence and University of Perugia, where he also received his BSc degree and MSc summa cum laude. He has been cooperating since 2014 with Hong Kong Baptist University, where he also spent exchange research periods. He lectured in Computer Science and e-learning training courses and recently has been part of the University of Perugia covid-19 task force for e-learning. Since 2015he has been carrying out research on AI, focusing on Emotion Recognition, Link Prediction in Complex Networks, Natural Language Processing,and e-learning, publishing works and sitting as workshop chair at international conferences. He is IEEE and AIxIA member.

Valentina Franzoni Ph.D. in Engineering for Computer Science at the Department of Computer, Control, and Management Engineering at La Sapienza University of Rome, Italy, Postdoc Researcher at Department of Mathematics and Computer Science in the University of Perugia, Italy,where she received her BSc degree and MSc summa cum laude and she has been a senior research assistant (2011–2015). Also collaborating with the Department of Computer Science, Hong Kong Baptist University since 2012, where she also was senior research assistant in 2012.
At Perugia University she also worked as a Contract Professor for the Master in Systems and Technologies for Information and Communication Security, and Assistant Professor for Multimedia Systems and Computer Science. She was trainer of teachers for the e-learning and blended learning platform Unistudium, and she is part of the covid-19 task force for e-learning. Her research interests focus on AI, in which she has a record of invited international speeches, publications and awards,with a main focus on affective computing (multimodal emotional models; emotion recognition from short text, face recognition, mouth recognition,crowd sounds), complex networks (models for bacterial diffusion in social networks, link prediction in social networks, emotion diffusion in social networks, semantic proximity measures and contexts), web usability using VR navigation devices, and advanced strategies for e-learning.She has been chairing several workshops and organizing international conferences, she is guest editor in international journals, indexed and with impact factor, and she is a member of several scientific associations, e.g.AAS, IEEE, AAAC, WIE, AIxIA, Ithea.

Damiano Perri: Ph.D. candidate in Computer Science in the joint Ph.D. Program in Mathematics, Statistics and Computer Science between the Departments of Mathematics and Computer Science of the Universities of Florence and University of Perugia, where he also received his BSc degree and MSc summa cum laude. His research activity if focused on GPGPU Computing, Parallel and Distributed environments, Virtual and Augmented Reality, Artificial Intelligence, Neural Networks.
In particular his recent research activity is aimed at investigating predictive models in order to optimize the most important operators in emerging Convolutional Neural Networks (CNNs). He is an expert developer of synthetic environments and their applications in everyday life. Since 2016 he has been the technical manager of the LibreEOL project (https://www.libreeol.org/info).

Osvaldo Gervasi, Ph.D., Associate Professor at the Department of Mathematics and Computer Science, University of Perugia. He teaches *Networks Architecture*, *Virtual Reality Systems* at the Bachelor in Computer Science of Perugia University and *Distributed Systems and High-Performance Computing* and *Human Computer Interaction* at the Master in Computer Science "Intelligent and Mobile Computing" of the Perugia University. His research interests span from Parallel and Distributed Computing, Cloud and High Performance Computing, GPGPU Computing, Artificial Intelligence, Virtual and Augmented Reality, e-Learning.

He published more than 100 papers on scientific Journals and peer reviewed international conferences Proceedings. He edited 80 books of international conference proceedings. He was invited seven times to deliver a Plenary Lecture at International Conferences. He is editor in Chief of the *International Journal of Grid and Distributed Computing*, ISSN: 2005-4262 (Print), ISSN: 2207-6379 (Online), Published by NADIA (Sandy Bay, Tasmania 7005, Australia). He was President of the Open Source Competence Centre (CCOS) of the Umbria Region (Italy), established under the Regional Law 11/2006 from 2007 to 2013. He is President of the Not-for profit organization ICCSA, which manage yearly the International Conference on Computational Science and Its Applications (http://www.iccsa.org). He served for the period 2016-2020 the Board of Directors of The Document Foundation (TDF), which releases the popular Open Source Office Suite, LibreOffice. He is the Project Leader of the Open Source e-assessment platform LibreEOL (https://www.libreeol.org/info). He is Senior Member of IEEE and ACM and member of the Internet Society.