Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Intelligent Systems with Applications 15 (2022) 200094

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Emotion recognition in the times of COVID19: Coping with face masks

Roberto Magherini, Elisa Mussi, Michaela Servi *, Yary Volpe

Department of Industrial Engineering, University of Florence, Via Santa Marta 3, 50139, Florence, Italy

ARTICLE INFO

Keywords:
Artificial intelligence
COVID19
Emotion recognition
Grad-CAM
Facial Expression Recognition
Non-verbal communication

ABSTRACT

Emotion recognition through machine learning techniques is a widely investigated research field, however the recent obligation to wear a face mask, following the COVID19 health emergency, precludes the application of systems developed so far. Humans naturally communicate their emotions through the mouth; therefore, the intelligent systems developed to date for identifying emotions of a subject primarily rely on this area in addition to other anatomical features (eyes, forehead, etc..). However, if the subject is wearing a face mask this region is no longer visible. For this reason, the goal of this work is to develop a tool able to compensate for this shortfall. The proposed tool uses the AffectNet dataset which is composed of eight class of emotions. The iterative training strategy relies on well-known convolutional neural network architectures to identify five sub-classes of emotions: following a pre-processing phase the architecture is trained to perform the task on the eight-class dataset, which is then recategorized into five classes allowing to obtain 96.92% of accuracy on the testing set. This strategy is compared to the most frequently used learning strategies and finally integrated within a real time application that allows to detect faces within a frame, determine if the subjects are wearing a face mask and recognize for each one the current emotion.

1. Introduction

One cannot not communicate, is the first axiom of psychologist Watzlawich (Watzlawick et al., 2011) who emphasizes that in the interaction between human beings, any gesture (words, silence, activity) has a message value, and that communication plays a key role in the foundation of civil society. Communication can be divided into 7% verbal communication, 38% paraverbal communication (i.e., voice analysis) and 55% nonverbal communication (communication that occurs through facial expressions, gestures, glances, etc.) (Lapakko, 1997); as a result, nonverbal communication strongly affects many aspects of daily life and the effectiveness of human interactions. Specifically, expression interpretation can generate positive feelings of trust, agreement, but also negative feelings of self-preservation such as fear, suspiciousness, etc. (Ratliff and Patterson, 2008). Facial expressions (supported by the voice (Przybyło, 2008, Oosterhof and Todorov, 2009)) are therefore the first tool to communicate our feelings to the outside world, interact with others and through which we perceive how others communicate with us.

The recent COVID-19 pandemic situation and the consequent obligation to regularly wear a face mask imposed by many states, complicates and inhibits non-verbal communication, which thus suffers an important loss of information. The facial mask in fact hides the mouth region that holds a fundamental role in the recognition of emotions, particularly happiness (Blais et al., 2012). This makes emotional manifestations difficult to interpret and therefore undermines social ties as we know them (Carragher and Hancock, 2020, Olivera-La Rosa et al., 2020, Bernstein and Yovel, 2015). Some studies show that facial masks have had a strong negative impact on the recognition of emotions, preventing rapid impressions and synchronies, and creating social distances (Bani et al., 2021, Marini et al., 2021).

Another important consequence is the impact these protection measures may have on automatic face and emotion recognition performed by computer systems (Mao et al., 2015, Minaee and Abdolrashidi, 2019, Kaur and Kulkarni, 2021). In general, systems for automatic emotion detection, i.e. machines and devices sensitive to moods, are critical for rich and robust human-computer interaction.

In recent years, there has been a growing interest in integrating emotion recognition with modern human-computer interface technologies (Oosterhof and Todorov, 2009). This area of research has also found wide application in fields such as animation, medicine, and security (Palagi et al., 2020, Hess and Fischer, 2013, Tramacere and Ferrari, 2016, Singh et al., 2021, He, 2022). An application, that the authors consider of great interest for the healthcare field, is the robot-assisted

* Corresponding author at: University of Florence, Department of Industrial Engineering, Via di Santa Marta 3, 50139, Florence, Italy.
E-mail addresses: roberto.magherini@unifi.it (R. Magherini), elisa.mussi@unifi.it (E. Mussi), michaela.servi@unifi.it (M. Servi), yary.volpe@unifi.it (Y. Volpe).

https://doi.org/10.1016/j.iswa.2022.200094
Received 7 February 2022; Received in revised form 16 June 2022; Accepted 20 June 2022
Available online 26 June 2022

therapy for children with autism spectrum disorders (ASD) (Richardson et al., 2018) or for elderly patients with dementia (Osaka et al., 2021, Cruz-Sandoval et al., 2020).

Therapy programs that rely on humanoid and pet robots for activities such as companionship, as an exercise coach, and as a daily living assistant are now increasingly popular (Cruz-Sandoval et al., 2020, Szymona et al., 2021). Research is moving in this direction to make such robots increasingly empathetic and able to read patients' moods using machine learning approaches. These tasks have become more difficult with the introduction of face masks that occlude some parts of the face. In this direction, while face detection systems have been efficiently implemented even in the presence of face masks (Wang et al., 2020, Machiraju et al., 2021), to the best of the authors' knowledge, systems capable of automatically detecting emotions have not been developed.

The purpose of this work is to test the feasibility of implementing a Convolutional Neural Network (CNN)- based system capable of recognizing some important categories of emotions from the face when a face mask is worn. Specifically, a learning strategy is proposed to achieve a high level of accuracy in distinguishing emotions. The implementation of intelligent systems capable of automatically recognizing emotions in case of partial coverage of the face could be widely used in the present pandemic circumstances in which face mask are mandatory and in general in cases where part of the face is obscured for example by a scarf.

Section 2 provides information regarding the dataset used to train the network and the proposed training algorithm; Section 3 reports the results obtained in identifying emotions and the description of an application developed to perform the task real-time. In Section 4 the obtained results are discussed and compared to off-the-shelf methods. Finally, conclusions are drafted in Section 5.

2. Materials and Methods

2.1. Dataset

In order to develop the emotion recognition system, the authors chose to use the AffectNet dataset (Mollahosseini et al., 2017), as it is one of the most numerous facial expression datasets available in literature. The original dataset is composed of ~320K manually annotated images belonging to eight different discrete categories of affects: Happiness, Disgust, Anger, Fear, Surprise, Sadness, Contempt and Neutral. The so composed dataset is unbalanced in that the classes have significantly different frequencies of examples, e.g. the Happiness and Disgust classes have a ratio in terms of frequency of about ~30:1. The images have a resolution of 224 × 224. The dataset is composed of significantly different images: RGB and BW images are present, with different brightness and very heterogeneous backgrounds. Moreover, the framed subjects appear mainly frontal portrayed and sometimes face occlusion elements such as hands, hair, hats, sunglasses, etc. are present. Therefore, the dataset covers a wide range of real situations and is therefore sufficiently descriptive of environmental conditions in which the tool the authors intend to realize can be used.

In the literature, previous studies of affect recognition have simplified the task by combining the emotion categories (Mehendale, 2020, JA, 1994), similarly in this work a first manipulation of the original dataset consisted in recategorize the images into the following five categories: Anger-Disgust, Fear-Surprise, Happiness, Sadness and Neutral and in eliminating the Contempt category. The reclassification is justified by the similarity of some features of the upper region of the face in the facial expressions related to the feelings of respectively Anger-Disgust and Fear-Surprise. Specifically in the case of Anger and Disgust the eyebrows go down while in the case of Fear and Surprise the eyebrows go up together. The elimination of the Contempt category is justified in two ways: 1) it is not a key emotion in communication and 2) it is an emotion in which the expressiveness is concentrated in the mouth region and therefore not detectable if the subject is wearing a face mask. This merging process also brings a benefit in terms of partial balancing of the dataset.

Since the specific task envisaged in this work consists in identifying the emotions of a subject in the presence of a face mask, in the second step it was necessary to recreate a suitable dataset in which a synthetic mask was applied to each portrayed subject. This processing was achieved through the use of the MaskTheFace algorithm (GitHub 2021). Briefly, this algorithm initially identifies the tilt of the face and places a mask, chosen from a database of available masks; the orientation of the mask is subsequently refined by extracting six features from the face; more details are described in (Anwar and Raychowdhury, 2020).

In Fig. 1 two examples of images before and after the application of the aforementioned algorithm.

2.2. Proposed Framework

This work proposes an end-to-end deep learning framework which consists of two main blocks (see Fig. 2).

The first block is dedicated to checking the validity of the image since some images have occlusion elements, such as sunglasses, hats, particular shadows etc. that do not make the forehead and eye region visible, making it impossible to recognize the emotion in the presence of a face mask. Therefore, the first block has the purpose of filtering these images, considering valid those in which the eyes region is well visible, i.e. the subject is not wearing sunglasses and there are no important shadows given by hats or something else.

In order to realize this block the authors have tested some of the main neural networks architectures typically used for transfer learning. The main advantage of transfer learning is that it reduces the time spent to develop and train a model by reusing the weights of already developed models, therefore, where possible, there is a tendency to use this type of approach. In this work, the best performing network resulted to be the ResNet50v2 (He et al., 2016). ResNet50 is a reduced version of residual networks with 50 layers; the network attempts to solve vanishing gradient and degradation problems by introducing residual learning blocks with identity mapping connections.

After the transfer learning step, a fine-tuning step was performed to improve the accuracy of the model. This step involved unfreezing the entire model and retraining it on the available dataset with a very low learning rate. This phase is typically used as it can potentially lead to significant improvements by incrementally adapting the pre-trained features to the new data.

For the image validation task the images were resized to 64 × 64; this resolution value was determined empirically by testing different values, looking for a trade-off between lower training time and higher accuracy.

The second block of the framework performs emotion recognition. As mentioned, this work aims to recognize the following five categories of emotions: Anger-Disgust, Fear-Surprise, Happiness, Sadness and Neutral. For this purpose it was used the well-known architecture Inception, which, after a preliminary test campaign, has revealed to be the most suitable for the task. Inception (Szegedy et al., 2016) has played and plays a fundamental role in the world of machine learning.

Inception modules are used in convolutional neural networks to enable more efficient computation and deeper networks through dimensionality reduction with stacked 1 × 1 convolutions. These modules are designed to solve the problem of computational overhead, as well as overfitting, among other problems. The solution, in short, is to take multiple sizes of kernel filters within the CNN, and rather than stacking them sequentially, sort them to operate on the same level. The Inception module has undergone many modifications and improvements over time, in this work the InceptionResnetV2 architecture was employed, developed with the aim of introducing residual connections that sum the output of the convolution operation of the start module to the input.

For the network training phase, the authors propose a strategy that deviates from the typical learning process: rather than performing a single learning phase on the five categories dataset, two successive

Fig. 1. Two examples of the application of the MaskTheFace algorithm on two images of the AffectNet dataset.

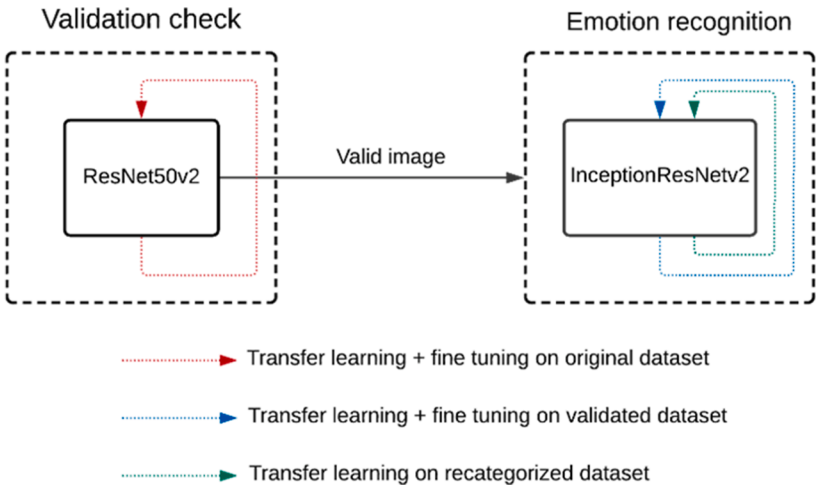

Fig. 2. Schematization of the proposed framework.

learning phases were performed, first on the original dataset (eight categories of emotions) and then on the recategorized dataset. Thus, this paper proposes an iterative strategy in which the training process is repeated twice.

In a first phase of training, a transfer learning step and a fine-tuning step were performed on the original dataset; the network thus trained

underwent a second training process that includes a new phase of transfer learning and fine tuning on the recategorized dataset. This strategy allows the network to approach the problem in a more generic way and then to specialize on the new dataset.

In the two phases of training, considering the unbalance of the dataset, the classes were weighted in such a way that the model pays more attention to the elements belonging to under-represented classes. To such purpose, for the calculation of the weight of every class the following formula was used:

$$W_i = \frac{\sum_{j=1}^{n} elems_j}{n * elems_i} \tag{1}$$

where W_i represents the weight assigned to class i, $elems_j$ represents the number of elements belonging to class j, and n represents the number of classes.

For both transfer learning and fine-tuning the Adam optimizer (Kingma and Ba, 2014) was used, as opposed to the often used Stochastic Gradient Descent (Reddi et al., 2019), as it allows to deal with more stable training variations which are due to the use of class-weights. In particular, for the fine-tuning phase the Adam optimizer in the amsgrad version was chosen, with a learning rate of 1e-04, beta_1 of 0.9, beta_2 of 0.999, and finally an epsilon value of 1e-07. Moreover, the categorical crossentropy loss was used.

For emotion recognition step the images were resized to 240 × 240, a higher resolution than the previous step to provide greater levels of detail, critical for the task. To reduce overfitting some data augmentation (such us flipping images and changing image brightness) were applied to the training dataset.

K-fold cross-validation and a Leave One Out (LOO) cross-validation were used to validate the method. More in detail, a k-fold cross validation (k = 5) was performed using the stratified k-fold version and a LOO validation was performed on a new dataset containing images of the five emotions relative to nine people outside of the AffectNet dataset.

3. Results

The tool presented in this paper was implemented in Python, using the Keras library (Ketkar and Ketkar, 2017), the high-level API of Tensor Flow (Abadi et al., 2016). A machine with Unix operating system, Nvidia 3060 graphics card and Intel Core i7 processor was used for training.

As described in the previous section, the first block consists of image validation, i.e., identifying images suitable for emotion recognition when the face mask is present. After the training process, the network for the validation check achieves an accuracy of 98.55% (on the testing set) in sorting valid images, i.e., with the eye region not occluded by sunglasses, shadows, or hats.

The images considered valid by the first module (264011 images) constitute the dataset for the second module (right block in Fig. 2), i.e., the validated dataset. These data were divided into training set (70%), validation set (10%) and testing set (20%).

As explained in the previous section, the model was first trained on the task of classifying eight emotions. Subsequently, as explained above it was performed a phase of training of the model on the recategorized data and an accuracy of about 93.53% on the training set, and 89.7% on the validation set was achieved after 10 learning epochs (Fig. 3). As said these results were obtained with transfer learning and fine-tuning phases. In fact, after some tests it was confirmed the necessity to use the transfer learning phase, since to train the model completely with the transfer learning less than 25 epochs were required; without this phase the learning times were not comparable.

Once the procedure of training was concluded, the final model was evaluated on the testing set according to the following metrics:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$F\_score = 2* \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

where TP = true positive, TN = true negative, FP = false positive, FN = false negative.

Table 1, a double entry table, shows for each row the true class and for each column the predicted class; the last two columns report the accuracy and F-score values obtained for each considered emotion.

3.1. K-fold cross-validation

To ensure the validity of the training method, k-fold cross-validation with k equals to 5 was used. In particular, to be able to perform a direct comparison between the results shown above and the k-fold results, the same test set was used. Therefore, the validation and training data were divided into 5 folders using the stratified version of k-fold cross-validation that tries to maintain the same percentage of elements for each class in the training and validation set. For each of the 5 folds 80%

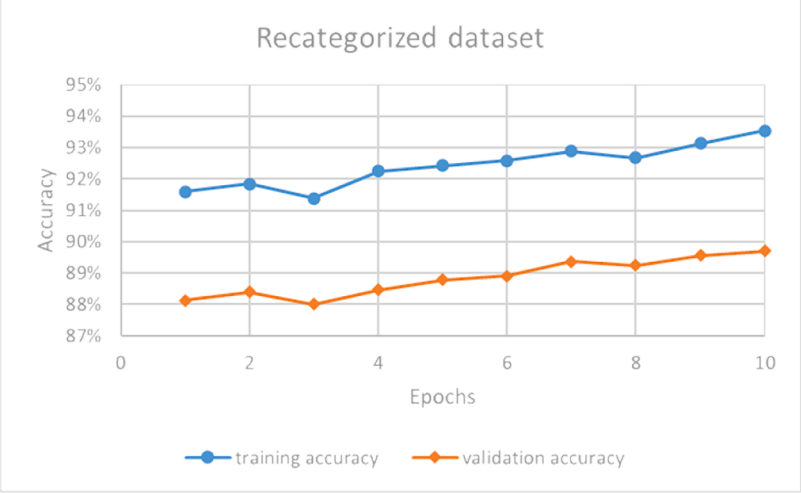

Fig. 3. Accuracy of the learning phase for the recategorized dataset.

Table 1
Accuracy and F-score of predicted classes on the testing set.

Class	A-D	F-S	H	N	S	TP	FP	TN	FN	Accuracy	F-Score
A-D	5222	45	30	70	36	5222	691	48773	181	99.34 %	96.65 %
F-S	26	3815	22	40	13	3815	706	50245	101	99.63 %	97.42 %
H	299	368	24196	1399	138	24196	259	28208	2204	91.96 %	91.65 %
N	322	244	188	13313	235	13313	1576	38989	989	96.39 %	93.08 %
S	44	49	19	67	4667	4667	422	49599	179	99.35 %	96.31 %

A-D = Anger-Disgust, F-S = Fear-Surprise, H = Happiness, S = Sadness, Acc = Accuracy

of data was used for the training phase and 20% for validation according to the procedure described before, i.e. first transfer learning then fine-tuning. In Fig. 4 are reported the categorical accuracies obtained using k-fold cross validation.

Finally in Table 2 are reported the final results, for the training, validation and test set of the 5 models.

Table 2
Accuracy values reached by each folder.

	K=1	K=2	K=3	K=4	K=5
Training	96.60%	96.62%	96.82%	96.87%	96.98%
Validation	96.64%	96.60%	96.71%	96.66%	96.70%
Testing	96.58%	96.32%	96.81%	96.86%	96.92%

3.2. Leave-One-Out Cross-Validation

To validate the generalizability of the solution, a Leave-One-Out cross-validation test was performed. This procedure is commonly used to estimate the performance of a model using data not included in the training phase. Specifically, to perform this test it was decided to collect new samples: images of all five emotions were collected from nine subjects wearing a face mask. The LOO procedure is as follows: one by one each subject was excluded from the training samples and the others were used for a small fine-tuning epoch of the best performing model (from Table 2 the best model is K=5), then the excluded subjects were used to test the performances of model, and finally the resulting model was validated again with the original testing set.

Table 3 shows the results of the LOO procedure: it can be seen that the only emotion misclassified by the model is Sadness and in general the model is able to correctly classify 40 out of 44 cases keeping the accuracy on the test set above 93%.

In Fig. 5 is reported an example of two subjects.

Table 3
Results of the LOO test. Wrong prediction is in reference to the excluded subject in the LOO iteration. Test accuracy is relative to the testing set described above.

Subject	Wrong Prediction	Test Accuracy
1	Sadness misclassified with Neutral	96.92%
2	Sadness misclassified with Neutral	96.92%
3	Sadness misclassified with AngerDisgust	96.92%
4	-	93.42%
5	-	94.63%
6	-	93.14%
7	Sadness misclassified with Neutral	96.92%
8	-	96.92%
9	-	96.92%

3.3. Gradient Map

In order to debug the model and visually detect what the network is "looking for" and "activating" within the image, the Grad-CAM algorithm was used (Selvaraju et al., 2016).

Grad-CAM works by finding the final convolutional layer in the network and then examining the gradient information flowing into that layer. The output of the algorithm is a heatmap visualization for a given class label that allows to visually verify where in the image the CNN is looking. Fig. 6 shows an example for each class of the application of the Grad-CAM algorithm.

3.4. Real-time Application

The good results obtained by the network pushed the authors to integrate the tool within a real time application, which using a webcam, aims to detect emotions of framed subjects. Specifically, the application identifies the subjects' faces in the scene, determines if the subjects are wearing a mask and detects five different classes of emotion.

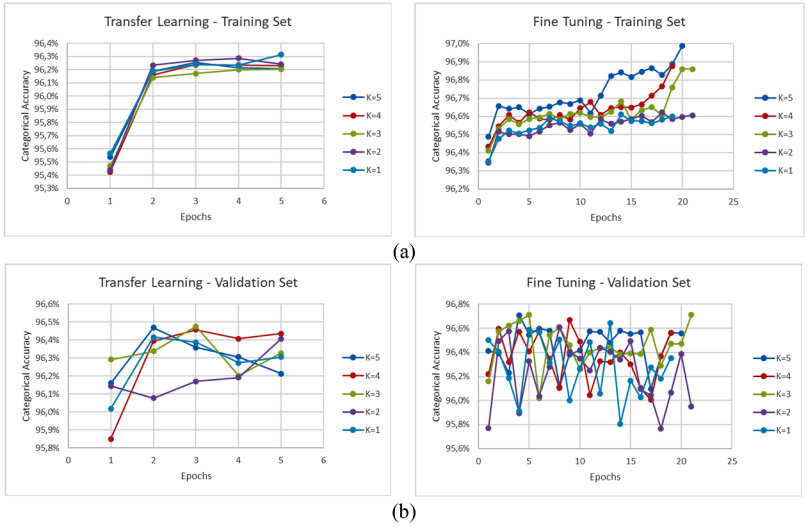

Fig. 4. (a) Transfer learning and (b) fine-tuning accuracy of training and validation sets for the 5 folders.

Fig. 5. Five emotions of two new subjects, from the left: a) Anger-Disgust, b) Fear-Surprise, c) Happiness, d) Neutral, e) Sadness.

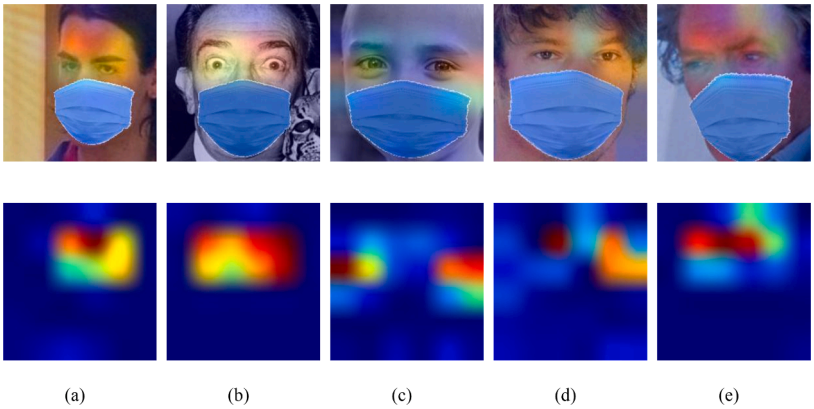

Fig. 6. Example of the results of the Grad-CAM algorithm on five images representing the five classes of emotion, from the left: a) Anger-Disgust, b) Fear-Surprise, c) Happiness, d) Neutral, e) Sadness.

The application was implemented using Python, with the support of a network developed with CaffeNet (Jia et al., 2014) for the detection of faces within the images, and Tensorflow was used for both the emotion classification network and for the validation network, which determines whether an input is valid on the basis of face occlusions. In Fig. 7 are shown examples of the application execution; specifically it is shown for each of the five categories of emotion the real-time response of the application, and also the ability of the application in detecting emotions on multiple faces in the same picture. In Fig. 8 the relative gradient map are shown.

4. Discussions

In Table 1 are reported the performances of the proposed model in terms of accuracy and F-score, before training with k-fold cross-validation; these metrics were chosen since in general the accuracy provides a simple and direct indication of the performance of the network, but it does not take into account the class imbalances, or the different costs of false negatives and false positives, which are considered in the F-score metric.

From the numerical results (Table 1) it is possible to see how the network is able to distinguish the three classes Anger-Disgust, Fear-Surprise, Sadness with high accuracy, respectively 99.34%, 99.63% and 99.35%. As regards, instead, the Happiness and Neutral classes, it is possible to notice how the network reaches the lowest values of Accuracy and F-Score. Specifically, the Happiness class reaches an accuracy value of 91.96%, which is the lowest, despite being the most frequently represented class. This phenomenon can be justified by what was mentioned in the introductory section, i.e. that the emotion of happiness generates a greater expressiveness in the region of the mouth. This also affects the accuracy of the Neutral class (96.39%) since the region analyzed by the network, that is the eyes, often has a connotation similar to that of Happiness.

The validation through k-fold cross-validation allowed to assess the validity of the proposed methodology. In fact the results obtained using k-fold cross-validation reached higher accuracy values. Better performances are due to a new and more balanced organization of the training and validation set, given by the use of the stratified version of k-fold cross-validation.

The LOO experiment, performed by collecting the expressions of nine new subjects with face masks, under different light and environmental conditions, showed the ability of the network to adapt to new inputs while maintaining stable performances. From Table 3 can be observed how, among the new cases, the only expression not always correctly classified is Sadness. The main reason for the misclassification of the Sadness class, which in the original dataset is classified with high accuracy, in the authors' opinion, lies in the fact that sadness is one of the most difficult emotions to pretend, and the nine subjects were not photographed in conditions of real sadness.

To visualize the connections activated when the emotion is identified with the proposed method, the authors used the Grad-CAM algorithm; in Fig. 8 it is possible to see for each example the relative heat map and which regions of the face are crucial for the discrimination of one emotion with respect to another. It is possible to observe that to identify the emotion of Anger-Disgust the contour of the eye is fundamental, while Sadness is recognized by analyzing the region of the forehead and

Fig. 7. Example of the real-time application integrating the emotion recognition network: a) Anger-Disgust, b) Fear-Surprise, c) Happiness, d) Neutral, e) Sadness, f) emotion recognition on multiple subjects.

Fig. 8. Grad-CAM algorithm applied to the results of Fig. 7, from the left: a) Anger-Disgust, b) Fear-Surprise, c) Happiness, d) Neutral, e) Sadness.

Happiness is detected by the eye's region. It can be observed that the active connections concern entirely the upper region of the face, demonstrating that the system has learned to identify the only anatomical region involved in the emotion expression.

In order to compare the proposed training strategy with the most common and classical learning methods used when facing a new task, in the following the results of two common training strategies implemented by the authors, using the same InceptionResnetV2 architecture, are shown. A first strategy (strategy I) involved training the model on the task of classifying the eight emotions originally composing the dataset: the first phase of transfer learning, equal to 30 epochs, and the following phase of fine-tuning, equal to 50 epochs, have seen the network reach an accuracy equal to 97.15% for what concerns the training set and 45.83% for what concerns the validation set.

For a second comparison strategy (strategy II), the InceptionResNetv2 architecture with a single training phase (including transfer learning and fine tuning) on the validated dataset already recategorized into the five classes was used. Therefore, unlike the proposed learning method, the training phase on the eight emotions was not performed, but rather a direct training on the dataset recategorized into five emotions. The training results are shown in Fig. 9.

As can be seen in Fig. 9, the accuracy on the validation set reaches a value of 70%, which appears significantly lower than the value obtained with the proposed method. The result presented in Fig. 9 is the best in terms of accuracy of a range of experiments performed varying the training set and the validation set with cross-validation.

The obtained results show a clear improvement in the ability of emotion recognition when the proposed method is applied. If the improvement between strategy I and the proposed method can be partly attributed to the higher simplicity of the task to be performed, between strategy II and the proposed method the improvement is due to the adopted learning strategy. In the proposed method the network learns initially to concentrate on important general features and then to focus on specific emotions and in particular on the differences between them. Table 4 summarizes the accuracy results obtained on the emotion recognition task, with the two described strategies and the proposed

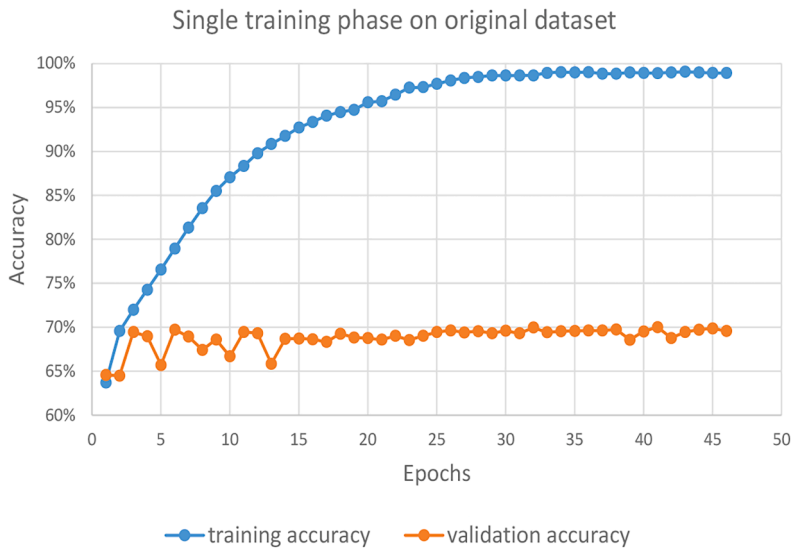

Fig. 9. Accuracy of training the InceptionResNetv2 directly on recategorized data.

Table 4
Comparison of the accuracy values reached by the three learning approaches

	Proposed method	Strategy I	Strategy II
	Training of model obtained firstly on eight-class dataset then on the five-class dataset	Training of InceptionResNetv2 on the eight-class dataset	Training of InceptionResNetv2 on the five-class dataset
Validation accuracy	96.71%	45.8%	70%

method, on the validation set.

5. Conclusions

Global social and health upheavals naturally affect the well-being of individuals. The COVID19 pandemic has made this phenomenon evident, affecting worldwide most areas of daily life. This work intends to address the problem of loss of communicative ability in social relationships related to the obligation to wear face masks, in order to limit the spread of the virus. Albeit the scientific community has been interested in the development of machine learning algorithms for emotion recognition, the new requirement to wear face masks compromises what has been developed so far. With this in mind, the authors propose a training method for deep neural networks to tackle this task. Specifically, starting from a dataset composed of eight emotions (Happiness, Disgust, Anger, Fear, Surprise, Sadness, Contempt, Neutral) an iterative learning strategy has been proposed in which a generic training phase is performed on the original dataset which is then recategorized into five classes (Anger-Disgust, Fear-Surprise, Happiness, Sadness, Neutral) to perform the second training phase. The proposed method proved to be a promising tool reaching an accuracy on the testing set of 96.92% in the recognition of the five emotions.

In light of these encouraging results, the authors developed a simple application able to perform the task real-time. The application first detects the faces in the scene, then determines the validity of the image (possible occlusions of the upper part of the face), observes and communicates the presence or absence of the facial mask and finally detects the emotion of the framed subject. The authors believe that the developed tool lays the foundation for the use of emotion detection in a variety of contexts, such as those in robot-assisted therapy for children with autism spectrum disorders (ASD) or elderly patients with dementia.

In conclusion, this study demonstrated the possibility of detecting a person's moods through machine learning techniques even when a face mask is worn. The task is achieved by analyzing features of the upper region of the face.

In situations where it is required to wear a face mask, such as in a hospital or laboratory, people detect the emotions of their interlocutors through the interpretation of the upper part of the face and are facilitated by the context. With this in mind, future development will involve the analysis of video sequences rather than single frames to allow the tool to also consider and learn the context. In addition, future developments will also include the analysis and study of the feasibility of employing Evolutionary Approaches (Bird et al., 2019) for feature selection and the improvement of the fine-tuning results.

Author's Contribution

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

CRediT authorship contribution statement

Roberto Magherini: Methodology, Software, Investigation, Data curation. **Elisa Mussi:** Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Michaela Servi:** Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing, Supervision. **Yary Volpe:** Conceptualization, Validation, Writing – review & editing, Supervision.

Declaration of Competing Interest

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

References

Abadi, M., et al. (2016). TensorFlow: A system for large-scale machine learning. In *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016* (pp. 265–283). May.

A. Anwar and A. Raychowdhury, "Masked Face Recognition for Secure Authentication," Aug. 2020.
Bani, M., et al. (2021). Behind the Mask: Emotion Recognition in Healthcare Students. *Medical Science Educator, 1*, 3. May.
Bernstein, M., & Yovel, G. (2015). Two neural pathways of face processing: A critical evaluation of current models. *Neurosci. Biobehav. Rev., 55*, 536–546. Aug.
Bird, J. J., et al. (2019). A Deep Evolutionary Approach to Bioinspired Classifier Optimisation for Brain-Machine Interaction. *Complexity*.
Blais, C., Roy, C., Fiset, D., Arguin, M., & Gosselin, F. (2012). The eyes are not the window to basic emotions. *Neuropsychologia, 50*(12), 2830–2838. Oct.
Carragher, D. J., & Hancock, P. J. B. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. *Cogn. Res. Princ. Implic., 5*(1). Dec.
Cruz-Sandoval, D., Morales-Tellez, A., Sandoval, E. B., & Favela, J. (2020). A Social Robot as Therapy Facilitator in Interventions to Deal with Dementia-related Behavioral Symptoms. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 161–169). Mar.
"GitHub - aqeelanwar/MaskTheFace: Convert face dataset to masked dataset." [Online]. Available: https://github.com/aqeelanwar/MaskTheFace. [Accessed: 21-Sep- 2021].
He, J. (2022). Algorithm Composition and Emotion Recognition Based on Machine Learning. *Computational Intelligence and Neuroscience*. Jun.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mappings in Deep Residual Networks. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 9908*, 630–645. LNCSMar.
Hess, U., & Fischer, A. (2013). Emotional Mimicry as Social Regulation. *Personality and social psychology review, 17*(2), 142–157. May.
JA, R. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol. Bull., 115*(1), 102–141. Jan.
Jia, Y., et al. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. In *MM 2014 - Proc. 2014 ACM Conf. Multimed* (pp. 675–678). Jun.
Kaur, S., & Kulkarni, N. (2021). Emotion recognition-a review. *International Journal of Applied Engineering Research, 16*(2), 103–110.
Ketkar, N., & Ketkar, N. (2017). Introduction to Keras. *Deep Learning with Python* (pp. 97–111). Apress.
Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc*. Dec.
Lapakko, D. (1997). Three cheers for language: A closer examination of a widely cited study of nonverbal communication. *Communication Education, 46*(1), 63–67. Jan.
Machiraju, S., Urolagin, S., Mishra, R. K., & Sharma, V. (2021). Face Mask Detection using Keras, Opencv and Tensorflow by Implementing Mobilenetv2. In *In2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 1485–1489). Dec.
Mao, Q. R., Pan, X. Y., Zhan, Y. Z., & Shen, X. J. (2015). Using Kinect for real-time emotion recognition via facial expressions. *Front. Inf. Technol. Electron. Eng., 16*(4), 272–282. Apr.
Marini, M., Ansani, A., Paglieri, F., Caruana, F., & Viola, M. (2021). The impact of facemasks on emotion recognition, trust attribution and re-identification. *Sci. Reports 2021 111, 11*(1), 1–14. Mar.
Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Appl. Sci. 2020 23, 2*(3), 1–8. Feb.
Minaee, S., & Abdolrashidi, A. (2019). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors, 21*(9). Feb.
Mollahosseini, A., Member, S., Hasani, B., Mahoor, M. H., & Member, S. (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. Aug.
Olivera-La Rosa, A., Chuquichambi, E. G., & Ingram, G. P. D. (2020). Keep your (social) distance: Pathogen concerns and social perception in the time of COVID-19. *Pers. Individ. Dif., 166*. Nov.
Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion, 9*, 128–133. Feb.
Osaka, K., Tanioka, R., Betriana, F., Tanioka, T., Kai, Y., & Locsin, R. C. (2021). Robot Therapy Program for Patients with Dementia: Its Framework and Effectiveness. *Information Systems-Intelligent Information Processing Systems*. Feb.
Palagi, E., Celeghin, A., Tamietto, M., Winkielman, P., & Norscia, I. (2020). The neuroethology of spontaneous mimicry and emotional contagion in human and non-human animals. *Neurosci. Biobehav. Rev., 111*, 149–165. Apr.
Przybyło, J. (2008). *Automatyczne rozpoznawanie elementów mimiki w obrazie twarzy i analiza ich przydatności do sterowania*. St. Staszica, Kraków: Akademia Górniczo-Hutniczna im.
Ratliff, M. S., & Patterson, E. (2008). Emotion recognition using facial expressions with active appearance models. In *Proc. of HRI. Citeseer*. Mar.
Reddi, S. J., Kale, S., & Kumar, S. (2019). On the Convergence of Adam and Beyond. In *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc*. Apr.
Richardson, K., et al. (2018). Robot enhanced therapy for children with autism (DREAM): A social model of autism. *IEEE Technol. Soc. Mag., 37*(1), 30–39. Mar.
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis., 128*(2), 336–359. Oct.
Singh, P., Mishra, R. K., Urolagin, S., & Sharma, V. (2021). Enhancing Security by Identifying Facial Check-in using Deep Convolutional Neural Network. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 1006–1010). Dec.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2818–2826). Dec.
Szymona, B., et al. (2021). Robot-Assisted Autism Therapy (RAAT). Criteria and Types of Experiments Using Anthropomorphic and Zoomorphic Robots. Review of the Research. *Sensors 2021, Vol. 21, Page 3720, 21*(11), 3720. May.
Tramacere, A., & Ferrari, P. F. (2016). Faces in the mirror, from the neuroscience of mimicry to the emergence of mentalizing. *Journal of Anthropological Sciences, 94*, 113–126. Jun.
Z. Wang et al., "Masked Face Recognition Dataset and Application, " Mar. 2020.
Watzlawick, P., Bavelas, J. B., & Jackson, D. D. (2011). *Pragmatics of human communication : a study of interactional patterns, pathologies, and paradoxes* (p. 284). New York, London: W.W. Norton & Company.