An Approach for Improving Automatic Mouth Emotion Recognition

An Approach for Improving Automatic Mouth Emotion Recognition

Giulio Biondi[1,2] (ORCID:0000−0002−1854−2196), Valentina Franzoni[2,3,4] (ORCID:0000−0002−2972−7188), Osvaldo Gervasi[2] (ORCID:0000−0003−4327−520X), and Damiano Perri[2] (ORCID:0000−0001−6815−6659)

[1] University of Florence, Dept. of Mathematics and Computer Science, Florence, Italy
[2] University of Perugia, Dept. of Mathematics and Computer Science, Perugia, Italy
[3] Sapienza University of Rome, Department of Computer, Control, and Management Engineering "Antonio Ruberti", Rome, Italy
[4] *corresponding author*

Abstract. The study proposes and tests a technique for automated emotion recognition through mouth detection via Convolutional Neural Networks (CNN), meant to be applied for supporting people with health disorders with communication skills issues (e.g. muscle wasting, stroke, autism, or, more simply, pain) in order to recognize emotions and generate real-time feedback, or data feeding supporting systems. The software system starts the computation identifying if a face is present on the acquired image, then it looks for the mouth location and extracts the corresponding features. Both tasks are carried out using Haar Feature-based Classifiers, which guarantee fast execution and promising performance. If our previous works focused on visual micro-expressions for personalized training on a single user, this strategy aims to train the system also on generalized faces data sets.

1 Introduction

In this work, we present a system for mouth-based visual emotion recognition. Our purpose is to lay the basis for a health-care system for people who suffer from severe disease, e.g., strokes, or conditions such as autism, who may benefit from automated support of emotion recognition. Such systems can detect basic emotions from smartphone or computer camera devices, to produce feedback, either text, audio or visual for other humans, or a digital output to support other connected services. Connecting such an architecture to appropriate services could help users to convey their or others' emotions more effectively, providing augmented emotional stimuli, e.g., in case of users affected from a pathology which involve social relationship abilities, or when users experiment difficulties in recognizing emotions expressed by others. The system could also call a human assistant, e.g., for hospitalized patients feeling intense pain. In this paper, we focus on the mouth expression in correctly determining the emotion

expressed by a subject. A crucial step is the selection of a reference model which classifies emotions, e.g., Ekman,[54] Plutchik, and Lovheim [15]. We selected a basic subset of the Ekmann emotions: *Joy* and *Disgust*, together with the *Neutral* condition (i.e., no emotion expressed). Joy is among the simplest emotions to recognize through face expression, thus an ideal candidate for results comparisons concerning the state-of-the-art. Disgust, instead, is present in much fewer instances in available data sets, because it is more difficult to stimulate, and it is a less ideal but more interesting example of computation. We include the neutral state as a control state for recognition results on both emotions.

2 Problem description and proposed solution

Our study exploits the high precision of CNN processing to process mouth images to recognize emotional states. On one hand, we expect the system's capability to exploit best on the single user with personalized training; on the other hand, in this work we also 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-centered implementation, the user trains the network on her/his facial expressions and 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. Then, we train the software also to recognize different users. In order to obtain optimized results, the ambient light setting needs a proper setup:

- **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**: an important problem is precisely that of the variation of light. In computer vision ([1]), the variation of the lens involves an alteration of the information ([41]. No complete control of the detected information is achieved: the system will be able to withstand variations in brightness without compromising the original information.

The proposed solution has been implemented in C++ and OpenCV graphics libraries; hence, it is compatible with all operating systems, with high reliability and constant support from the community.

3 State of the Art

Deep Learning and Image Classification

The recent scientific focus on Deep Learning towards the end of the XX century has contributed to the rebirth of significant interest in neural networks. The real impact of Deep Learning began in the context of speech recognition around the year 2010, when two Microsoft Research employees, Lil Deng and Geoenix Hinton, realized that using large amounts of data for training a deep neural network resulted in lowering error rates far below the state of the art [20]. Discoveries in the field of hardware have certainly contributed to the rise of interest in Deep Learning. In particular, the ever-powerful GPUs seem to be able to perform the countless mathematical calculations of matrices and vectors in Deep Learning [11,12,13]. Actual GPUs allow reducing workout times from the weeks to a day. Recently, deep learning has been used for several types of research aiming at the classification of images and learning, trying to solve the limitations of machine learning, which reside in overfitting and domain dependence, with image adaptation, kernel randomization [14] and transfer learning [21]. Commitment has been dedicated by researchers to exploit domain dependence as a feature, where personalized classification can quickly exploit a particular user or entity, especially for smart-home systems [35] and microblog sentiment tagging [37]. Alternative approaches consider evolutionary algorithms,[27][58][59] random walks on semantic networks of images [25][26][60] and max-product neural networks.

History and Description of Neural Networks

Convolutional Neural Networks are among the most used methods for affective image classification [22][2] thanks to their flexibility for transfer learning, and easy tools available on the Web [34]. An artificial Neural Network (NN), composed of artificial neurons 1, or nodes, can be used for solving artificial intelligence (AI) problems. NNs are biologically inspired, where a neural network is a network or circuit of neurons in the brain. The connections of the biological neuron are modeled as weights: a positive weight reflects an excitatory connection, while negative values mean inhibitory connections. In 1983 Geoff Hinton, now an emeritus professor at University of Toronto, co-invented Boltzmann machines, [45] one of the first types of neural networks to use statistical probabilities, then updating the strength of the connections within a neural network with backpropagation. [46] In the late-1970s and early-1980s, Hinton began working with neural networks when they were deeply unfashionable, because most computer scientists believed the technique was a dead end, while a better approach to Artificial Intelligence (AI) could be to explicitly encode human expertise in rules sets. Today we know that deep neural networks using backpropagation underpin most advances in AI, from Facebook friends automatic tagging, to the voice recognition capabilities of Amazon Alexa and Google Home, to its translation capability from previously difficult languages, such as Mandarin. LeCun, then,

was a post-doc with Hinton's supervision, developing Convolutional Neural Networks as an improvement of the work on backpropagation. Bengio, who worked with LeCun on computer vision at Bell Labs, applied neural networks to natural language processing, leading to enormous advances in computer translation. Recently, he also built a model to allow neural networks to create novel and realistic images. In March 2019, Hinton, LeCun and Bengio received together the Turing Award, considered the Nobel prize for computing, for their advances in Artificial Intelligence with Deep Learning. [47] As we can see in Fig. 1, a single perceptron (i.e., the NN atom) takes several binary inputs, $x_1, x_2, \ldots, x_n$ and produces a single binary output. Weights $w_1, w_2, \ldots, w_n$ can be introduced

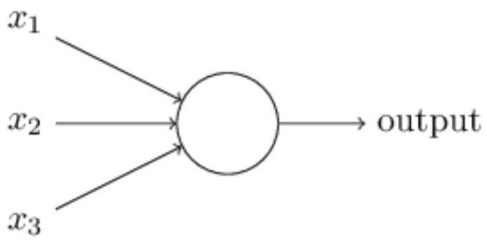

Fig. 1. A simple example of a 3-input perceptron

to express the importance of the respective inputs to the output. The neuron's output, 0 or 1, is determined whether the weighted sum $\sum_j w_j x_j$ is less/greater than a threshold parameter value of the network:

$$output = \begin{cases} 0 & if \sum_j w_j x_j \\ 1 & if \sum_j w_j x_j \end{cases} \tag{1}$$

By varying the weights and threshold, we can get different models of decision-making, thus different devices device capable to make decisions by weighting up evidence. A real NN will have several perceptrons in each column, and several cascade columns, where each columns is called a *layer*. A several-layers NN of perceptrons can engage sophisticated decision-making, adding variations to the comparison to the threshold. Several types of layers can be adapted to different calculation aims.

Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a class of deep neural networks, most commonly applied to analyzing visual content, with excellent results on image recognition, segmentation, detection and retrieval.[10,44] The key enabling factors behind such relevant results were principally techniques to scale up the networks to millions of parameters, where labeled data sets are needed to support

the learning process. CNNs are able, under such conditions, to learn powerful and interpretative image features. Convolutional layers apply an operation of convolution to the input, which emulates the response of an individual neuron to visual stimuli, processes data only for its receptive field. A set of kernels (i.e., learning parameters), with a small receptive field, extend through the full depth of the input volume. A forward pass convolutes each filter across width and height of the volume in input, calculating the dot product between the filter entries and the input, thus producing a 2-dimensional activation map. The network results to learn filters activating when some specific type of feature is detected in a particular position of the input. Although fully connected neural networks can be used to learn features as well as classify data, a relevant amount of neurons is necessary due to the large input sizes of images, where compression is not always a good idea because any pixel may be relevant. E.g., a fully connected layer for an image of size 100 x 100 will have 10000 weights for each neuron. The operation of convolution offers a great solution to the problem, so tat the network can be deeper with fewer parameters. E.g., tiling regions of size 5 x 5, regardless of image size, having the same shared weights, require just 25 kernels. The problem of exploding gradients in traditional NNs with many layers is solved using backpropagation. Convolutional networks may also include local or global pooling layers, to reduce data dimension using a combination of the outputs of neuron clusters obtaining one neuron in the following layer.

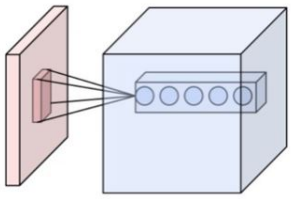

Fig. 2. A single CNN layer

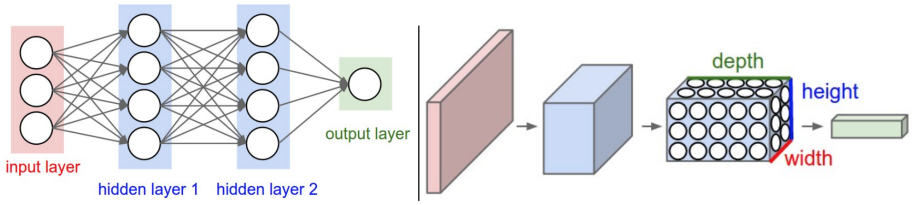

Fig. 3. Left: A regular 3-layer Neural Network. Right: A Convolutional Network arranges its neurons in three dimensions (width, height, depth). The input layer holds the image.

The neurons in a layer will be connected to a small region of the previous layer, as illustrated in Fig. 2, instead of all of the neurons, as happens in the fully-connected layer, which connects all the neurons in one layer to every neuron in another layer. A simple CNN is a sequence of layers, each of which transforms one volume of activations to another through a differentiable function. Typically, three types of layers are used: Convolutional Layer, Pooling Layer, and Fully-Connected Layer, which are then stacked together to form a full CNN architecture (see Fig. 3). In this way, CNN transforms the original image layer by layer from the original pixel values to the final class scores. Note that some layers contain parameters, and others do not. [43] In particular, the convolutional and the fully-connected layers perform transformations as a function of both the activations in the input volume, both the the weights and biases of the neurons (i.e. the parameters). On the other hand, the pooling and $RELU$ layers, which apply an element-wise activation function, such as the $\max(0, x)$, will implement a fixed function.

Artificial Intelligence assisting Health Care

For computerized health care assisting, multidisciplinary studies in Artificial Intelligence, Augmented Reality and Robotics stressed out the importance of computer science for automatizing real-life tasks for assistive and learning objects, [56] such as detecting words from labial movements (i.e. automated lip detection) [29], Virtual reality for prosthetic training [24] or neural telerehabilitation of patients with stroke [6], vocal interfaces for robotics applications [30]. As an application of complex networks, it is possible to predict bacteria diffusion patterns, [36][61] as well as epidemiology data [23], having a viral spread. To be mentioned, huge advances are happening on medical image recognition and multi-stage feature selection for classification of cancer data [32], and of text corpora for medical or patient feedback in social networks. One of the most promising advances of recent years for AI-assisted health care is the opportunity to have light-implementation Mobile Apps, that can be quickly developed to be used in a friendly manner [5], to assist and support disabled users for communication and learning tasks. Such applications can be run directly on personal smartphones or wearable devices, for health monitoring and prognosis [8] as well as for interactive support for people with disabilities or conditions that can influence communication and learning, such as autism spectrum disorders [7]. Using cloud services or networks in the Internet of Things (IoT), makes possible both to connect such devices to high capability servers, both to collect data in a distributed collaborative perspective [19], in order to feed big knowledge-bases, and increase the capability of the single object, i.e. of its owner, as a member of a vast interactive collective dynamic knowledge (i.e. a Big Data) network.

Affective Computing and Emotion Recognition

Multidisciplinary approaches recently stressed out the importance of recognizing and extracting affective and mental states, in particular emotions, for commu-

nication, understanding, and supporting humans in any task with automated detectors and artificial assistants having machine emotional intelligence [3]. In real-life problems, individuals transform overwhelming amounts of heterogeneous data in a manageable and personalized subset of classified items. The process of recognition of moods and sentiments is mostly complex. Recent research underlines that primary emotional states such as happiness, sadness, anger, disgust, or neutral state [38] can be recognized based on text,[31] physiological clues such as heart rate, skin conductance and face expression, differently from sentiment, moods and affect, which are more complex states and can be better managed with a multidimensional approach [39] [15]. Since Rosalind Picard defined the challenges for Affective Computing in 2003 [4], numerous advances have been made in the task of emotion recognition, such as defining collective influence of emotions expressed online [9], stating that emotional expressiveness is the crucial fuel that sustains communities; studying cultural aspects of emotions in art [16] and its variations; create emotionally engaging experiences in games [33], where affective changes are crucial to the conscious experience of the world around us. Some of the more ethical and critical challenges defined by Rosalind Picard, however, remain open. For example, many of the modalities for emotion recognition (e.g., blood chemistry, brain activity, neurotransmitters) are not readily available, commercial tools are limited [18], data sets for training are not general [17] and people's emotion is so idiosyncratic and variable, that there is difficult to recognize an individual's emotional states from available data [4]. Moreover, the challenge to use Affective Computing to help people, e.g., with self-aid tools is not widely faced in research, preferring applications to marketing. [55]

4 The proposed Emotion Recognition engine

We based our Emotion Recognition Engine on popular open-source libraries: the image processing features are provided by OpenCV [48], version 3.1.0. First, the software recognizes the presence of a face in the image; when a face is found, the algorithm looks for the mouth location and extracts the corresponding subframe. [28] Both tasks are carried out using Haar Feature-based Classifiers [49], which guarantee fast execution and promising performance. A face detection pre-trained classifier is integrated into OpenCV; the mouth classifier, instead, is the one used in [50]. During the training phase, samples of the subject are captured from the camera at regular intervals (or fed from disk) and used to produce a set of mouth snapshots. Such snapshots will form, after shuffling, the training, validation, and test set, with the first two used to train the networks with the Deep Learning framework Caffe [34]. The remaining images are subsequently used to test the performances of the networks. The system has been designed to perform both offline and online recognition, i.e. recognize emotions from a series of pre-stored images or directly from a video feed.

4.1 The structure of the EmEx2 CNN

In order to have a direct approach to the world of conundrum neural networks, the EmEx [42] approach, which we used in part of our tests, focuses on detecting user-centered emotions from the mouth. Network layers are set up to extract the specific information of the image data, accurately setting the parameters to have a valid recognition. In our CNN, training data are labeled with emotions, and the results of the layer computation are evaluated in terms of accuracy and loss. The neural network architecture is based on the popular *LeNet-5*,[62], consisting of several layers connected. The main level is for the *data set* and the corresponding tags. The second layer is the *convoluted layer*, where the convolution operations are performed on the input images, extracting features about each frame in each class. Then, a *pooling layer* is used to reduce the parameter magnitude, reducing in width and height, the volume of previously created data, with a time gain in computation. For scaling, a max function of a variable set is used. Different convolutional and pooling layers follow. An *inner product layer innerProduct* groups the information in a single numeric value, to be processed again in the following phases. The system is now capable of returning a vector representation of neurons, and it will no longer be possible to apply unambiguous layers. Another layer of innerProduct is then applied to put the layers in sequence. Thus, the last one will have an output parameter that equal to the number of classes needed for the classification. The K final values will be the parameters of a probability function that allows the final classification. In the training phase, the network ends with an *Accuracy layer* for network accuracy calculation, and with a *Loss layer* for the calculation of the error function needed for a correct and useful training phase. In the classification phase, instead, a *SoftMax layer* is the final layer to classify new images, which are not included in the data set (i.e., the test set images in our experiments). This layer calculates the likelihood of the most appropriate class in the grading phase, and therefore, its output represents the final solution.

4.2 The structure of the AlexNet CNN

AlexNet [10], the second network that we used in our experiments, is a network presented as a winner of the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), on the ImageNet [52] data set, which includes ≈ 1.2 million pictures representing 1000 different objects in over 22000 categories. [53] Feed-forward networks could offer the power needed for such a huge data set, requiring much preprocessing work. Using still modern techniques, such as data augmentation and dropout, AlexaNet exploited the benefits of CNN and backed them up with record-breaking performance in the competition. The AlexaNet CNN is used in several applications; it consists of five convolutional layers, followed by three fully connected layers. Also, three max-pooling layers are inserted after, respectively, the first, second, and fifth convolutional layer, while the first two fully-connected layers are followed by a dropout layer, to avoid overfitting.

5 Image collection and Training phase

The first data set is a generalized faces data set, including faces from different ethnicity, gender, age: the *10k US Adult Faces database* [57] from the Maryland Laboratory of Brain and Cognition of the USA National Institute of Mental Health, which includes 637 faces images labeled with *Neutral* state and 1511 with *Joy* emotion.
For the experiments regarding a single user, we collected images for *Joy*, *Disgust*, and *Neutral*. During the training phase, the subject was presented with a list of videos and pictures, selected to elicit a particular emotion in the audience. In particular, a set of 62 short videos was used to elicit *Joy*, whereas 140 images were selected for *Disgust*. The participants were asked to sit, one by one, and watch the videos/images, while their reactions were recorded by the camera. Later, samples were extracted from the sections of the videos, which showed an evident reaction to the stimuli, at a rate of three frames per second. For the *Neutral* state, the test subject's expression was recorded watching relaxing images, where no particular emotion was elicited. As a basic rule, we decided that no media could be watched more than once, as the reaction would not be spontaneous anymore in case of multiple views. The collected samples were then shuffled, to equitably distribute frames belonging to the same sequences between training set, validation set, and test set.

6 Experiments design and results

Three experiments were performed for this work, using the previously described two networks, i.e., AlexNet and Emex. In the first experiment, the networks were trained and tested on the data set composed of samples that we collected from our test subjects; in the second experiment, the same test was repeated on the *10k US Adult Faces* database. Finally, a cross-domain experiment was conducted, testing the networks trained on 10k US Adult Faces database sample to recognize emotions in our single-users data sets.

6.1 Single-user test

The first test was conducted on the samples collected from each test subjects, in order to see how the networks perform training and testing on the same user in different conditions, e.g., before and after a degenerative pathology, which may prevent the patient from expressing his own emotions and related needs in words. Results, shown in Table 1, show that both networks easily overfit. During the training phase, a perfect accuracy (i.e., 1) was achieved after a few iterations: 50 for the EmEx network, and 150 for the Alexnet network. Further iterations were not necessary, because both networks showed a constant behavior, correctly classifying all the test images. The different training time to obtain the best performances is due to the much higher complexity of the AlexNet network with respect to EmEx, [40][42] in terms of the number of parameters to be optimized.

Table 1. Results of tests using both test networks on the single-user data set

Network	Training Steps	Accuracy	Micro-Averaged F1
AlexNet	50	0.5437	0.5437
AlexNet	100	0.5340	0.5340
AlexNet	150	1	1
AlexNet	200	0.6484	0.6484
EmEx	50	1	1
EmEx	100	1	1
EmEx	150	1	1
EmEx	200	1	1

AlexNet was originally designed for a much more complex problem, as stated in section 4.2, i.e., the identification of objects belonging to an extremely high number of classes. It is worth noticing that, although our task was quicker to tackle, inter-class differences may be less evident for our task than for the original ImageNet data. Therefore, our problem is more difficult to solve. Furthermore, the number of training samples used in our experiment was purposefully small, to assess performances in a context where high computing capabilities and data sets are not available, e.g., where the user can train emotional expressions on a mobile environment or a common desktop/laptop, with a relatively small number of images. Moreover, if such images are shot through a video, they will have less intra-dataset differences.

6.2 Multiple-users test

The second test was performed on the *10k US Adult Faces database* [57], including multiple-users images. This test includes only *Joy* and *Neutral*, due to the lack of enough training samples for *Disgust*. For both classes, 444 samples were included in the training set, for a total of 888 images, while the validation test set comprised 56 images per class. All the images that were left out, i.e., 814 for joy and 56 for neutral, were used to calculate the metrics. The settings used for the network training, which differ from the original ones, are:

- Train batch size: 10
- Test batch size: 16
- Test iterations: 7
- Test interval: 50
- Maximum number of iterations: 1000
- Random state seed: 1234

The ADAM [51] optimizer was used on both networks. *AlexNet* achieved a maximum training accuracy of ≈ 0.84 after 400 iterations, while *EmEx* peaked at a higher ≈ 0.91 score after 800 iterations. However, as shown in the results, AlexNet holds a better generalization ability, thanks to its complexity, achieving both higher accuracy and F1 scores with respect to the EmEx network. Complete results are reported in table 2 and figure 4.

Table 2. Results of tests using the two networks on the 10k US Adult Faces database images. Best figures for AlexNet in bold; for EmEx in italic bold

	AlexNet		EmEx	
Steps	Accuracy	F1	Accuracy	F1
100	0.0644	0.0000	0.8264	0.8993
200	0.0644	0.0000	0.8161	0.8922
300	0.8713	0.9272	0.7621	0.8549
400	0.8506	0.9140	0.7816	0.8684
500	0.8172	0.8923	0.7575	0.8517
600	0.8897	0.9385	0.7540	0.8491
700	**0.8966**	**0.9426**	0.8276	0.8992
800	0.7770	0.8651	***0.8517***	***0.9145***
900	0.8207	0.8949	0.8115	0.8886
1000	0.6931	0.8044	0.7701	0.8603

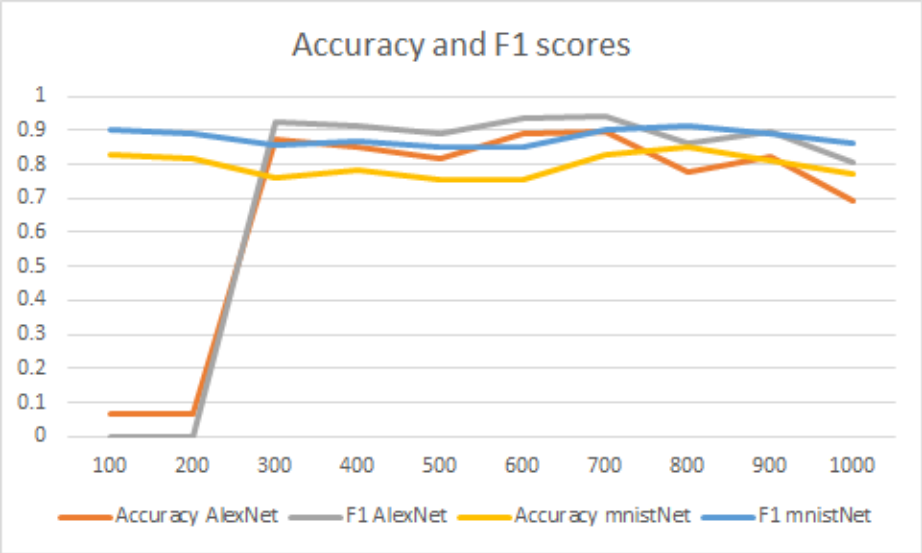

Fig. 4. Recognition performance of the two networks on the test samples

6.3 Cross Test

Both the AlexNet and EmEx networks trained on the 10k US Adult Faces database were tested on the data set created for the Single-user test in section 6.1; results are reported in table 3. Interestingly, the EmEx network performed

Table 3. Results of tests using the two networks trained on the 10k US Adult Faces database on the Single-user data set. Best figures for AlexNet in bold, for EmEx in italic bold

	AlexNet		EmEx	
Steps	Accuracy	F1	Accuracy	F1
100	0.5054	0	0.6344	0.6793
200	0.7098	0.5846	0.4301	0.4647
300	0.6237	0.3860	0.8065	0.7568
400	0.6990	0.5625	0.8602	0.8354
500	0.5484	0.2500	***0.8925***	***0.8781***
600	0.6667	0.4918	0.8495	0.8205
700	0.6667	0.4918	0.8280	0.7895
800	0.6129	0.3571	0.8172	0.7733
900	**0.7204**	**0.6061**	0.8172	0.7733
1000	0.5699	0.2308	0.8172	0.7733

consistingly better than AlexNet, showing a better generalization capability in a completely different environment from the one it was trained for, e.g., with respect to light, user position, and image quality. This is probably due to the ability of AlexNet to better adapt to the peculiar characteristics of the data set it is trained on, thanks to its complexity, but having problems in a fairly different settings when no samples are given.

7 Conclusions and future work

In this work, we described a framework for Emotion Recognition from mouth expressions. Experiments, conducted on both single-user and generalized faces data sets, show good recognition performances of the framework, which can correctly identify the chosen emotions, using limited computational resources and doing it both online and offline. Results show that mouth expressions play an essential role in defining the emotion conveyed by the subject, and can be exploited with low computational power and complexity of systems. Both the tested networks achieved high recognition performances, with the AlexNet network better adapting to the single data set, and a seemingly better ability of the EmEx network to generalize the domain. Future works will investigate the ability of the framework in recognizing more emotions, and include the publication of our single-users image collection.

References

1. T. F. Cootes and C. J. Taylor, D. H. Cooper and J. Graham, Active Shape Models-Their Training and Application, Computer Vision and Image Understanding, Vol. 61, N. 1, pp. 38–59, DOI:1077-3142/95, 1995.
2. Rainer Stiefelhagen and Jie Yang and Alex Waibel: A Model-Based Gaze Tracking System,International Journal of Artificial Intelligence Tools, Vol. 6, N. 2, pp. 193–209, 1997.
3. Picard, R.W., Vyzas, E., Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, Issue: 10, pp. 1175–1191, 2001.
4. Picard, R.W. Affective computing: Challenges, International Journal of Human Computer Studies, Vol. 59, Issues: 1-2, pp. 55–64, DOI:10.1016/S1071-5819(03)00052-1, 2003.
5. Franzoni, V., Gervasi, O., Guidelines for web usability and accessibility on the Nintendo Wii, Transactions in Computer Science VI, part of Lecture Notes in Computer Science Vol. 5730, pp. 19–40, DOI:10.1007/978-3-642-10649-1_2, 2009.
6. Gervasi, O., Magni, R., Zampolini, M., Nu!RehaVR: Virtual reality in neuro tele–rehabilitation of patients with traumatic brain injury and stroke, Virtual Reality, Vol. 14, Issue: 2, pp. 131–141, DOI:10.1007/s10055-009-0149-7, 2010.
7. 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, Vol. 14 Issue:7, pp. 663–680, DOI:10.1007/s00779-010-0294-8, 2010.
8. Pantelopoulos, A., Bourbakis, N.G., A survey on wearable sensor-based systems for health monitoring and prognosis IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, Vol. 40 Issue:1, art. no. 5306098, pp. 1–12, 2010.
9. Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A., Hołyst, J.A., Collective emotions online and their influence on community life PLoS ONE, Vol. 6 Issue:7, art. no. e22207, pp. 1–8, DOI:10.1371/journal.pone.0022207, 2011.
10. Alex Krizhevsky, Ilya Sutskever and Geoff Hinton: Imagenet classification with deep convolutional neural networks, 25th International Conference on Advance in Neural Information Processing System,pag.1106–1114, 2012.
11. Gervasi, O., Russo, D., Vella, F,: The AES Implantation Based on OpenCL for Multi/many Core Architecture. 2010 International Conference on Computational Science and Its Applications, Fukuoka, ICCSA 2010, pages 129–134, Washington, DC, USA,IEEE Computer Society, DOI: 10.1109/ICCSA.2010.44, 2010.
12. Vella, F., Neri, I., Gervasi, O., Tasso, S. A simulation framework for scheduling performance evaluation on CPU-GPU heterogeneous system, Lecture Notes in Computer Science, 7336, ICCSA 2012, pp. 457–469, Springer, DOI: 10.1007/978-3-642-31128-4_34, (2012).
13. Mariotti, M., Gervasi, O., Vella, F., Cuzzocrea, A., Costantini, A.: Strategies and systems towards grids and clouds integration:A DBMS-based solution, Future Generation Computer Systems, vol. 88, pp. 718–729, http://dx.doi.org/10.1016/j.future.2017.02.047, 2018.
14. Neumann, M., Patricia, N., Garnett, R., Kersting, K. Efficient graph kernels by randomization Lecture Notes in Computer Science, 7523 LNAI (PART 1), pp. 378–393, 2012.

15. Franzoni, V., Poggioni, V., Zollo, F. Automated classification of book blurbs according to the emotional tags of the social network Zazie CEUR Workshop Proceedings, Vol. 1096, pp. 83–94, DOI:10.13140/RG.2.1.3194.7689, 2013.
16. Bertola, F., Patti, V. Emotional responses to artworks in online collections UMAP Workshops Proceedings, 997, 2013.
17. Saif, H., Fernandez, M., He, Y., Alani, H., Evaluation datasets for Twitter sentiment analysis a survey and a new dataset, the STS-Gold,CEUR Workshop Proceedings, 1096, pp. 9–21, 2013.
18. Cieliebak, M., Dürr , O., Uzdilli, F. Potential and limitations of commercial sentiment detection tools, CEUR Workshop Proceedings, 1096, pp. 47–58, 2013.
19. Tasso, S., Pallottelli, S., Rui, M., Laganá, A., Learning objects efficient handling in a federation of science distributed repositories, Lecture Notes in Computer Science, 8579 LNCS (PART 1), pp. 615–626, DOI:10.1007/978-3-319-09144-0_42, 2014.
20. LeCun, Yann and Bengio, Yoshua and Hinton, Geoffrey: Deep learning, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved, ISBN: 0028-0836, DOI:10.1038/nature14539, 7553, pag.436-444, vol.521, 2015,
21. Patel, V.M., Gopalan, R., Li, R., Chellappa, R. Visual Domain Adaptation: A survey of recent advances, IEEE Signal Processing Magazine, Vol. 32, Issue: 3, art. no. 7078994, pp. 53–69, 2015.
22. Peng, K.-C., Chen, T., Sadovnik, A., Gallagher, A., A mixed bag of emotions: Model, predict, and transfer emotion distributions, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015, art. no. 7298687, pp. 860–868, 2015.
23. Voirin, N., Payet, C., Barrat, A., Cattuto, C., Khanafer, N., Regis, C., Kim, B.-A., Comte, B., Casalegno, J.-S., Lina, B., Vanhems, P., Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital, Infection Control and Hospital Epidemiology, Vol. 36, Issue: 3, pp. 254–260, 2015.
24. Phelan, I., Arden, M., Garcia, C., Roast, C., Exploring virtual reality and prosthetic training 2015 IEEE Virtual Reality Conference, VR 2015 - Proceedings, art. no. 7223441, pp. 353–354, 2015.
25. Franzoni, V., Milani, A., Pallottelli, S., Leung, C.H.C., Li, Y., Context-based image semantic similarity 12th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2015, art. no. 7382127, pp. 1280–1284, 2015.
26. Pallottelli S., Franzoni V., Milani A., Multi-path traces in semantic graphs for latent knowledge elicitation, Proceedings - International Conference on Natural Computation, 2016-January, 281-288, 2016, DOI:10.1109/ICNC.2015.7378004
27. Franzoni V., Milani A., Semantic context extraction from collaborative networks, Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2015, 131-136, 2015, DOI:10.1109/CSCWD.2015.7230946
28. Lewis, Trent W. and Powers, David M. W.: lip contour detection techniques based on front view of face, Journal of Global Research in Computer Science, Vol. 2, N. 5, pag.43-46, ISSN: 2229-371X, 2011.
29. Gervasi, Osvaldo, Magni, Riccardo and Ferri, Matteo: A Method for Predicting Words by Interpreting Labial Movements, Lecture Notes in Computer Science, vol. 9787, pages 450–464, ICCSA 2016, Beijing (China), Springer, DOI: 10.1007/978-3-319-42108-7_34, 2016.
30. Bastianelli, E., Nardi, D., Aiello, L.C., Giacomelli, F., Manes, N., Speaky for robots: the development of vocal interfaces for robotic applications, Applied Intelligence, Vol. 44, Issue:1, pp. 43–66, DOI:10.1007/s10489-015-0695-5, 2016.

31. Giulio Biondi, Valentina Franzoni, Yuanxi Li,Alfredo Milani: Web-based similarity for emotion recognition in web objects, Proceedings of the 9th International Conference on Utility and Cloud Computing, UCC 2016, Shanghai, China, December 6-9, 2016, pag. 327–332, 2016.
32. Alkuhlani, A., Nassef, M., Farag, I. Multistage feature selection approach for high-dimensional cancer data, Soft Computing, Vol. 21, Issue: 22, pp. 6895—6906, 2017.
33. Canossa, A., Badler, J., El-Nasr, M.S., Anderson, E. Eliciting emotions in design of games - A theory driven approach, CEUR Workshop Proceedings, 1680, pp. 34-40, 2016.
34. Caffe Framework:
Github:`https://github.com/BVLC/caffe`, last visited on Sept 12, 2018.
35. Lou, Y., Wu, W., Vatavu, R.-D., Tsai, W.T., Personalized gesture interactions for cyber-physical smart-home environments, Science China Information Sciences, Vol. 60, Issue: 7, DOI:10.1007/s11432-015-1014-7, 2017.
36. Franzoni, V., Chiancone, A., Milani, A., A multistrain bacterial diffusion model for link prediction, International Journal of Pattern Recognition and Artificial Intelligence, Vol.31, Issue 11, World Scientific, DOI:10.1142/S0218001417590248, 2017.
37. Cui, W., Du, Y., Shen, Z., Zhou, Y., Li, J., Personalized microblog recommendation using sentimental features, 2017 IEEE International Conference on Big Data and Smart Computing, BigComp 2017, art. no. 7881756, pp. 455–456, DOI:10.1109/BIGCOMP.2017.7881756, 2017.
38. P. Angelov, X. Gu, J. A. Iglesias, A. Ledezema, A. Sanchis, O. Sipele, and R. Ramezani, Cybernetics of the Mind, Learning Individual's Perceptions Autonomously, IEEE Systems, Man, and Cybernetics Magazine, Vol. 3, N.2., pp. 6–17, ISSN:2333-942X, DOI:10.1109/MSMC.2017.2664478, 2017.
39. Franzoni, V., Milani, A., Vallverdu, J., Emotional Affordances in Human-Machine Interactive Planning and Negotiation. Proceedings of WI 2017, Workshop on Affective Computing and Emotion Recognition (ACER), pp. 924–930, DOI:10.1145/3106426.3109421, 2017.
40. Riganelli M., Franzoni V., Gervasi O., and Tasso S., EmEx, a Tool for Automated Emotive Face Recognition Using Convolutional Neural Networks, ICCSA 2017, Workshop on Emotion Recognition, Lecture Notes in Computer Science, vol. 10406, pp. 692-704, DOI:10.1007/978-3-319-62398-6_49, 2017.
41. V. Franzoni, Autonomous Hexapod Robot With Artificial Vision and Remote Control by Myo-Electric Gestures. Cyber-Physical Systems for Next-Generation Networks (2018) pp. 143-162 DOI: 10.4018/978-1-5225-5510-0.ch007
42. O.Gervasi, V. Franzoni, A. Riganelli, S. Tasso, Automating facial emotion recognition. Web Intelligence, Volume 17, Issue 1, (2019) pp. 17-27 DOI: 10.3233/WEB-190397
43. G.Mezzetti, Design and Experimentation of Target-Driven Visual Navigation in Simulated and Real Environment via Deep Reinforcement Learning Architecture for Robotics Applications, Master laurea thesis, University of Perugia (2019)
44. Farabet, Clement and Couprie, Camille and Najman, Laurent and LeCun, Yann, Learning hierarchical features for scene labeling. IEEE transactions on pattern analysis and machine intelligence v. 35 n.8 (2013)
45. Fahlman, Scott E., Geoffrey E. Hinton, and Terrence J. Sejnowski. "Massively parallel architectures for Al: NETL, Thistle, and Boltzmann machines." National Conference on Artificial Intelligence, AAAI. 1983.

46. Plaut, D. C., S. J. Nowlan, and G. E. Hinton. "Experiments on learning by Back-propagation Technical Report CMU–CS–86–126." Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA (1986).
47. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.
48. Bradski, G.. The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000).
49. Viola, P., and Jones, M.. Rapid object detection using a boosted cascade of simple features. (2005)
50. Castrillón Santana, M., Déniz Suárez, O., Hernández Sosa, D., and Lorenzo Navarro, J.. Using Incremental Principal Component Analysis to Learn a Gender Classifier Automatically. In 1st Spanish Workshop on Biometrics. Girona, Spain (2007)
51. Kingma, D. P., and Ba, J. Adam: A Method for Stochastic Optimization. (2014)
52. Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. . ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition. (2009)
53. . AlexNet Caffe Implementation. [online] Available at: https://github.com/weiliu89/caffe/tree/ssd/models/bvlc_alexnet [Accessed 2019].
54. Ekman, P. (1992). An Argument for Basic Emotions. Cognition and Emotion. https://doi.org/10.1080/02699939208411068
55. Franzoni, V., Milani, A., Emotion Recognition for Self-aid in Addiction Treatment, Psychotherapy, and Nonviolent Communication. Lecture Notes in Computer Science, ICCSA Conference, 2019.
56. Franzoni, V., Milani, A., Nardi, D., Vallverdú, J.: Emotional machines: The next revolution. Web Intelligence 17(1): 1-7 (2019)
57. Bainbridge, W.A., Isola, P., Oliva, A., The intrinsic memorability of face images. Journal of Experimental Psychology: General. Journal of Experimental Psychology: General, 142(4), 1323-1334. (2013)
58. Milani A., Poggioni V., Planning in reactive environments, Computational Intelligence, 23, 4, 439-463, (2007), Blackwell, DOI:10.1111/j.1467-8640.2007.00315.x
59. Baioletti M., Milani A., Poggioni V., Rossi F., Experimental evaluation of pheromone models in ACOPlan, Annals of Mathematics and Artificial Intelligence, 62, 43528, 187-217, 2011, DOI:10.1007/s10472-011-9265-7
60. Ukey N., Niyogi R., Singh K., Milani A., Poggioni V., A Bidirectional Heuristic Search for web service composition with costs, International Journal of Web and Grid Services, 6, 2, 160-175, (2010), DOI:10.1504/IJWGS.2010.033790
61. Chiancone A., Franzoni V., Niyogi R., Milani A., Improving Link Ranking Quality by Quasi-Common Neighbourhood, Proc. of 15th ICCSA 2015, 21-26, (2015), DOI:10.1109/ICCSA.2015.19
62. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791 (1998)