Building an ecologically valid facial expression database – Behind the scenes

Building an ecologically valid facial expression database – behind the scenes

Francesca Nonis[1], Luca Ulrich[*,1], Nicolò Dozio[2], Francesca Giada Antonaci[1], Enrico Vezzetti[1], Francesco Ferrise[2], and Federica Marcolin[1]

[*] Corresponding author: luca.ulrich@polito.it
[1] Department of Management & Production Engineering, Politecnico di Torino
[2] Department of Mechanical Engineering, Politecnico di Milano

Abstract. Artificial Intelligence (AI) algorithms, together with a general increased computational performance, allow nowadays exploring the use of Facial Expression Recognition (FER) as a method of recognizing human emotion through the use of neural networks. The interest in facial emotion and expression recognition in real-life situations is one of the current cutting-edge research challenges. In this context, the creation of an ecologically valid facial expression database is crucial. To this aim, a controlled experiment has been designed, in which thirty-five subjects aged 18-35 were asked to react spontaneously to a set of 48 validated images from two affective databases, IAPS and GAPED. According to the Self-Assessment Manikin, participants were asked to rate images on a 9-points visual scale on valence and arousal. Furthermore, they were asked to select one of the six Ekman's basic emotions. During the experiment, an RGB-D camera was also used to record spontaneous facial expressions aroused in participants storing both the color and the depth frames to feed a Convolutional Neural Network (CNN) to perform FER. In every case, the prevalent emotion pointed out in the questionnaires matched with the expected emotion. CNN obtained a recognition rate of 75,02%, computed comparing the neural network results with the evaluations given by a human observer. These preliminary results have confirmed that this experimental setting is an effective starting point for building an ecologically valid database.

Keywords: Facial Expression Recognition, ecologically-valid data, 3D facial database, basic emotions, affective database, human-robot interaction.

1 Introduction

Automatic classifiers are spreading in a variety of fields and are strongly used even in facial expression recognition applications. The main problem in using supervised classifiers to perform facial expression recognition is to find valid data to train machine learning algorithms. Data that have to be inputted for the training phase must be labelled, and to do such an operation, a scientific approach to describe emotions is necessary.

The first study dealing with quantification of emotions was initiated by Wundt [1] and continued by Schlosberg [2], that introduced a three-dimensional model which dimensions were pleasant-unpleasant, tension-relaxion, and excitation-calm. Ekman [3] recommended to merge the two last dimensions because they resulted too similar each other, and Russell [4] developed the Circumplex Model of Affect, that has been taken as a reference in the present work and is shown in **Fig. 1**.

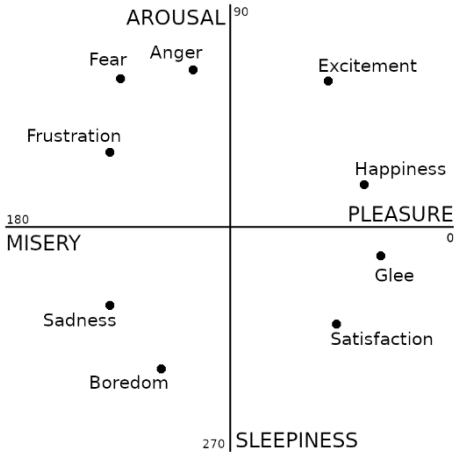

Fig. 1. Circumplex Model of Affect. On the left the eight affect concepts are arranged in a circular order, as well as the twenty-eight affect words that are displayed on the right.

Ecological validity refers to the possibility to generalize the data collected from observed behavior in the laboratory to the natural behavior in the real world [5]. In the current study, the facial expression of a subject must be due to a certain stimulus and not to boundary or artificial conditions. Within the context of an experiment, this means to find a trade-off between the experimental rigor, necessary to compare results obtained from different subjects, and the comfortability of the subjects themselves, who should express feelings by their nature only as a consequence to the stimulus received, not conditioned by constraints imposed by the experiment.

An experiment that aims to study facial expressions can be set up in different ways. First of all, participants can be asked to act or to express their spontaneous feelings as has been requested for the current work; then, the format of results must be established [6] and, consequently, the necessary equipment must be procured (a standard video camera, an RGB-D camera, sensors to obtain physiological data, a database to store answers to a questionnaire, etc.); moreover, stimuli to obtain spontaneous reactions must be defined.

Among the variety of stimuli raising emotions, such as audio-visual [7], movie clips [8], music tracks and game scenarios [9], for the present work images stored in IAPS and GAPED affective databases have been chosen.

The International Affective Picture System (IAPS) is the most known affective database. The version of the database used for this experiment is composed of 1182

images (the database has been subjected to updates through the years) subdivided into semantic categories to arouse different emotions. IAPS has been largely used in psychiatric applications. Taskiran et al. [10] have studied the responses to emotional stimuli in patients affected by attention-deficit hyperactivity disorder, Migliore et al. [11] have conducted a similar study in Relapsing-Remitting Multiple Sclerosis (RRMS) patients, Moret-Tatay et al. [12] and Bekele et al. [13] have dealt with middle-aged adults and older men respectively affected by Schizophrenia, Peter et al. [14] have compared emotional responses in subjects with personality disorders, cluster-C personality disorders and non-patients. Furthermore, the ability of processing emotions after traumatic situations has been investigated, such as post-earthquake distress by Pistoia et al. [15] and violated women by Navarro Martinez [16], but also the dependence from alcohol can have implications in the way of processing emotions according to Dominguez-Centeno et al. [17]. The vast majority of the above-mentioned studies focused on evaluating subjects' emotions through the analysis of physiological signals: electroencephalography (EEG), electrocardiography (ECG) and magnetic resonance imaging (MRI).

The Geneva Affective Picture Database (GAPED) is composed of 730 images. The intent in building this new dataset was to overcome a problem that arouses in using IAPS extensively: the impact of those images seemed to drop in terms of efficacy both for positive and negative emotions. In particular, regarding the negative ones, GAPED designers subdivided images into four classes: two of them represents animals (snakes and spiders), one concerns the violation of the social norms (defined by legality), one concerns the violation of personal norms (determined by morality). According to Dan-Glauser and Scherer [18], estimation of low tolerability of the stimuli related to social norms becomes relevant in the elicitation of anger, but also in disgust, pity, guilt, shame, and contempt. There are predictable dissimilarities in valence marks among the positive, neutral, and negative categories, but also in arousal rates, indeed it is usually possible to find a correlation since valence scores are rarely independent from arousal levels.

To quantify an emotion is a critical task that has been largely discussed; what people refer to when using the term *feeling* is only the conscious experience of an emotion. Nonetheless, Mehrabian and Russell [19] have developed the Semantic Differential Scale to assess objects, events, and situations by using 18 opposite couples of adjectives related to three independent dimensions:

- Valence/pleasure: it describes the positivity or negativity of an emotion. In the work of Mehrabian and Russell, adjectives used to label valence were not only in the range happy-unhappy; the concept of positivity was also associated to pleasantness, satisfaction, content, hope and relaxation, while the concept of negativity was associated to annoyance, unsatisfaction, melancholy, despair, and boredom.
- Arousal: it describes the level of activation inducted by the received stimulus, in terms of psychophysical response. The continuum ranges from the lowest level associated with a status of boredom and sleepiness, to the highest level of frantic excitement. Adjectives used to label this dimension are aroused-unaroused, stimulated-relaxed, excited-calm, awake-sleepy, frenzied-sluggish, jittery-dull.

- Dominance: it describes how a subject feels with regards to the aroused emotion in terms of submission-dominance. It is the most critical dimension to define because of possible misinterpretations; for example, one subject may consider her/his sense of control in the situation presented, while another subject could consider whether the pictured object is perceived in control or not of that situation.

The Self-Assessment Manikin (SAM) [20] is a solution that maps the three dimensions into three non-verbal pictorial scales in order to directly assess the pleasure, arousal, and dominance associated in response to an object or event.

Valence ranges from pleasant to unpleasant; in the SAM implementation selected for this experiment, the lowest value is represented by a frowning figure, while the highest value is represented by a smiling figure.

Arousal ranges from calm to excited; in the SAM implementation selected for this experiment, the lowest value is represented by a sleepy figure, while the highest value is represented by a wide-eyed figure.

Dominance has not been used in this experiment not to move the focus of the participants forcing them to answer a too demanding questionnaire. Furthermore, in literature and in describing images in affective databases, it is the least used dimension.

SAM has been the selected scale representation in this study.

As reported in [21], participants responses have been collected on a 9-point rating scale for each dimension.

In the experimental setup for the current work, participants have seen images chosen from affective databases and have answered a short questionnaire to express their own feeling elicited by each image. They have been recorded using the RGB-D camera Intel RealSense SR300. The final aim of this project is to build an ecologically valid dataset within which RGB-D images are stored. These images should represent spontaneous facial expressions, indispensable to train deep learning neural networks, such as Convolutional Neural Network, or other supervised machine learning algorithms.

The result that has been obtained up to now is twofold: on one side, a comparison between elicited emotions and expected emotions has been drawn up; on the other side, a remarkable recognition rate of a CNN trained with the obtained images has been achieved.

2 Methods

In this Section, all the components involved in the experiment design and implementation phase are presented (**Fig. 2**).

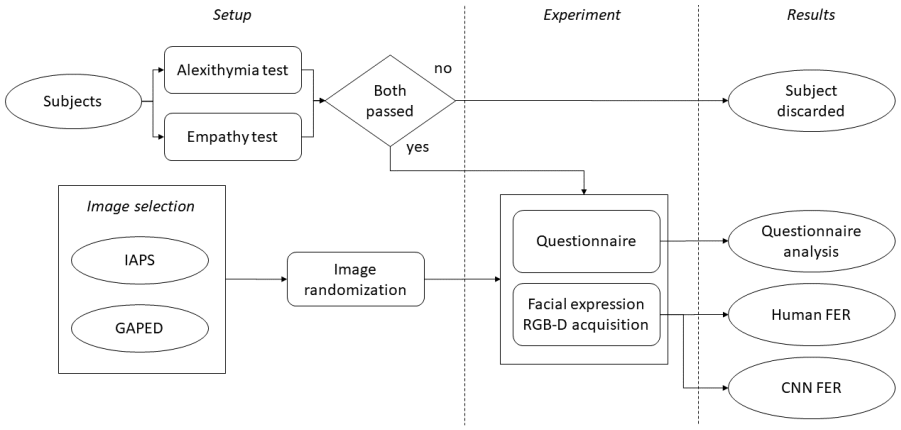

Fig. 2. Methodology steps.

2.1 Participants

Participants have been selected among students and PhD students of Politecnico di Torino, aged between 18 and 35, for a total number of thirty-five participants, fourteen female and twenty-one male subjects.

The experiment's nature has required to ensure that participants had at least a standard level of empathy and were not alexithymic; thus, every participant filled-in two tests before attending the experiment itself.

Empathy identifies the ability to identify and to understand others' points of view, thoughts, intentions, and beliefs and is fundamental to build interpersonal relationships [22]. Empathy has two main components: the affective one and the cognitive one. The first refers to the affective reaction to another person's emotional state, and the latter refers to the cognitive capacity to take the perspective of the other person [23]. The Balanced Emotional Empathy Scale (BEES) proposed by Mehrabian and translated into Italian by Meneghini et al. [24] has been adopted to evaluate participants' empathy. This test is composed of thirty questions, and each answer requires a choice between strongly disagree and strongly agree on a seven-point scale. Results are subdivided into three ranges, namely below the average, standard and above the average.

Alexithymia identifies a reduced ability in recognizing, describing, and understanding one's own emotions [23]. It goes without saying that alexithymia and empathy are interlinked because if one has difficulties recognizing his own emotions, that person will have difficulties in recognizing others' emotions. The Toronto Structured Interview for Alexithymia (TAS-20) has been adopted to evaluate participants' alexithymia [25]; this test has a 3-factor structure: the first evaluates difficulty in recognizing feelings, the second evaluates the difficulty in describing feelings, the third one considers externally oriented thinking. There are twenty questions, and each answer requires a choice between strongly disagree and strongly agree on a five-point scale. Results are subdivided in three ranges, namely non-alexithymic, borderline and alexithymic.

Subjects' empathy and alexithymia results have been displayed in **Fig. 3**.

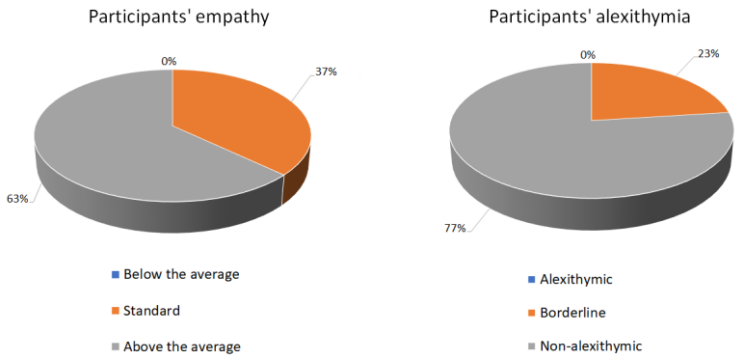

Fig. 3. On the left: participants' alexithymia according to the Toronto Structured Interview for Alexithymia (TAS-20). On the right: Participants' empathy according to the Balanced Emotional Empathy Scale (BEES).

2.2 Images selection

The choice of static visual stimuli, i.e., images, has been done to obtain an important advantage, which is to identify the exact moment during which the image is shown to the participants, namely the moment from which it is reasonable to search for a facial expression in the video acquired with the camera.

Databases from which to gather the pictures have been IAPS and GAPED. This choice has been made after the literature review and, also, a trial day dedicated to identifying weaknesses in the planned design of the experiment. The chosen pictures have been selected to arouse the widest range of stimuli possible and to represent a selection of a comprehensive sample of contents across the entire affective space.

Unfortunately, the literature lacks a unique validated system to relate the Russell model (**Fig. 1**), the six Ekman's expressions (happiness, sadness, anger, fear, disgust, and surprise), and the images stored in the IAPS database, that have been classified according to valence and arousal values. Nonetheless, an attempt to put in relation these dependent dimensions has been done, according to the theory of emotions elaborated by Plutchik [26]. This effort has been done to be able to select the IAPS proper images to arouse specific stimuli without having images organized by emotions (there is no emotion label on IAPS images), but only by valence and arousal values. Results of this mapping operation are shown in **Fig. 4** and **Fig. 5**.

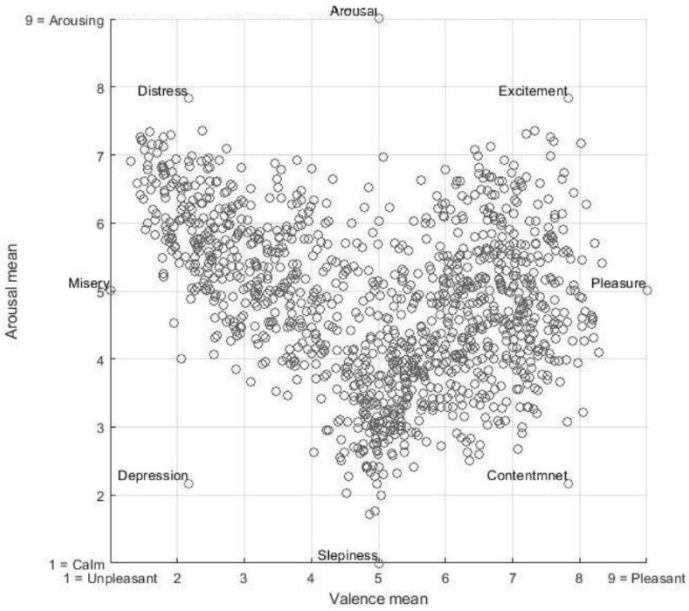

Fig. 4. Valence-arousal and Russel's model mapping (affective space).

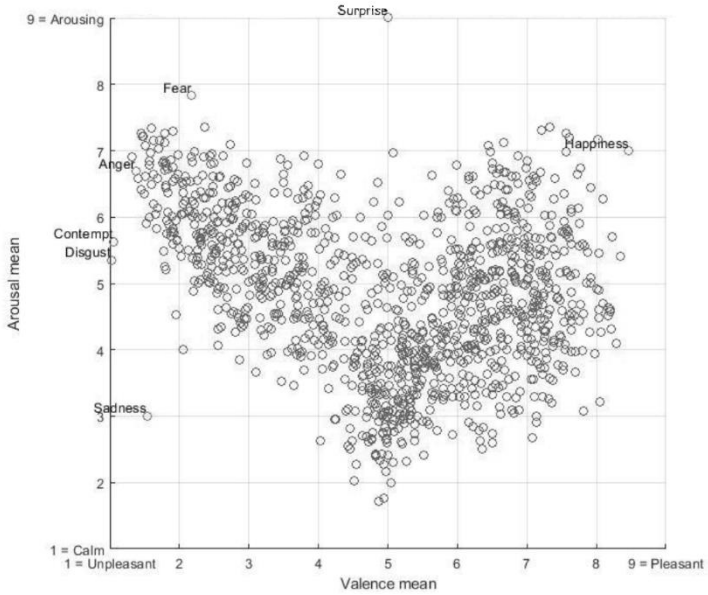

Fig. 5. Valence-arousal and Plutchik's theory of emotions mapping.

Images stored in the GAPED database had both a labelling system to describe which emotions should arouse and the scores of valence and arousal, even if ranging from 0

to 100, so a normalization has been performed. Despite this operation, a poor correlation between values of IAPS and GAPED associated with positive emotions has been obtained. Some examples of inconsistency are reported in **Table 1**.

Table 1. Valence and arousal comparisons for positive images in IAPS and GAPED databases.

	IAPS		GAPED	
Images subject	Valence	Arousal	Valence	Arousal
Puppies	8.34	5.41	8.68	3.3
Baby	7.86	5	8.37	3.2
Seal	8.19	4.61	8.61	1.68

These issues in making correspondences between the two datasets can be explained by the different characteristics of the images, IAPS ones look older, even if GAPED is only six years more recent the most used version of IAPS, but mostly because of cultural factors of people that evaluated the pictures; indeed, IAPS has been developed by the National Institute of Mental Health Center for Emotion and Attention at the University of Florida and images have been evaluated by one hundred people aged 18-24, while GAPED comes from Geneva, Switzerland, and images have been evaluated by sixty people aged 19-34.

To solve the issue, all the images have been carefully selected one-by-one.

GAPED images have been selected more easily since the database is arranged in folders (humans, animals, neutrals, spiders, snakes, and positive categories).

IAPS images have been selected assigning the correct labels of emotions to each picture considering in a first instance the work of Bradley & Lang [27, 28], and Bradley et al. [29]. The contents that generate the same emotion in the two genders has been considered, and the emotion with the major percentage has been taken as predominant (**Table 2** has been reported from [28]).

Table 2. List of the most frequent specific emotion descriptors reported in [28]-[29] and linked to the IAPS picture categories selected for our experiment. The table includes also the proportion of men and women selecting that specific emotion to describe their affective experience.

Picture category	Women		Men	
Families	Happy (.79)	Loving (.78)	Happy (.58)	Loving (.58)
Pollution	Disgust (.56)	Irritated (.43)	Disgust (.34)	Irritated (.26)
Loss	Sad (.79)	Pity (.56)	Sad (.61)	Pity (.59)
Illness	Pity (.67)	Sad (.69)	Pity (.58)	Sad (.51)
Contamination	Disgust (.88)	Irritated (.50)	Disgust (.78)	Irritated (.40)
Accidents	Sad (.63)	Pity (55)	Pity (.50)	Sad (.49)
Mutilation	Disgust (.81)	Sad (.47)	Disgust (.75)	Pity (.42)
Animal Threat	Afraid (.69)	Anxious (.31)	Afraid (.42)	Anxious (.23)
Human Threat	Afraid (.67)	Angry (.42)	Afraid (.37)	Angry (.35)

A final check with the affective space has been done before the following list of images was confirmed (24 IAPS images in **Table 3**, 24 GAPED images in **Table 4**). Images used in the training phase have not been reported.

Table 3. Selected IAPS images.

Description	Valence	Arousal	Emotion
Beaten woman	2.31	6.38	Anger
Soldiers	2.10	6.53	Anger
Soldier	1.51	7.07	Anger
Mutilation #1	1.79	7.26	Disgust
Mutilation #2	1.79	7.12	Disgust
Mutilation #3	1.80	6.77	Disgust
Mutilation #4	1.70	7.03	Disgust
Mutilation #5	1.48	7.22	Disgust
Mutilation #6	1.58	6.97	Disgust
Baby with tumor	1.46	7.21	Disgust
Injury	1.56	6.79	Disgust
Snake	3.79	6.93	Fear
Dog attack	3.09	6.51	Fear
Shark	3.85	6.47	Fear
Kiss	7.27	5.16	Happiness
Mushroom #1	5.42	3.00	Neutrality
Mushroom #2	5.15	3.69	Neutrality
Spoon	5.04	2.00	Neutrality
Bowl	4.88	2.33	Neutrality
Lamp	4.87	1.72	Neutrality
Toddler	1.79	5.25	Sadness
Sad child	1.78	5.49	Sadness
Injured child	1.80	5.21	Sadness
Car accident	2.34	6.63	Sadness

Table 4. Selected GAPED images.

Description	Valence	Arousal	Emotion
Animal mistreatment #1	2.12	5.89	Anger
Animal mistreatment #2	2.08	6.46	Anger
Animal mistreatment #3	2.40	6.88	Anger
Animal mistreatment #4	1.15	7.23	Anger

Animal mistreatment #5	1.71	7.46	Anger
Snake #1	4.94	6.09	Fear
Snake #2	2.44	6.5	Fear
Spider #1	4.21	5.44	Fear
Spider #2	4.85	6.4	Fear
Spider #3	3.94	5.63	Fear
Baby #1	8.07	3.38	Happiness
Baby #2	8.03	2.86	Happiness
Baby #3	8.21	2.72	Happiness
Puppies #1	8.19	3.37	Happiness
Puppies #2	8.68	3.3	Happiness
Baby fox	7.83	3.11	Happiness
Kitten	7.77	3.10	Happiness
Antenna	5.4	2.97	Neutrality
Chairs	5.01	2.06	Neutrality
Lamp and sofa	5.84	2.10	Neutrality
Animal in captivity #1	3.20	6.43	Sadness
Animal in captivity #2	1.80	7.48	Sadness
Animal in captivity #3	2.11	6.38	Sadness
Animal in captivity #4	2.08	5.68	Sadness

The final list is also the result of an analysis conducted after the trial day; the number of 48 images was defined instead of the initial 60, a trade-off to use the greatest number of pictures preserving the participants' attention, and some images considered too dated (belonging to IAPS database) have been substituted. Images are uniformly distributed among the basic emotions: anger, disgust, fear, happiness, sadness, and neutrality. Moreover, **Fig. 6** identifies the images of the final dataset onto the Valence-Arousal plane. Surprise is not present among the labelled images.

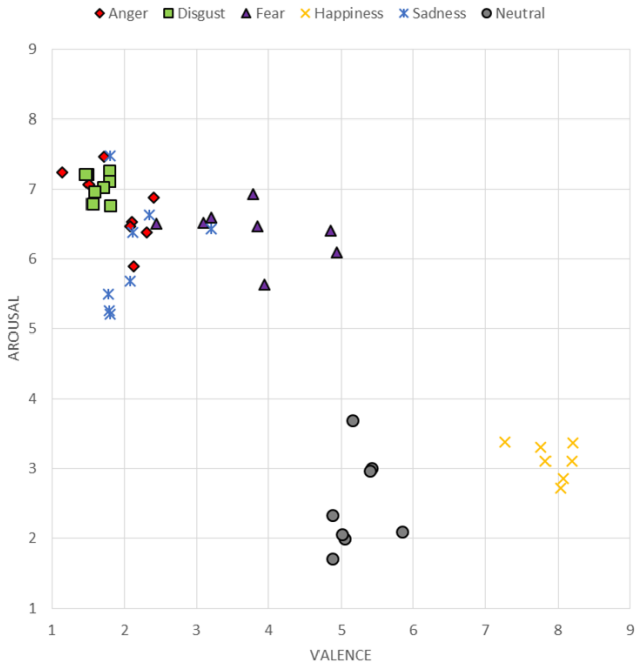

Fig. 6. IAPS and GAPED distribution in terms of valence and arousal.

2.3 Experimental Setup

Participants were asked to fill-in the empathy and alexithymia tests before coming to the laboratory.

Regarding the part of the experiment held in presence, at the beginning participants have attended a presentation to become familiar with the context and to receive the main indications on what to do during the experiment, without influencing their emotionality in any way not to corrupt the results. They have been warned about the presence of images that could have potentially bothered their sensibility. The possibility of abandoning the experiment due to any kind of discomfort has been clarified.

The experiment has taken place in two phases: training and testing. The structure of both the phases has been the same: in a first instance one image provided by affective databases was displayed in full-screen mode (**Fig. 7**), then the participants had to fill-in the questionnaire about valence, arousal and the prevalent felt emotion (**Fig. 8**). It has to be noticed that the label surprise has been inserted in the questionnaire, to let participants free of choosing the most proper basic emotion they felt, independently from the fact that images arousing surprise have not been inserted in the final dataset of 48 images.

Fig. 7. Example of image content selected to arouse happiness [30].

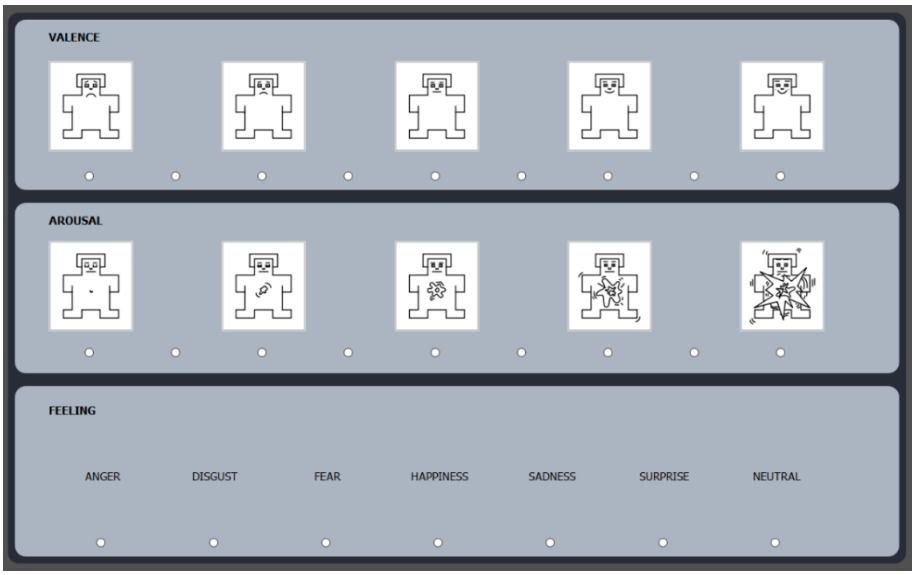

Fig. 8. Screenshot of the questionnaire used for the experiment.

The training phase has been useful mainly to get participants familiar with the questionnaire, because answers were forced to be given in no more than 15 seconds, to favor spontaneity. SAM icons are intuitive, but not so easy to interpret if never seen before.

The testing phase is composed of 48 images which are randomized for every participant and lasts about twenty minutes.

An ad-hoc software has been necessary to deal with both the management of images and questionnaire, maximizing the user experience not to distract the user from his task and the management of the RGB-D camera recording. Indeed, the Intel RealSense SR300 has been connected to the same application using a different thread and has been set up to record user's expression from the moment during which the affective image appears on the screen to two seconds after it disappears, to be sure not to lose any expression. An idle interval of 2000 milliseconds between the affective image and the questionnaire and between the questionnaire and the next affective image has been introduced. The affective image lasts 6000 milliseconds on the screen; nonetheless, a smaller frame has been inserted next to the questionnaire as a reminder for the user.

In **Fig. 9** the experimental setup is illustrated.

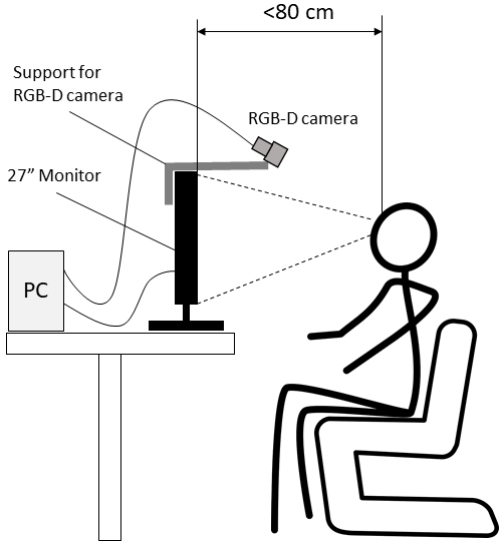

Fig. 9. Experimental setup.

2.4 Facial Expression Recognition via Deep Learning

The RGB-D camera was used to record spontaneous facial expressions aroused in participants, storing both the color and the depth frames, with the purpose of creating a facial depth map database to be adopted for testing novel face/facial expression recognition methodologies. Thus, it was necessary to preliminarily test the acquired data to verify that could be suitable for feeding a Convolutional Neural Network (CNN), the most used neural networks to identify objects and faces within frames. Deep learning methods provide some advantages, including the automatic features extraction and the possibility of retraining existing networks for other recognition activities, with cutting-edge recognition results [31].

Facial Expression Recognition has been performed starting by VGG-16, a convolutional neural network model proposed by Simonyan and Zisserman [32], trained on the Imagenet dataset [33], a database of images of generic objects. Then, the recognition

task has been readapted to faces by training the CNN on the BU-3DFE database [34], a 3D facial expression database that presently contains 100 subjects (56% female, 44% male), ranging age from 18 years to 70 years old, with a variety of ethnic/racial ancestries. All the subjects involved in the database performed seven expressions in front of the 3D face scanner. Excluding the neutral expression, the other six basic emotions (i.e., anger, disgust, happiness, fear, sadness, and surprise) are represented with four levels of intensity resulting in a total of 2,500 3D facial expression models and the corresponding 2D texture images. The training phase is the most demanding step when dealing with deep learning because a huge amount of data is required [35].

Keras was used, an open-source neural-network library written in Python, running on top of TensorFlow, on Windows 10 Pro with NVIDIA GeForce RTX 2060.

Data obtained from the acquisitions, i.e., color and depth frames, must be processed before being used as input for the neural network. Three main steps can be identified: frame capture to manually extract the most significant frames (Maximum Criterion variation) to be analyzed through the neural network; Color to Depth alignment to both temporally and spatially synchronize the frames; Face Detection to identify the face only in its oval shape. The input layer of the network requires RGB images having a size of 224x224.

The testing phase results are discussed in the next Section, together with the other experimental results.

3 Results and Discussion

To evaluate the effectivity of the acquired data in order to create a facial database for emotion recognition, it was core to compare the emotions expected to be aroused and the emotions pointed out in the questionnaire by the participants. This comparison aims to verify if the images chosen from the affective database have been effective.

In **Table 5** the six considered emotions are displayed in the first column. From the second column to the last one, the indication of the emotion pointed out by the 35 users has been reported.

Table 5. Comparisons between expected emotions (first column) and questionnaire answers.

	Emotions reported in the questionnaire						
	Happiness	Neutrality	Sadness	Disgust	Anger	Fear	Surprise
Happiness	79%	16%	1%	0%	0%	0%	4%
Neutrality	8%	75%	2%	1%	0%	1%	13%
Sadness	6%	7%	67%	1%	4%	7%	8%
Disgust	0%	3%	15%	67%	3%	5%	7%
Anger	0%	9%	23%	14%	44%	3%	7%
Fear	4%	26%	0%	26%	0%	27%	17%

It can be noticed that the prevalent emotion found in the questionnaires matches with the expected emotion in every case.

To be coherent with the CNN training and the literature, surprise has been maintained as an available option to choose, even if not directly present among the affective images. In some cases, participants have chosen this emotion instead of neutrality because they did not know how to react. Anyways, 75% of matching between expected and aroused neutrality is remarkable, as well as the 79% of happiness.

Obtained results are perfectly coherent with **Table 2**. For instance, mutilations should arouse disgust both for men and women, then sadness in women and pity in men. Pity is not a basic emotion, the closest one is sadness, and in our study, the mutilations that have been chosen to arouse disgust, have aroused disgust in the 67% and sadness in 15% of the participants.

Anger images have been mostly evaluated as anger (44%) or sadness (23%) or disgust (14%) confirming the not so clear area of the affective space occupied by these three emotions.

Fear has been the emotion aroused with less success (27%). According to Edwards et al. [36], disgust can be part of the emotional reaction to certain phobic stimuli. This explains why it has been chosen from the 26% of the participants, as well as the neutrality, simple to explain that 26% of the participants have not felt these images frightful enough.

After that, emotions have been analyzed, a comparison between valence and arousal values expected from one side, valence and arousal pointed out in the questionnaires on the other side has been carried on.

The 48 images have been represented in the affective space (singularly in **Fig. 10**, compacted in **Fig. 11**), both with valence and arousal values reported in affective databases and with valence and arousal values given by participants' answers to the questionnaire. In this last case, valence and arousal have been averaged among the 35 participants for every image, and to choose the emotion that each valence-arousal couple represents, the most selected emotion by the participants has been used.

Surprise has not been reported in the graphs because, as expected, it has been chosen a few times by the users and not significative for this comparison.

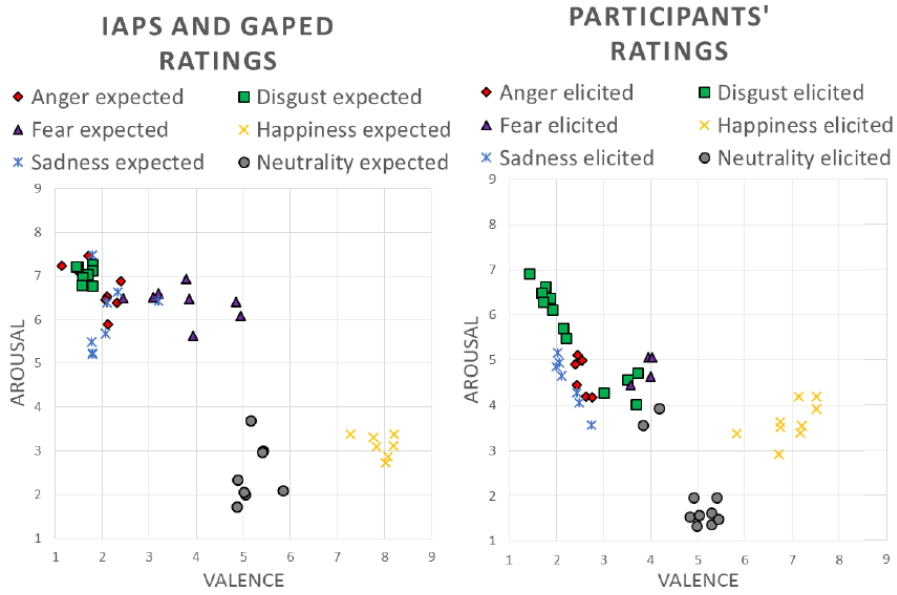

Fig. 10. Valence and arousal values comparisons in affective databases (on the left) and obtained during the experiment (on the right).

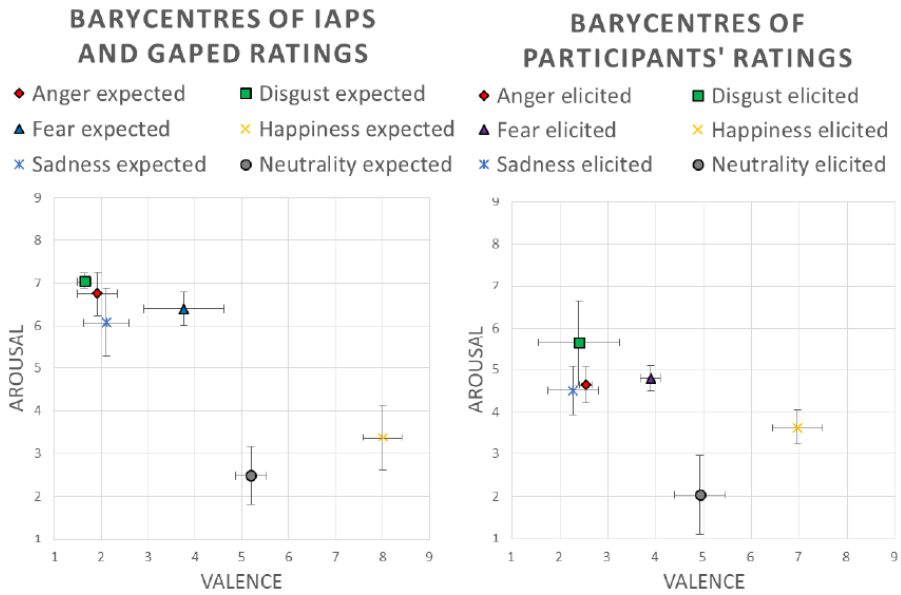

Fig. 11. Valence and arousal barycenter comparisons in affective databases and obtained during the experiments by participants' ratings. Lines represent the standard deviation.

Then, the CNN was run in order to have a preliminary classification of the emotions acquired by the RGB-D camera. An heterogeneous focus group was created to label users' emotions frame by frame. This allowed us to obtain a percentage accuracy.

The use of both the RGB and the 3D channels has returned three classifications for each image. The first one takes into account only the RGB data; the second one only the 3D data. Instead, the third classification results in an average between the two previous classifications. Results considering only depth images have not been satisfactory (55,20%), while RGB and depth's combined usage has led to a 72,65% of agreement. Best results have been obtained considering only RGB images, correctly recognized 3 times out of 4 (75,02%). Some considerations about these results must be made.

The participants' head orientation was not the most suitable one because the RGB-D camera has not been positioned in the center of the field of view of the participants; otherwise, the monitor would not have been visible. The need to keep the monitor size large enough was urgent, to guarantee the user the most immersive experience possible to ensure the spontaneity of facial expressions. Hence, to ensure the ecological validity of the experiment, a compromise with the data visualization has been requested. The result is that faces have been framed with a slight tilt angle relative to the camera; particularly, the CNN trained with depth images labeled many images as angry since areas highlighting anger have been pointed out this way (wrinkles between eyes). This consideration leads to the second one: the dataset for training is too limited, especially for negative emotions. The training set is uniformly split, nonetheless, negative emotions are really close to each other in the affective space. This means that the CNN, as well as most classifiers, needs more images to properly run, especially those images belonging to classes critical to be recognized. Finally, the participants in the experiment, the focus group, and the CNN were asked to evaluate the images assigning only one single emotion. Nonetheless, some images could have led to feeling more than an emotion.

The final recognition rate of 75,02% can be considered satisfactory for the purpose of testing the adoptability of our data for facial expression recognition purposes.

4 Conclusions

The present work has aimed to realize and test an ecologically valid facial expression database, realized by recording spontaneous facial expressions elicited during the designed experiment. Each person has been recorded by an RGB-D camera. Information about her/his emotions has been obtained through a questionnaire and a visual analysis provided by human observers, and an automatic recognition method based on CNN. The results reflect the distribution of the arousal and valence values in Russell and Plutchik's affective spaces taken as reference.

The next step will be the CNN training with an enlarged dataset both for RGB and Depth images. To improve the recognition rate of the basic emotions and increase the range of emotions, further images will be provided by different databases consisting of posed and spontaneous expressions. Nowadays, state-of-art results are obtained using a multimodal approach, a new promising research direction; the building of an

ecologically valid RGB-D dataset that benefits from the accuracy of 2D and the flexibility of 3D data, aims to obtain cutting-edge results in the Facial Expression Recognition field.

Another improvement will be creating immersive virtual reality environments to arouse more pronounced and clear expressions, preserving the dataset's ecologically valid nature, chiefly for those emotions that occupy similar areas in the affective space. Furthermore, the usage of these environments will change the paradigm of the experiment, converting the user experience from passive to proactive.

References

1. Wundt, W.: Outlines of Psychology. In: Rieber, R.W. (a c. di) Wilhelm Wundt and the Making of a Scientific Psychology. pagg. 179–195. Springer US, Boston, MA (1980)
2. Woodworth, R.S., Barber, B., Schlosberg, H.: Experimental psychology. Oxford and IBH Publishing (1954)
3. Ekman, P.: A Methodological Discussion of Nonverbal Behavior. The Journal of Psychology. 43, 141–149 (1957). https://doi.org/10.1080/00223980.1957.9713059
4. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology. 39, 1161 (1980)
5. Schmuckler, M.A.: What is ecological validity? A dimensional analysis. Infancy. 2, 419–436 (2001)
6. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pagg. 3438–3446 (2016)
7. Gross, J.J., Levenson, R.W.: Emotion elicitation using films. Cognition & emotion. 9, 87–108 (1995)
8. Soleymani, M., Pantic, M., Pun, T.: Multimodal emotion recognition in response to videos. IEEE transactions on affective computing. 3, 211–223 (2011)
9. Chanel, G., Rebetez, C., Bétrancourt, M., Pun, T.: Emotion assessment from physiological signals for adaptation of game difficulty. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 41, 1052–1063 (2011)
10. Taskiran, C., Karaismailoglu, S., Cak Esen, H.T., Tuzun, Z., Erdem, A., Balkanci, Z.D., Dolgun, A.B., Cengel Kultur, S.E.: Clinical features and subjective/physiological responses to emotional stimuli in the presence of emotion dysregulation in attention-deficit hyperactivity disorder. Journal of Clinical and Experimental Neuropsychology. 40, 389–404 (2018)
11. Migliore, S., Curcio, G., Porcaro, C., Cottone, C., Simonelli, I., D'aurizio, G., Landi, D., Palmieri, M.G., Ghazaryan, A., Squitieri, F.: Emotional processing in RRMS patients: Dissociation between behavioural and neurophysiological response. Multiple Sclerosis and Related Disorders. 27, 344–349 (2019)
12. Moret-Tatay, C., Rueda, P.M., Bernabé-Valero, G., Gamermann, D.: Emotional Recognition in Schizophrenia: An Analysis of Response Components in Middle-Aged Adults. Psychiatric Quarterly. 90, 543–552 (2019)

13. Bekele, E., Bian, D., Zheng, Z., Peterman, J., Park, S., Sarkar, N.: Responses during facial emotional expression recognition tasks using virtual reality and static iaps pictures for adults with schizophrenia. In: International Conference on Virtual, Augmented and Mixed Reality. pagg. 225–235. Springer (2014)
14. Peter, M., Arntz, A., Klimstra, T.A., Faulborn, M., Vingerhoets, A.: Subjective emotional responses to IAPS pictures in patients with borderline personality disorder, cluster-C personality disorders, and non-patients. Psychiatry research. 273, 712–718 (2019)
15. Pistoia, F., Conson, M., Carolei, A., Dema, M.G., Splendiani, A., Curcio, G., Sacco, S.: Post-earthquake distress and development of emotional expertise in young adults. Frontiers in behavioral neuroscience. 12, 91 (2018)
16. Martínez Navarro, A.M.: Medición de las Respuestas Emocionales en la Violencia contra las Mujeres: Una Revisión Sistemática, (2019)
17. Dominguez-Centeno, I., Jurado-Barba, R., Sion, A., Martínez-Maldonado, A., Castillo-Parra, G., López-Muñoz, F., Rubio, G., Martínez-Gras, I.: Psychophysiological Correlates of Emotional-and Alcohol-Related Cues Processing in Offspring of Alcohol-Dependent Patients. Alcohol and Alcoholism. (2020)
18. Dan-Glauser, E.S., Scherer, K.R.: The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. Behavior research methods. 43, 468 (2011)
19. Mehrabian, A., Russell, J.A.: An approach to environmental psychology. The MIT Press, Cambridge, MA, US (1974)
20. Bradley, M.M., Lang, P.J.: Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. Journal of behavior therapy and experimental psychiatry. 25, 49 (1994)
21. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International Affective Picture System (IAPS)(1997). Technical Manual and Affective Ratings. NIMH Center for the Study of Emotion and Attention. (1997)
22. Mul, C., Stagg, S.D., Herbelin, B., Aspell, J.E.: The feeling of me feeling for you: Interoception, alexithymia and empathy in autism. Journal of Autism and Developmental Disorders. 48, 2953–2967 (2018)
23. Moriguchi, Y., Decety, J., Ohnishi, T., Maeda, M., Mori, T., Nemoto, K., Matsuda, H., Komaki, G.: Empathy and judging other's pain: an fMRI study of alexithymia. Cerebral Cortex. 17, 2223–2234 (2007)
24. Meneghini, A.M., Sartori, R., Cunico, L.: Adattamento italiano della Balanced Emotional Empathy Scale (BEES) di Albert Mehrabian [The Italian adaptation of the Balanced Emotional Empathy Scale (BEES) by Albert Mehrabian]. Florence, Italy: Giunti Organizzazioni Speciali. (2012)
25. Bagby, R.M., Parker, J.D., Taylor, G.J.: The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. Journal of psychosomatic research. 38, 23–32 (1994)
26. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Theories of emotion. pagg. 3–33. Elsevier (1980)
27. Coan, J.A., Allen, J.J.: Handbook of emotion elicitation and assessment. Oxford university press (2007)
28. Lang, P., Bradley, M.M.: The International Affective Picture System (IAPS) in the study of emotion and attention. Handbook of emotion elicitation and assessment. 29, 70–73 (2007)

29. Bradley, M.M., Codispoti, M., Sabatinelli, D., Lang, P.J.: Emotion and motivation II: sex differences in picture processing. Emotion. 1, 300 (2001)
30. Kundu, S.: https://unsplash.com/@senjuti
31. Nonis, F., Dagnes, N., Marcolin, F., Vezzetti, E.: 3D approaches and challenges in facial expression recognition algorithms—A literature review. Applied Sciences. 9, 3904 (2019)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. (2014)
33. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pagg. 248–255. Ieee (2009)
34. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pagg. 211–216. IEEE (2006)
35. Nonis, F., Barbiero, P., Cirrincione, G., Olivetti, E.C., Marcolin, F., Vezzetti, E.: Understanding Abstraction in Deep CNN: An Application on Facial Emotion Recognition. In: Progresses in Artificial Intelligence and Neural Systems. pagg. 281–290. Springer (2020)
36. Edwards, S., Salkovskis, P.M.: An experimental demonstration that fear, but not disgust, is associated with return of fear in phobias. Journal of Anxiety Disorders. 20, 58–71 (2006)