World Academy of Science, Engineering and Technology
International Journal of Mechanical and Mechatronics Engineering
Vol:12, No:6, 2018

Affective Robots: Evaluation of Automatic Emotion Recognition Approaches on a Humanoid Robot towards Emotionally Intelligent Machines

Silvia Santano Guillén, Luigi Lo Iacono, Christian Meder

***Abstract*—One of the main aims of current social robotic research is to improve the robots' abilities to interact with humans. In order to achieve an interaction similar to that among humans, robots should be able to communicate in an intuitive and natural way and appropriately interpret human affects during social interactions. Similarly to how humans are able to recognize emotions in other humans, machines are capable of extracting information from the various ways humans convey emotions—including facial expression, speech, gesture or text—and using this information for improved human computer interaction. This can be described as *Affective Computing*, an interdisciplinary field that expands into otherwise unrelated fields like psychology and cognitive science and involves the research and development of systems that can recognize and interpret human affects. To leverage these emotional capabilities by embedding them in humanoid robots is the foundation of the concept *Affective Robots*, which has the objective of making robots capable of sensing the user's current mood and personality traits and adapt their behavior in the most appropriate manner based on that. In this paper, the emotion recognition capabilities of the humanoid robot Pepper are experimentally explored, based on the facial expressions for the so-called basic emotions, as well as how it performs in contrast to other state-of-the-art approaches with both expression databases compiled in academic environments and real subjects showing posed expressions as well as spontaneous emotional reactions. The experiments' results show that the detection accuracy amongst the evaluated approaches differs substantially. The introduced experiments offer a general structure and approach for conducting such experimental evaluations. The paper further suggests that the most meaningful results are obtained by conducting experiments with real subjects expressing the emotions as spontaneous reactions.

***Keywords*—Affective computing, emotion recognition, humanoid robot, Human-Robot-Interaction (HRI), social robots.

I. Introduction

It is a common assumption that humans prefer an interaction with a machine in a similar way to how they interact with one another instead of having to adapt themselves to machines. At the present time, we are accustomed to continuously interacting with machines, since they are already part of our daily life. It has by some means promptly become completely natural to depend on devices for diverse tasks since early morning to tell us the most appropriate route we should drive to work to avoid traffic as well as to interact with machines like the cash register or the ATM instead of with real humans. The interaction with smart devices spreads across the whole day: it is routine that we are automatically and continuously notified of events such as meetings of the day, delays on our flights, new releases of products we like and reminders for birthdays and even rely on these notifications and reminders so we do not need to focus on remember these pieces of information. Concrete examples of devices designed specifically to carry out such kind of functions are Amazon Echo and Google Home, relatively recently released voice-enabled wireless speakers. These make use of intelligent personal assistants, i.e. software agents capable of understanding user requests expressed in natural language and perform the required actions, Amazon Alexa and Google Assistant, respectively. On the one hand, it is fair to say these make indeed helpful assistants. On the other hand, the kind of interaction we experience is utterly different from how humans intuitively communicate. Benefiting from these services currently requires humans to behave more like machines instead of having machines adapt to our needs. We learned to interpret the impersonal why how devices that handle our most personal information and details about our lives communicate with us: we click on icons on a screen, and understand something requires our attention when we see a phone vibrating or an LED light blinking and mostly could never understand whether the user is happy with the results or contrarily very frustrated. Emerging social humanoid robots, on the contrary —such as personal assistant or companion robots— can also use intelligent assistants and moreover are equipped with numerous sensors that make it possible for them to engage in effective emotional interaction, leading to substantial improvements in their performance in scenarios such as education or elderly care. The rest of this paper is organized as follows: Section II presents the theoretical background in respect to emotions and emotion recognition. Then, the literature concerning algorithms and methods for automatic emotion recognition is reviewed in Section III. As no comparative study on the emotion recognition accuracy of available technologies is present in the literature, Section IV introduces the experiments designed and carried out in terms of this paper. In Section V we discuss the results obtained before concluding in Section VI.

II. Foundations

A. Emotions

Successful real-time emotion recognition plays a fundamental role in human social interactions and is

Silvia Santano Guillén is with the Cologne University of Applied Sciences and inovex GmbH, Karlsruhe, Germany (e-mail: ssantano@inovex.de).

Luigi Lo Iacono is with the Cologne University of Applied Sciences, Germany.

Christian Meder is with the inovex GmbH, Karlsruhe, Germany.

considered a component of the *Emotional Quotient* (EQ) by the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) [1]. The interpretation by humans is a very complex process that involves extensive and diverse areas in the brain. Because of the accuracy of the human visual system, it would have a big impact on the development in computer vision and machine learning if science could completely describe the algorithm used by our visual system, which seems to have identified a set of robust features that facilitate rapid categorization of emotions [2]. A particularly interesting type of facial expression is the so-called micro expressions that last just a fraction of a second. Their fundamental characteristics are that they are involuntary and occur extremely fast. Over the last years these expressions have been gaining more attention because, among other reasons it is thought that their correct detection can help to determine true feelings or emotions and, for example, determine whether someone is telling the truth or not since the person is not able to avoid these micro-expressions even when consciously trying to hide the underlying emotion.

An emotion is considered to be a complex state of feelings that arises in response to stimuli and causes physiological as well as psychological changes. Regarding the arduous task of defining the concept and components of emotion, researchers have not found consensus. Scherer already pointed out "*the question 'What is an emotion?' rarely generates the same answer from different individuals, scientists or laymen alike*" [3]. As stated in [4], emotion theorists generally agree that emotional responses are composed of several, partially independent components: arousal (such as a pounding heart, sweating, blood rushing to the face), expression (the outward sign that an emotion is being experienced e.g. fainting, a flushed face, muscle tensing, facial expressions, tone of voice, rapid breathing, restlessness, and any other body language) and subjective experience (which refers to the way each individual person experiences feelings). Others include also cognitive and neurophysiological components [3]. From these, only the first two can be measured. Automatic emotion recognition—as well as the human ability to discern emotive gestures in others—is in most cases—including the approaches investigated and proposed in this research—based on the second component, the expressive behavior, because of its easily measurable characteristics with current technology.

B. Automatic Emotion Recognition

Although over the last decades researchers have proposed different and often competing models of emotional expression, all theorists agree that all normally-functioning humans express their emotions, mainly with their voices, faces, and bodies. Among these, the position of the muscles of the face are a central organ in the expression of emotion. The facial movements are caused by the movement of muscles that connect to the skin and fascia in the face. When these muscles move, they move the skin and create lines and folds. Most of the facial expression's information can be found in the position of the eyes' muscles as well as the eye contact. In addition, facial expression has great importance in sign languages, where it is used to convey specific meanings. Thus, research in Automatic Emotion Recognition comprises a variety of related fields, such as robotics, computer vision, speech analysis, cognitive psychology, and computational learning theory. The relying input sources may come from different signals, e.g. visual, audio, text and neurophysiological signals or a combination thereof. Emotion Recognition methods based on visual cues rely mostly on facial expression as their source.

A previous step is emotion classification, the means by which emotions can be categorized and distinguished from one another. It is a contested issue in emotion research and affective science, the scientific study of emotion or affect. The classification of emotions has been researched for several decades mainly from two distinct viewpoints: the discrete emotion theory and the dimensional models of emotion theory. Discrete emotion theory maintains that for all human beings there is a set of basic emotions which are innate and expressed and recognized across cultures and which combined produce all others. One of Paul Ekman's most influential studies concludes with the finding that facial expressions can be universally recognized and officially put forth six basic emotions in 1971: anger, fear, disgust, surprise, happiness, and sadness [5]. On the other hand the dimensional models of emotion theory maintain that emotions can be characterized on a dimensional basis in groupings and pursue the conceptualization of human emotions, defining where they are located in two or three dimensions. Most of these dimensional models incorporate dimensions of valence and arousal or intensity.

Each emotion needs to be characterized by models that relate them to actual measurable cues, e.g. muscle positions. The Facial Action Coding System (FACS) [6] published in 2002 tries to systematically categorize the expressions based upon these physical features. It defines Action Units (AU) as units of muscular activity, the combination of which forms expressions. FACS is the most widely used and standardized method for expression analysis in several fields, including computer vision.

III. Related Work

In recent years the field of machine learning has made extraordinary progress in addressing image classification tasks. In particular—based on recent research—the model called deep convolutional neural network seems to be achieving excellent results and is able to match and in some domains even exceed human performance, outperforming traditional methods in visual recognition. The objective of emotion recognition can be tackled in several ways. For the case of recognition from the facial expression, the most common approach is based on the discrete emotion theory: to identify the expression from a set of basic emotions, which falls under the category of classification. In this case, supervised learning commends itself as a natural method and labeled images are provided as training data.

Researchers have demonstrated steady progress, often validating their work against ImageNet, an academic benchmark for computer vision [28]. *Deep convolutional* *neural network* (DCNN) models are being continuously evolved and improved, producing new state-of-the-art outcomes in very short times. Examples of this are the models QuocNet, AlexNet, Inception (GoogLeNet), Inception-v2 and its latest evolution, the Inception-v3.

Neural networks are known for their need of large amounts of data for training and validation of the models. To this extent, large curated image and video databases are compiled in academic environments. These databases should include relevant variation in the data, e.g. different poses and actions, occlusion, changes in the illumination, intensity of the expressions and timing, as well as preferably some individual differences among the subjects. Some well-known and widely used databases for research purposes are: the extended Cohn-Kanade Dataset (CK+) [7], FER-2013 [8] or the Japanese Female Facial Expressions (JAFFE) [9].

Many recent studies such as [11]-[13] submitted to the *2015 Emotions in the Wild* (EmotiW 2015) [10] contest for static images all used DCNNs. In [11], the authors focus on the task of automatically classifying a set of static images from the SFEW 2.0 dataset into 7 basic emotions. The authors use an ensemble of multiple DCNNs pre-trained with FER-2013 and fine-tuned with SFEW 2.0 achieving a 61.29% test accuracy. The research carried out in [13] also exhibited significant improvement in facial emotion recognition using CNNs with a test accuracy up to 54.56% where the authors approached training with a small amount of data and appearance variation usually caused by variations in illumination. The research in [14] presents a different kind of network, a deep belief network with hierarchical face parsing which the authors train with the JAFFE and the CK+ databases. The work in [16] is also based on a DCNN that used the FER-2013 dataset and presented results with an average accuracy of 67% on emotion classification along with capabilities to recognize race, age and gender from images of faces. The research in [17] explored three neural network architectures used in previous studies which are customized and trained. Afterwards, the best performing network—a similar model as in [16]—was further optimized. The results are in comparison satisfactory while using limited resources, obtaining about 63% accuracy on FER-2013 database and 71% with RafD database images. The results also show remarkable difference between the emotions: while it works best for ’happy’ with 90% accuracy, the worst detected is ’sad’ with 28%.

Nevertheless, neural networks are not the only way to address this issue but algorithms such as Support Vector Machine classifiers, Rule-Based systems, Conditional Random Fields and Hidden Markov Models are also among repeatedly used algorithms. For instance, in [15], the authors use Gabor filtering for image processing and the Support Vector Machine (SVM) algorithm for the classification task using images from the CK+ database. The accuracy of the implementation on emotion recognition varies from 88% to 100%, provided that the data is preprocessed such that every image complies to a strict format.

The development and use of humanoid robots and robots able to coexist and collaborate with humans is currently a very popular matter of research. However, the pursuit of this goal is not new but started already several decades ago as it can be seen on [22] and [23], that describe the development of some of the first robots of this kind.

Social Robotics Research that concretely explores the use of automatic emotion recognition in robots can be found e.g. in [19] where the main finding was that two-way interaction, possessing thoughts, feelings and emotions, and being capable of sensing the social environment are the most essential parts of social behavior to pursue for social robots. In [18], the authors investigated how a robot capable of mood detection may be beneficial in relationships. The contribution from [20] is a robotic platform and a vision system based on a Microsoft Kinect Sensor to recognize the emotion of the user and use the provided face points to extract visual features based on the action units of the FACS. A fuzzy classifier is then used to detect the emotion and generate a response on a commercial humanoid robot platform.

IV. Experiments

A. Methodology

To assess how the emotion recognition of humanoid robots perform compared to other state-of-the-art solutions, an experiment has been designed and carried out. The robotic platform used is Pepper [26] (see Fig. 1) from Softbank Robotics, an autonomous humanoid programmable robot. This 1.20m tall humanoid robot was specifically developed to interact with humans, what he does through speech and natural body movements which make the communication intuitive. To express itself better and for support purposes, when speech communication is not possible, a tablet is mounted in front of its breast. With its cameras and a 3D sensor, Pepper is able to detect and recognize faces and the environment. Moreover, its great amount of sensors all around its body allow it to move safely and autonomously: infrared, sonars, laser and bumpers. NaoQi is the Unix-based operative system it runs and its behavior is fully programmable in several languages such as C++ and Python, so that new features can be implemented. It connects to the Internet, which allows to integrate any cloud services to extend its capabilities, as well.

Fig. 1 The Humanoid Robot called Pepper and developed by Softbank Robotics [26]

The first method investigated is therefore Pepper’s ALMood Module [29] (from NaoQi version 2.4) which is part of the

Open Science Index, Mechanical and Mechatronics Engineering Vol:12, No:6, 2018 publications.waset.org/10009027.pdf

standard libraries for Pepper that returns several values about the instantaneous emotion perception in humans. ALMood properties rely upon a combination of the information returned from several sources, which in turn are based on the data captured with Pepper's cameras, microphones and touch sensors. The sources include head angles, facial expressions, semantic analysis from speech, acoustic voice emotion analysis, sound level and energy level of noise, touch sensors and movement detection. Concretely, ALMood's expression cues rely on OMRON's Real-Time Facial Expression Estimation Technology [30], that combines the companys proprietary 3D model-fitting technology and a statistical classification method based on a database of facial images. Facial expression values are returned with the same format as the output of a softmax function, i.e. real values in the range of 0 to 1 normalized so that they add up to 1. The softmax function often appears in the output layer of neural networks dedicated to classification tasks to facilitate the final comparison as it helps to highlight the largest values and conceal those that are significantly below. From the basic emotions it is capable of recognizing the following: neutral, happy, angry, sad and surprised, as well as attention, valence and ambiance state (activity/calm). The valence value, although not evaluated on this experiment, can yield very useful information about the user's emotional reaction over the following 3 second period, for cases where no specific expression needs to be identified but mainly whether the reaction is rather positive, neutral or negative.

The second approach is an implementation of a DCNN programmed making use of the TFLearn library on top of TensorFlow [27]. For the preprocessing, the library OpenCV was used in order to detect the faces in the image using a Cascade Classifier. Afterwards, OpenCV was also used to reshape the image and resize it to the format that the network will take as input, 48x48. The network model employed has been used by several research studies and is based on a slight variation of the Alexnet model as used in [16], [17].

The convolutional network model is composed of a total of 9 layers, all of them containing ReLu (Rectified Linear Units), an element-wise non-linear operation applied to each pixel that replaces by zero the negative values. The input layer is 48x48, as the input data. Afterwards, come a convolutional, a local contrast normalization and a max-pooling layers, followed by two more convolutional layers with a max-pooling layer in between in order to reduce the number of parameters. It finishes with a fully connected layer, to which dropout was applied, and as output layer the soft-max function of size 7, to return the likelihood that each of the following seven emotions: happiness, sadness, anger, surprise, disgust, fear and neutral is present on the main face of the image, the biggest face, as normalized values. The network has been trained with the images from the FER-2013 dataset [8].

The third approach is version 1.0 of Google Cloud Vision [31], a service part of Google's neural network based machine learning platform that provides various artificial intelligence services with pre-trained models. The services can be used through APIs with a JSON REST interface, either by making direct HTTP requests to the server or with client libraries offered in several programming languages. The so-called *annotate* method runs the image detection and annotation for configurable features in one or more images and returns the likelihood for a set of emotions: joy, sorrow, anger, surprise.

The next investigated solution is part of Microsoft's Cognitive Services—formerly known as Project Oxford—a set of APIs, SDKs and services of Microsofts machine learning based features: the version 1.0 of the Emotion REST API [32]. It is capable of recognizing the following emotions: anger, contempt, disgust, fear, happiness, neutral, sadness and surprise present in an image, which can be provided either via a URL or within the API request as data. The response includes an array of all face entries encountered in the image and for each face the associated emotion scores. These values are normalized in the range (0, 1).

The last approach evaluated in this experiment is the version 2.0 of Kairos' Emotion Analysis software [33], which according to the authors' description is capable of detecting positive, negative and neutral sentiments and the following emotions: anger, disgust, fear, joy, sadness and surprise. Moreover, some other features the implementation is capable of detecting from the faces are: attention time, number of glances and blinking. After having uploaded a media file, video or image, the response includes the confidence for the set of emotions as not normalized integer values in the range from 0 to 100.

Unfortunately, the approaches and algorithms behind the considered services are not publicly known.

The performance of each of the solutions has been evaluated by capturing images with the robot's cameras of:

- Emotion-tagged photos from the Cohn-Kanade (CK+) database [7]
- Real subjects

and extracting the predominant emotion detected by each approach.

In both experiments the analyzed emotions were happy, sad, neutral, surprised, and angry since only these basic emotions are recognized by Pepper's software.

B. Experiment Design

The evaluation was divided in three parts, conducted in the same scenario where special attention was paid to maintain the same conditions in every test to fairly and accurately compare the emotion recognition results. The room was prepared for the experiment by covering the sources of daylight and using in every case the same light sources in the same position, two 800 W Arris lamps with tripods. Moreover, the distance from the robot to the face of the participants and the head orientation remained—to the extent possible—stable in every test.

The first part was based on detecting the expressions in images from ”standard” expressions, i.e. images part of a large database of subjects curated in academic environments showing facial expressions with a combination of position of the facial muscles as the experts determined to be standard expressions of emotions across cultures. The second and third part of the evaluation were performed with real subjects, voluntarily taking part in this research. In order to ensure

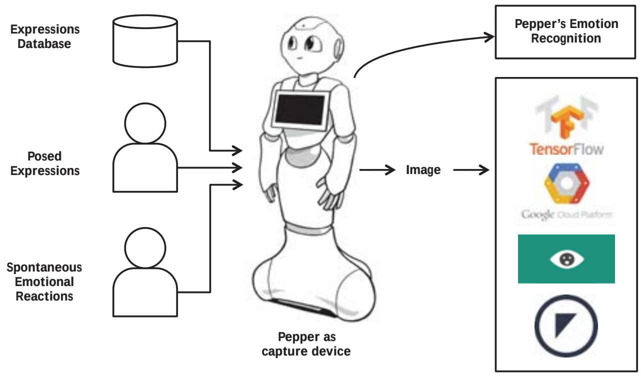
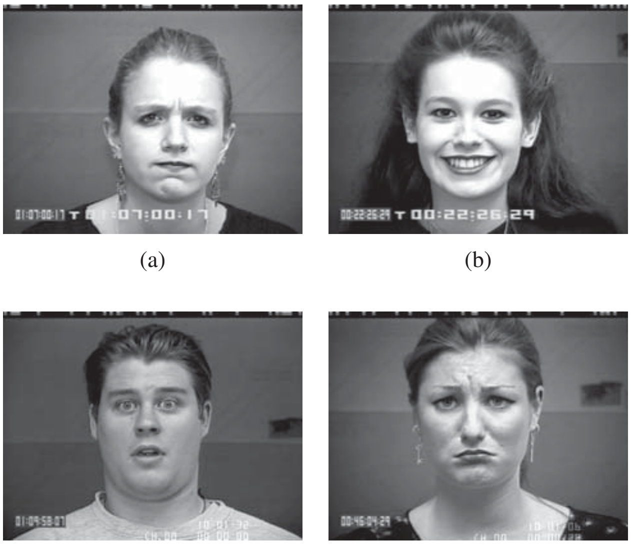


Fig. 2 Method employed for the experiments

some variation in the data, among the $N = 19$ subjects that took part in the experiment there were adults from several different nationalities, n% females and (100-n)% males, in the age range from 25 to 49 (mean=35, stdev=6). Figure 2 depicts the evaluation methodology used for the experiments.

C. Study with Academic Data Set

For emotion recognition several databases are available for research, varying in the quantity of subjects and the quantity of media files as well as technical aspects such as the resolution, the quality and the 'cleanness' of the images. Furthermore, some exhibit posed expressions while others are spontaneous or in the wild. The access is often generally restricted with the exception of research purposes. The choice for this experiment is to use the images from the CK+ database [7] (see Fig. 3), because of the large amount of images and their variety and since it includes both posed and non-posed expressions.



(a) (b)

(c) (d)

Fig. 3 Images extracted from the Cohn-Kanade AU-Coded Expression (CK+) database [7] from which the expression of the subjects was labeled as a) anger, b) happiness, c) surprise, d) sadness

Since its creation, the Cohn-Kanade (CK) database [7] is intended for promoting research on the topic *facial expressions detection*. Since its release in 2000, it has been the choice for numerous studies, both for algorithm development and evaluation, what at present make it one of the most-widely used test-beds. It is for instance the choice in many previous studies, including e.g. [14], [15]. The meta data linked to each image describes the target expression fully FACS coded as well as validated emotion labels. The data set is composed of sequences where each sequence begins with a neutral expression and proceeds to a target expression. The last image from the sequences is the one that will be used for this experiment exhibiting the final expression. Five images per emotion out of the hundreds of images from the database have been selected among those the subjects allowed to use and publish. The selected images have been printed out with the size of a real human head and subjected to the automatic emotion recognition methods at the same position and distance where the participants of the second part would sit.

D. Study with Real Subjects

The experiment was carried out with each of the participants individually, who were asked to sit in front of Pepper for the recognition phase. It was divided in two distinct parts: posed expressions—i.e. in absence of an underlying emotional state—and spontaneous reactions—i.e. congruent with an underlying emotional state. The order in which both are carried out would vary to avoid having them influence one another. For posed expressions, the subjects were simply asked to imitate the expressions corresponding to the basic emotions in front of the robot, whereas for spontaneous emotional responses an emotion elicitation technique was used: emotional film clips, a method commonly used in psychology research. During this part, the participants were given a different task: they were asked to fill a report selecting which emotions were possibly exposed by the characters in each of the videos, explaining that the intention of this part was to evaluate whether the perceived emotion is the same by every individual. This was intended to help distract them from thinking they might be observed, which was explained only once the experiment concluded. The images were taken at a frequency of one per second and the total duration of the videos was about 30 minutes.

It is based on findings from previous research from Ekman, Hager and Friesen [24] and Hess and Kleck [21] that two separate experiments have been conducted. According to them, dynamic aspects of the expression such as the speed at which onset and offset occur and the degree of irregularity of the movements reveal whether the expression was deliberate or spontaneous. The reason why spontaneous and deliberate facial expressions differ relies on the fact that spontaneous and deliberate facial expressions are indeed mediated by different neurological pathways. Aroused emotional reactions are believed to be more like a reflex, smooth and ballistic. On the other hand, when subjects are asked to pose an expression what they do is use the specific view of an appropriate expression that they have in their mind and in this way attempt to exhibit that expression using a closed-control loop approach which disrupts the smooth dynamics of the expression. In the literature, several elicitation procedures were employed for psychological experiments, for instance using emotional film clips. The technique—subjects watch an emotionally evocative video episode—proved to be one of the most popular and effective methods of emotion elicitation. As explained on [21]—compared to other methods—exposure to an emotional film excerpt has several advantages: First, it is one of the easiest techniques to implement in a laboratory. Second, it has been widely observed that film excerpts can elicit strong subjective and physiological changes. Third, the dynamic nature of film scenes provides an optimal artificial model of reality without the ethical and practical problems of real-life techniques. Fourth, it seems to be the most powerful technique to elicit emotion in a laboratory supported by many other previous studies.

Most of the videos chosen for this experiment have been selected from the list of videos used in the experiments introduced in [21], [25].

V. Evaluation

TABLE I
Percentages of Correct Emotion Recognition from a Printout of Images from a Specialized Academic Database

	Pepper	Google	Microsoft	Kairos	DCNN
ANGRY	80	40	0	40	0
HAPPY	100	100	100	0	80
SAD	80	40	80	20	80
SURPRISED	100	100	100	40	100
TOTAL	90	70	70	25	65

Table I shows the accuracy of the algorithms, represented with percentages of correct detections of each algorithm for each emotion with images from the academic dataset as well as the overall accuracy. In general, the performance of most algorithms is very impressive, especially high in images showing a joyful or surprised expression, for which the algorithms from Pepper, Google, Microsoft and the DCNN implementation scored 100%. For other expressions the numbers are lower although still rather satisfactory. The reason why they can recognize these expressions so accurately may very presumably lie in the fact that these expressions are ”pure” according to experts which characterized them as forms of the standard. Another noteworthy fact is that Peppers algorithm always scored the highest or among the highest, with percentages that rate from 80% to 100%, even for expressions such as “anger”, which all others seemingly had the most trouble with.

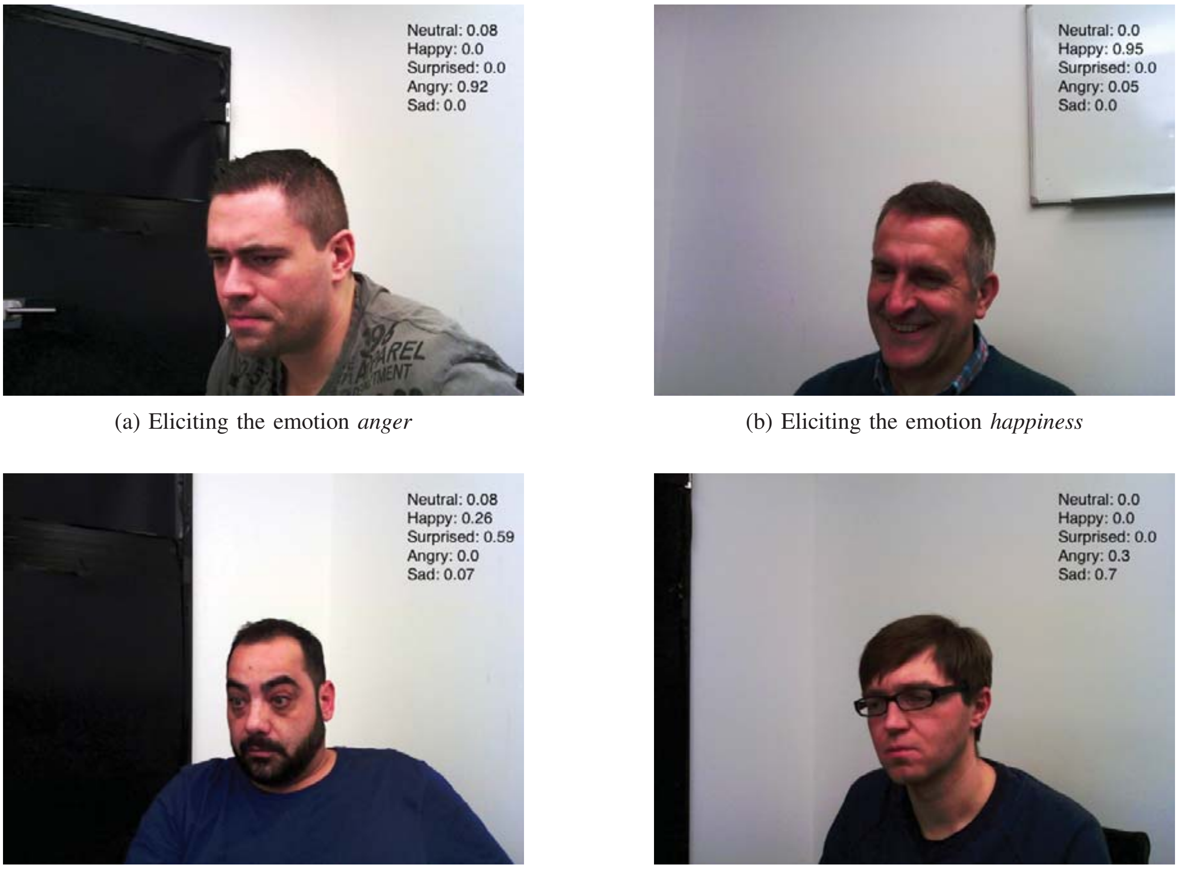
The second part of the experiment, posed expressions from real subjects, included more than two hundred pictures of all 19 participants, two to three pictures per emotion and subject. Table II shows the accuracy of the algorithms during this experiment. Considering the large amount of images collected and analyzed, these rates provide solid information about their performance. Besides, because it is based on diverse subjects and each of them is showing the expressions the way they interpret them without trying to imitate an example, these present differences, showing the variety of characteristics everyone personally expresses. The same pattern as with the database images can be observed in these results: the most easily detected expression is a happy expression with a detection rate that goes up to 100% with Microsoft’s algorithm, which seems to be perfectly well optimized for this emotion. Peppers and Googles algorithms also perform very satisfactorily with a 92.16% and 94.12% of success for happy faces respectively. Anger, sadness and surprise expressions appear to be the hardest to identify, and that the evaluated algorithms frequently mix up one with the other, as emphasized by the rather low correct detection rates of 13.04% for angry expressions in the case of Microsoft’s service or 17.07% of correct detections by Google’s when it comes to looks of surprise. In both cases, the implemented and self-trained DCNN performed a little bit better. Obviously from the results, Kairos’ performance using still images as media input can not reach such rates at all, not even closely. It seems Kairos has been tuned for best performance using video, where characteristics of the individuals can be learned over time, rather than static images. In general, Pepper was able to correctly identify the majority of expressions with an overall rate of success of 63.44% among all expressions, significantly higher than the other tested ones. Even if some of them are remarkably accurate for a certain emotion, the inefficacy with others such as sadness or surprise results in much lower overall rates, which is not the case for the algorithms behind Pepper’s emotion recognition which generally copes well with them. This remarkable difference with the results obtained using database images may very presumably lie in the fact that these expressions obtained in the laboratory with real persons are not completely homogeneous but present at least subtle differences with the descriptions experts made about standard expressions.

TABLE II
Percentages of Correct Emotion Recognition from a Real Subject when Posing a Certain Emotional Expression

	Pepper	Google	Microsoft	Kairos	DCNN
ANGRY	56.52	34.78	13.04	13.04	19.57
HAPPY	92.16	94.12	100	5.88	88.24
NEUTRAL	66.67	84.62	89.74	0.0	82.05
SAD	46.0	32.0	28.0	26.0	26.0
SURPRISED	53.66	17.07	43.9	31.71	21.95
TOTAL	63.44	52.86	54.63	15.42	47.58

For the last part of the experiment, based on spontaneous emotional reactions from the participants only images apparently containing the expression of emotions during the time participants were watching the videos were manually selected from the thousands of pictures taken by the robots cameras and analyzed by the algorithms. Seemingly, the film clips displayed to the participants had the desired effect and aroused emotions in many cases causing them to physically react without even noticing. In spite of the fact that every person reacted differently to the content—as there were two different clips targeting each of the emotions—enough content could be recorded containing spontaneous expressions as intended with a total of 437 images, an average of 23 per subject.

Examples of the images selected are shown in the Figs. 4a-4d with the predominant emotions recognized as angry, happy, surprised and sad expressions respectively as well as the scores from Pepper for each of the emotions.



(a) Eliciting the emotion *anger*

(b) Eliciting the emotion *happiness*

(c) Eliciting the emotion *surprise*

(d) Eliciting the emotion *sadness*

Fig. 4 Scenes captured during the spontaneous emotional reactions experiments where the emotions were elicited by showing emotional video clips. On the upper right corner, the respective scores for each emotion as recognized by Pepper's algorithms

Table III shows the accuracy of the algorithms, represented with percentages of correct detections. The expectations before the experiment were met: "natural" expressions are often more subtle and of shorter duration making them harder to identify. Moreover, more than one could arise simultaneously and even mask one another. This is reflected in the results where the clear tendency is to perform worse for each and every algorithm. Peppers emotion recognition did recognize 59.83%.

TABLE III
PERCENTAGES OF CORRECT EMOTION RECOGNITION FROM A REAL SUBJECT WHEN HAVING A SPONTANEOUS EMOTIONAL REACTION

	Pepper	Google	Microsoft	Kairos	DCNN
ANGRY	65.38	4.81	1.92	2.88	14.42
HAPPY	80.17	85.95	93.39	1.65	58.68
NEUTRAL	56.45	83.87	96.77	1.61	37.10
SAD	35.48	6.45	9.68	4.84	19.35
SURPRISED	49.43	1.15	1.15	10.34	1.15
TOTAL	59.83	38.22	41.42	3.64	28.46

The emotion values are updated in the background in Pepper so that accessing the values does not consume any time. Regarding the performance of the other algorithms, both Google Vision and Microsoft Emotion return their results in under a second including the picture upload so that the complete process usually has a total latency of around 700ms. The inference time of the trained DCNN is around 100 ms when executed on a computer with standard resources. Unfortunately, many of the images taken could not be used even if an expression was shown in the face of the participants because an automatic emotion recognition algorithm would not be able to identify it. From the observation during the experiments most of these cases were due to the following reasons: the subject was touching or involuntarily hiding parts of the face, looking away or from too high an angle with the camera and the face can not be properly detected or is moving or too close, involuntarily preventing the camera from capturing the complete face. Other difficulties faced during the realization of the experiments was that the robot could get distracted by sounds or movements in the environment and look away from the participant. However, these situations are what can be expected when used in real situations out of laboratory conditions.

VI. CONCLUSION

Being able to recognize human emotions is a first step towards emotionally intelligent machines. This feedback can be used to adapt the robot's behavior and thus improve the quality of human-robot interaction. The major contributions of this paper are the analysis of emotion recognition accuracy with a special focus on humanoid robots as well as the introduction of an evaluation methodology. After conducting experiments where the Pepper robot's emotion recognition was evaluated and compared to other solutions, remarkable differences in the recognition of emotions could be noted. Angry, sad and surprised expressions seem to be much harder to identify than happy expressions, which supports the results from previous studies that also suggest not every emotion can be recognized with the same accuracy, being joy the easiest to identify although the reasons for that could not yet be deducted.

A difference in the recognition accuracy can be noted between the three experiments: while the images from the academic database were interpreted correctly in most of the cases, notably more difficult was to identify posed expressions by real subjects and the rates are even lower for the subjects' spontaneous reactions. Presumably, this lies in the fact that the expressions from the database were categorized as standard by experts, while the others were not. However, Pepper's algorithm maintains a compelling accuracy of almost 60% even with these more subtle expressions above all other approaches investigated. It needs to be taken into account that to match the resolution Pepper's algorithm uses the pictures taken had a resolution of 640x480 px which—depending on the position and distance—might in some cases not be sufficient to differentiate an emotion. Presumably, higher resolution images could yield even better results. However, for fair comparison all algorithms were fed with the same image quality.

The fact that the results obtained with the DCNN-based implementation are not that different from the rest of the solutions even if the data used for the training differed from the images captured in the evaluation leads to the assumption that this algorithm could perform really well when trained with data more similar to what it will be later used with.

To improve these results, a possibility for future research is to experiment with a multi-modal approach, i.e. one that relies on other signals in addition to the facial expression. These relate to different aspects of the subject's communication, such as vocal expressions, words, utterances and pauses, or physiological cues like heart rate and skin temperature or gestures. The combination of several sources would result in a more reliable output, decreasing the probability of misinterpreting signals (false positives and true negatives).

References

[1] Mayer JD, Salovey P, Caruso DR, *Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT) users manual*, 2.0. Toronto, Canada: MHS Publishers, 2002.

[2] J. Liu, A. Harris, N. Kanwisher, *Stages of processing in face perception: an meg study*, Nat Neurosci, vol. 5, pp. 910916, 09 2002.

[3] Klaus R. Scherer, *Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT) users manual*, v. 44, 695-729 Social Science Information, 2005.

[4] E. Kennedy-Moore, J. Watson, *Expressing Emotion: Myths, Realities, and Therapeutic Strategies*. Emotions and social behavior, Guilford Press, 1999.

[5] P. Ekman, *Universals and Cultural Differences in Facial Expressions of Emotion*. University of Nebraska Press, 1971.

[6] P. Ekman, W. V. Friesen, and J. C. Hager, *The facial action coding system*, in *Research Nexus eBook*, 2002.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews, *The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression*, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2010, San Francisco, CA, USA, 13-18 June, 2010*, pp. 94–101, 2010.

[8] *Challenges in representation learning: Facial expression recognition challenge*, https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge (Last accessed: in April 2018)

[9] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, *Coding facial expressions with gabor wavelets* in *3rd International Conference on Face & Gesture Recognition (FG'98), Nara, Japan*, pp. 200–205, 1998.

[10] *The Third Emotion Recognition in The Wild (EmotiW) 2015 Grand Challenge*, http://cs.anu.edu.au/few/emotiw2015.html (Last accessed: April 2018)

[11] Z. Yu and C. Zhang, *Image based static facial expression recognition with multiple deep network learning* in *ICMI' 15 Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, Seattle, WA, USA, pp. 435–442, 2015.

[12] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, *Hierarchical committee of deep convolutional neural networks for robust facial expression recognition* in *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, 2016.

[13] G. Levi and T. Hassner, *Emotion recognition in the wild via convolutional neural networks and mapped binary patterns* in *ICMI' 15 Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, Seattle, WA, USA, pp. 503–510, 2015.

[14] Y. Lv, Z. Feng, and C. Xu, *Facial expression recognition via deep learning*, in *SMARTCOMP*, IEEE Computer Society, pp. 303–308, 2014.

[15] T. Ahsan, T. Jabid, and U.-P. Chong, *Facial expression recognition using local transitional pattern on gabor filtered facial images*, *IETE Technical Review*, vol. 30, no. 1, pp. 47–52, 2013.

[16] A. Gudi, *Recognizing semantic features in faces using deep learning*, *CoRR*, vol. abs/1512.00743, 2015.

[17] E. Correa, A. Jonker, M. Ozo, and R. Stolk, *Emotion recognition using deep convolutional neural networks.*, 2016.

[18] M. Hashemian, H. Moradi, and M. S. Mirian, *How is his/her mood: A question that a companion robot may be able to answer*, in *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings* (A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, eds.), pp. 274–284, Springer International Publishing, 2016.

[19] M. M. A. de Graaf, S. Ben Allouch, and J. A. G. M. van Dijk, *What makes robots social?: A user's perspective on characteristics for social human-robot interaction*, in Proceedings of *Social Robotics: 7th International Conference, ICSR 2015*, Paris, France, pp. 184–193, Springer International Publishing, 2015.

[20] A. Meghdari, M. Alemi, A. G. Pour, and A. Taheri, *Spontaneous human-robot emotional interaction through facial expressions*, in *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings* (A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, eds.), (Cham), pp. 351–361, Springer International Publishing, 2016.

[21] U. Hess and R. E. Kleck, *Differentiating emotion elicited and deliberate emotional facial expressions*, *European Journal of Social Psychology*, vol. 20, no. 5, pp. 369–385, 1990.

[22] M. Hirose, T. Takenaka, H. Gomi and N. Ozawa, *Humanoid robot*, *Journal of the Robotics Society of Japan*, vol. 15, no. 7, pp. 983–985, 1997.

[23] K. Hirai, M. Hirose, Y. Haikawa and T. Takenaka, *The Honda humanoid robot: development and future perspective*, *Industrial Robot: An International Journal*, vol. 26, no. 4, pp. 260–266, 1999.

[24] P. Ekman, J.C. Hager, W.V. Friesen, *The symmetry of emotional and deliberate facial actions*, Psychophysiology, 18: 101-106, 1981.

[25] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, *Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers*, *Cognition and Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[26] Aldebaran (Softbank Robotics), *Pepper robot*, https://www.ald.softbankrobotics.com/en/robots/pepper (Last accessed: April 2018)

[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, *Tensorflow: A system for large-scale machine learning* in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, 2016.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database* in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.

[29] Softbank Robotics, *ALMood Module*, http://doc.aldebaran.com/2-4/naoqi/core/almood.html (Last accessed: April 2018)

[30] OMRON Corporation, *Facial Expression Estimation Technology*, https://www.omron.com/media/press/2012/10/e1023.html (Last accessed: April 2018)

[31] Google Inc., *Cloud Vision API*, https://cloud.google.com/vision/ (Last accessed: April 2018)

[32] Microsoft Corporation, *Emotion API*, https://azure.microsoft.com/en-us/services/cognitive-services/emotion/ (Last accessed: April 2018)

[33] Kairos AR, Inc.,*Human Analytics*, https://www.kairos.com/features (Last accessed: April 2018)