Neural Computing and Applications (2018) 29:359–373
https://doi.org/10.1007/s00521-018-3358-8

S.I. : EANN 2016

A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots

Ariel Ruiz-Garcia[1] · Mark Elshaw[1] · Abdulrahman Altahhan[1] · Vasile Palade[1]

Received: 19 December 2016 / Accepted: 4 October 2017 / Published online: 8 February 2018

Abstract
We have recently seen significant advancements in the development of robotic machines that are designed to assist people with their daily lives. Socially assistive robots are now able to perform a number of tasks autonomously and without human supervision. However, if these robots are to be accepted by human users, there is a need to focus on the form of human–robot interaction that is seen as acceptable by such users. In this paper, we extend our previous work, originally presented in Ruiz-Garcia et al. (in: Engineering applications of neural networks: 17th international conference, EANN 2016, Aberdeen, UK, September 2–5, 2016, proceedings, pp 79–93, 2016. https://doi.org/10.1007/978-3-319-44188-7_6), to provide emotion recognition from human facial expressions for application on a real-time robot. We expand on previous work by presenting a new hybrid deep learning emotion recognition model and preliminary results using this model on real-time emotion recognition performed by our humanoid robot. The hybrid emotion recognition model combines a Deep Convolutional Neural Network (CNN) for self-learnt feature extraction and a Support Vector Machine (SVM) for emotion classification. Compared to more complex approaches that use more layers in the convolutional model, this hybrid deep learning model produces state-of-the-art classification rate of 96.26%, when tested on the Karolinska Directed Emotional Faces dataset (Lundqvist et al. in The Karolinska Directed Emotional Faces—KDEF, 1998), and offers similar performance on unseen data when tested on the Extended Cohn–Kanade dataset (Lucey et al. in: Proceedings of the third international workshop on CVPR for human communicative behaviour analysis (CVPR4HB 2010), San Francisco, USA, pp 94–101, 2010). This architecture also takes advantage of batch normalisation (Ioffe and Szegedy in Batch normalization: accelerating deep network training by reducing internal covariate shift. http://arxiv.org/abs/1502.03167, 2015) for fast learning from a smaller number of training samples. A comparison between Gabor filters and CNN for feature extraction, and between SVM and multilayer perceptron for classification is also provided.

Keywords Deep Convolutional Neural Networks · Emotion recognition · Gabor filter · Socially assistive robots · Support Vector Machine

1 Introduction

Roboticists have long anticipated and foreseen the arrival of intelligent machines capable of illustrating human behaviours without human assistance. Social robots are now able to perform a number of tasks autonomously, which led to a steady introduction of social robots into society. The kinds of social robot applications that are becoming available include those that offer therapy to children with autism [1], exercise coaches [2] and social robots that aid elder people with dementia [3, 4]. However, these machines are still unable to effectively interact with users in a human-like manner, which is an issue that has

✉ Ariel Ruiz-Garcia
ariel.ruiz-garcia@coventry.ac.uk

Mark Elshaw
mark.elshaw@coventry.ac.uk

Abdulrahman Altahhan
abdulrahman.altahhan@coventry.ac.uk

Vasile Palade
vasile.palade@coventry.ac.uk

[1] School of Computing, Electronics and Mathematics, Faculty of Engineering, Environment and Computing, Coventry University, Coventry, UK

proved to be difficult to overcome. Similar to the way human–human interaction (HHI) can be inhibited by the lack of initiative from one of the parties, human–robot interaction (HRI) can fail if there is limited or no engagement by the robot or the human user. Though, in order for the robot to take initiative, it has to be able to determine what actions to execute and to what degree. Moreover, it has to adapt its behaviours according to the user's responses. In this work, we address the first stage of an automated empathic behaviour system, which is the recognition of human emotions. We extend the work published in [5] by introducing a new hybrid architecture that combines a Deep Convolutional Neural Network, for self-learnt feature extraction and representation, with a Support Vector Machine for emotion recognition. This hybrid architecture offers novelty over similar approaches in that it has a simplified configuration, less hyperparameters, and is relatively faster to train, due to the use of batch normalisation (BN) [6] and its reduced number of layers, compared to the current state-of-the-art architectures employing CNNs. This architecture is also able to learn from smaller amounts of data than many of the similar state-of-the-art approaches.

As presented in [5], Gabor filters are common image pre-processing methods in the field of emotion recognition due to their ability to extract salient features. In the case of facial expression images, these are able to highlight areas around the mouth, eyes and eyebrows, all of which play an important role in the recognition of emotions. Considering the popularity and success of CNNs in image classification tasks, this work compares CNN to Gabor filters in terms of feature extraction for emotion recognition. However, since traditional CNN models are constrained by the efficiency of their MLP component to classify the features extracted by the convolutional layers, in this work we look at an alternative to the classifier component in an attempt to increase classification performance. In the case of Gabor filters, it was shown in [5] that SVM outperforms MLPs when classifying the feature vector produced by a bank of Gabor filters. Therefore, this work explores SVM as an alternative to the MLP component of traditional CNNs.

In addition, we explore the impact of moving the testing of the novel hybrid architecture from a controlled manner using images from available corpora, to a less controlled context by applying the model on a robot in a real-world environment. Hence, we present the initial results on real-time emotion recognition performed by our NAO robot <https://www.aldebaran.com/en/cool-robots/nao> using this new hybrid emotion recognition model. Such a real robot test will identify the performance of the emotion recognition model in a real-world scenario and will identify some of the factors that will need to be addressed to achieve levels of emotion recognition performance that will be required by human users.

The following section of this paper presents some background and literature review on social robots, and the current state of the art of computational emotion recognition from facial expressions. Section 3 presents our improvements to the emotion recognition models originally presented in [5] and the new hybrid model. Section 4 presents the results obtained with the emotion recognition models on the Karolinska Directed Emotional Faces (KDEF) dataset [7] and the results of the best performing model from Sect. 3 when tested on the Extended Cohn–Kanade (CK+) dataset [8]. Section 5 describes the preliminary emotion recognition experiments carried out with a NAO robot. Section 6 presents the conclusions of the work presented in this paper along with the future direction of this research. Section 6 also highlights the importance of training robots with realistic data obtained in out-of-the-lab environments. This is followed by a list of references.

2 Background and literature review

The introduction of social robots into society is inevitable, and the transition into a lifestyle that constitutes interacting with robots on a daily basis can be difficult if these machines do not fulfil user's interactive expectations. Therefore, social robots need to be equipped with the intelligence and skills necessary to build social, and to some extent intimate, relationships with end users. Consequently, our work encompasses the development of artificial mechanisms intended to endow social robots with the skills required to succeed in building social relations with human users. The capacity to build this social engagement between the robot and the human user will be particularly significant when considering scenarios where the interactions will be long term, for instance when the robot is a companion and a helper with the goal of ensuring that the senior citizen can remain in their own home as long as possible.

In this work, we make progress towards the development of a robot that is able to recognise emotions through the users' facial expressions. Although in this paper we focus on the goal of developing a model that is able to perform emotion recognition from facial expression, this is only the first step of our overall goal to create an empathic robot. Such a robot will not only recognise emotions, but also automatically and autonomously produce and associate responses to specific emotional states. We begin by targeting the most essential characteristic of an empathic robot: the ability to recognise human emotions. This section studies existing advancements in the field of social robotics and existing machine learning approaches to perform emotion recognition.

2.1 Socially assistive robots

Although the concept of autonomous robots that can learn was introduced decades ago, we are just now starting to witness their introduction into society. The technological limitations that were once the main obstacle of intelligent machines are no longer an impediment. The growing trend for the use of social assistive robots can be attributed in part to the numerous benefits that these machines can offer to individuals and society. For example, scientists in Japan are focusing their efforts in developing socially assistive robots that can serve as companions for the elderly. Moreover, studies show that elderly citizens in Japan prefer in certain cases to obtain assistance from robotic caretakers rather than their human counter-parts [9]. Although this can be interpreted as a worry for some, in a place where the population is in decline and ageing rapidly, citizens can benefit from social robots by employing them as companions and caretakers for the elderly.

Other uses in the health and social care sector include the use of social robots to assist care receivers and caregivers. For example, social robots can be used to help with the coaching of care receivers to eat healthy and take exercise [10]. Fasola and Mataric [2] have created a robot that encourages senior citizens to take exercise by recognising their arm movements, coaching behaviours that plan the exercise to be performances and spoken interactions with the participants. However, this robot fails to consider the mood of the user that could offer a great deal of information on why, for example, the user is not exercising to their full potential.

Social robots can also be employed by the wider spectrum domains in society. In the educational sector, social robots can be used as personal tutors and teachers and can assist in getting students to engage and learn in a proactive way. Castellano et al. [11] have identified a number of crucial points for the success of empathic robotic tutors, including the need for mechanisms for creating social bonds even if not all the features of the robot are anthropomorphic. In fact, existing studies show that social robots can increase children's interest in engineering, increase engagement in learning experiences and improve language skills development [12]. Kory Westlund et al. [13] have used a socially assistive robot called Tega to help children learn a second language. In the learning session, Tega and the child learn Spanish from a virtual teacher agent in the form of a Toucan. This approach does offer emotion recognition from faces by using the Affectiva [14] system and emotion production using the behaviour and speech of the robot. Nevertheless, these emotions rely on valance to improve the engagement of the child, and so the emotions considered are limited to positive, neutral and negative, which could mean that certain subtle, but significant emotion shifts might be lost.

Studies have shown that social robots can be a useful resource for social skills and communication therapies for children with autism [15]. Rabbitt et al. [16] argue that parents view social robots as a very acceptable form of treatment for children with disruptive behaviour problems. As a result of the positive feedback obtained from existing case studies, researchers continue to develop frameworks designed for robots to engage in social interactions with potential users. The KSERA project is an example of this, and it is specially designed to assist the elderly with conditions such as chronic obstructive pulmonary disease [17]. The KSERA robot, based on a combination of neural network learning approaches, offers activities to aid the elderly such as person recognition, speech recognition and production and navigation. Similarly, GeriJoy, a virtual care companion, offers wellness coaching, therapeutic programs, reminders, safety supervision, companionship and care for the elderly [18]. However, unlike the other approaches considered in this subsection, this system does not rely on machine learning, but is controlled by a human operator.

As social robots start to become more commonplace, the need to provide these machines with the necessary skills to build long-term social relations becomes more evident. Leite et al. [19] suggest that empathy supports the creation and development of social bonds between people and that people respond better to robots whom behave emphatically towards them. According to de Graaf et al. [20], people interact with social robots in the same way that humans interact with each other. This implies that robots need to adequately illustrate signs of intelligence and self-awareness to some extent in order to develop long-term relationships with users. Nonetheless, in order to develop an empathic robot we need to first address the focus of the research in this paper, the development of a model for emotion recognition that can perform in a real-world environment. The next section reviews current advancements in the field of computational intelligence, existing state-of-the-art emotion recognition models.

2.2 Machine learning approaches for emotional face recognition

As stated above, to make progress towards an empathic robot we need to address the first step which is emotion recognition. Hence, we will here explore the current state of the art in emotion recognition.

According to Duffy et al. [21], building machines that can be as intelligent and versatile as humans, and with the ability to socialise and interact as if they were humans themselves, requires employing the human frame of reference to a certain extent [21]. Artificial neural networks (ANN) are computational methods intended to model, to some degree, the way the human brain works and can be used to classify facial expression images as a given emotion [22, 23]. Variants of this method have also succeeded in classifying facial expression images, e.g. self-organising maps [24, 25], Support Vector Machines (SVM) [26, 27], convolutional neural networks (CNN) [28], amongst others. In this work, we first discuss and compare the performance of SVM and multilayer perceptron (MLP) neural networks to classify facial expression images obtained from the KDEF dataset [7] as the following emotions: sad, surprised, neutral, happy, fear, disgust and angry. These models rely on hard-coded feature selection using Gabor filters. In what is the main claim of this paper, we then explored an alternative approach that considered the above by replacing the Gabor filters with an approach that offers more self-learnt feature selection by using a Deep CNN.

Considering that the performance of classifier algorithms heavily relies on the quality of the feature vector representing the image and thus the emotional state, it is essential that the optimum image pre-processing method is applied to the images used for training. Gabor filter is one of the most popular methods in image processing due to its ability to detect edges. This process resembles the perception in the human visual system [29] and is characterised by multi-resolution and multi-orientation properties. Chelali and Djeradi [30] have proposed an approach which relies on the magnitude vector produced by Gabor filter. The authors propose applying discrete wavelet transform (DWT) to images as a pre-processing step before being classified by MLP and radial basis function networks. Chelali and Djeradi [30] obtain a peak performance of 95% accuracy on the Computer Vision database and 85% on the ORL database. Mehta and Jadhav [31] use a combination of Log–Gabor filters, PCA, and Euclidean distances on the JAFFE <http://www.kasrl.org/jaffe.html> dataset and obtain a performance of 93.57%. Ahsan et al. [29] also use Gabor filters and combine them with local transitional patterns (LTP): the authors apply Gabor wavelet filter on images and then obtain LTP codes by comparing transition of intensity change at different levels of neighbouring pixels in different directions. The resulting feature vector is then classified with a SVM, which produces an average accuracy rate of 95% on the Cohn–Kanade (CK) [32] database.

Facial expression classification using SVM has also been done by Sohail and Bhattacharya [27]. The authors proposed a method which includes identifying 15 different feature points and measuring the Euclidean distances between these and the feature points representation of a neutral face. The authors employ a SVM for classification and obtain an average recognition rate of 92% on the JAFFE dataset and 86.33% on the CK [32] dataset. These facial expression classification results using SVMs as a classifier positively illustrate the strengths of SVMs for emotion recognition. SVMs have also been employed for other classification tasks such as face recognition problems [33, 34] and face formalisation [35].

MLPs have also proven to be efficient for facial expression classification. Hewahi and Baraka [36] designed a MLP which makes use of ethnic background information to produce an accuracy rate of 83.3%. When no ethnic background information is used, the authors obtain a performance rate of 75% with the same model on the MSDEF dataset. Khashman [37] obtain 87.78% accuracy rate when using global pattern averaging as an image pre-processing step and a MLP for facial expression classification.

Due to the success of CNNs in image classification problems, CNNs are becoming of interest in facial expression classification problems. Ouellet [28] uses a CNN with five convolutional layers to obtain a feature vector and then feed it to a SVM for classification. The author obtains 94.4% accuracy rate on the Extended Cohn–Kanade Dataset. Similarly, Burkert et al. [38] have set a benchmark with 99.6% accuracy using a Deep CNN. The authors use a CNN with seven convolutional layers and test it on the CK dataset using a tenfold cross-validation method.

Although the models discussed in this section offer a good degree of accuracy, it is unclear what features determine classification performance. According to Beaudry et al. [39], the eyes and eyebrows play a bigger role for the recognition of sadness and the mouth is more influential to recognise happiness. In contrast, the authors determined that a holistic processing could be called upon fear, but could not determine the best approach to recognise other emotions. Identifying what facial features determine specific emotions could be of great value if applied to emotion recognition models. In this work, we try to emphasise on areas of interest such as the mouth, eyes and eyebrows to classify facial expression images. As a result, in this paper, we explore Gabor filters as an image pre-processing technique to highlight these areas within an image. Moreover, we use CNN in our new architecture to extract features and compare them against those obtained with Gabor filters. The following section describes our feature extraction methods and classification models.

3 Methodology

This section presents the emotion recognition approach originally published in [5] and introduces an architecture that employs a CNN for automatic feature extraction and representation. The latter architecture is less complex to train, needs less data to learn and achieves results at a similar level of performance compared to more complex and larger architectures. A visualisation of the filters learnt by the CNN and those obtained with the Gabor filters is provided. Both models use SVMs and MLPs as classifiers.

3.1 Emotional facial expression corpus

In this work, we train and test our emotion recognition models on the Karolinska Directed Emotional Faces [7] database. The corpus contains a set with 70 individuals: 35 males and 35 females aged between 20 and 30 years, each displaying seven different emotional expressions in five different angles. All images were taken under a controlled environment, and faces were centred with a grid by positioning eyes and mouths in fixed image coordinates [7].

During our experiments, we only use front angle images, a subset containing 140 front angle images for each one of the seven emotions. In order to obtain a feature vector, we located the face and cropped irrelevant spatial features such as background. Face images were grayscaled and resized to a standard 120 × 110 in the case of Gabor filter and 100 × 100 in the case of CNN to speed up the training. Figure 1 illustrates sample face images obtained from the KDEF database.

We also test the CNN+SVM model on unseen data using the Extended Cohn–Kanade dataset [8], a popular dataset in the field of emotion recognition and face detection. This corpus includes a set of 327 sequences from 118 participants. In this work, we only use the peak frame of each sequence since it contains the most emotion-related information. The same pre-processing applied on the KDEF dataset is applied to this dataset for consistency, as shown in Fig. 2. Testing on the CK+ dataset allows us to compare our hybrid model to the work of [28] who also uses a combination of CNN+SVM.

3.2 Gabor filter

Gabor filters are powerful image processing algorithms that resemble the perception in the human visual system [29] and facilitate edge detection on images. Facial expression classification heavily relies on the shape of facial features such as the mouth, eyes and eyebrows [39], and Gabor filters can be used to emphasise these areas. Therefore, for our first experiment we convolved our dataset with a bank of Gabor filters to obtain image representations that highlight these areas of interest. Each Gabor filter used is essentially a sinusoidal modulated by a Gaussian kernel function [30] in which orthogonal directions are represented by real components.

In [5], we used a bank of 40 filters with five scales and eight orientations, split the resulting feature vector into four and treated each sub-vector as an input sample in order to reduce overfitting. However, because the previous approach quadrupled the amount of data, in this work we only use a bank of ten filters expanding over eight orientations and two dimensions. This is done to provide a fair comparison against the model using a CNN as a replacement for the bank of Gabor filters and using the same number of training and testing samples. Moreover, we did not observe any improvements in classification performance when using more filters, and using a smaller number of filters allowed us to craft each one to highlight areas of interest such as the mouth, eyes and eyebrows.

The Gabor filters applied also down sample the original image over a scale of four, producing a reduced feature vector. Using a feature vector with reduced dimensionality allows for faster training of the SVM model and also produces similar results to full scale vectors. Once the Gabor filters were applied, the feature vector values were normalised in the range zero to one. Moreover, as done by Chelali and Djeradi [30], we only used the magnitude information given that it contains the most relevant information and discards the effect of noise.

We tried using a combination of real and imaginary components as a complex component; however, we obtained lower classification results compared to using only the real component. Let λ represent the frequency of the sinusoidal, θn represent the orientation and σ represent the standard deviation of the Gaussian over x and y dimensions of the sinusoidal plane; our Gabor filter applied to an image with dimensions x and y is defined as follows:

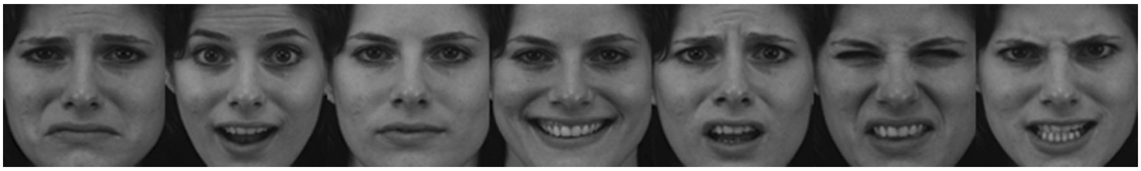

Fig. 1 Subject F07 from the KDEF [7] dataset, displaying seven emotions: sad, surprised, neutral, happy, fear, disgust and angry

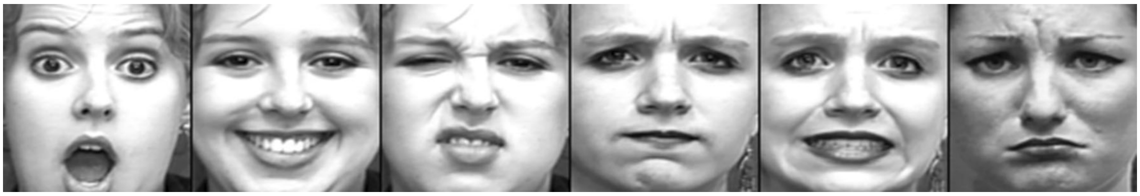

Fig. 2 Left to right, subjects S52, S55 and S116 from the CK+ [8] dataset, displaying six emotions: surprise, happy, disgust, angry, fear and sad. Note that no participant in the dataset provided consent for *contempt* images to be distributed

$$g_{\lambda,\theta}(x,y) = \exp\left[-\frac{1}{2}\left\{\frac{x^2_{\theta n}}{\sigma^2_x} + \frac{y^2_{\theta n}}{\sigma^2_y}\right\}\right]\cos(2\pi * \theta n * \lambda)$$

where

$$x_{\theta n} = x(\sin\theta n) + y(\cos\theta n)$$
$$y_{\theta n} = x(\cos\theta n) + y(\sin\theta n) \tag{1}$$

After trying a number of parameters, we concluded that initialising the Gabor filter with the following parameters produces the best magnitude response vector for emotion classification: $\theta = 2pi/3$, $\lambda = 6$, $\gamma = 0.5$ and $\sigma = 4$. This response vector is given by:

$$||g_{\lambda,\theta}(x,y)|| = \sqrt{\Re^2\{g_{\lambda,\theta}(x,y)\} + \Im^2\{g_{\lambda,\theta}(x,y)\}} \tag{2}$$

where $\Re$ denotes the real component of the filter and $\Im$ the imaginary one.

3.3 Facial expression classification using SVM and MLP

Once the feature vector was obtained, we feed it to a SVM and a MLP to be classified as one of the seven emotions: angry, disgust, fear, happy, neutral, sad and surprised. SVMs are non-probabilistic binary classifiers known for performing notably well in image classification problems. We tested the one-vs-one and one-vs-all approaches for multi-class classification and obtained better performance with the one-vs-one approach using a linear kernels. Let b represent the bias, K be a linear kernel function, and our facial expression classification is determined by:

$$f(x) = \operatorname{sgn}\left(\sum_{i+l}^{l} y_i a_i | K(x_i, x) + b\right) \tag{3}$$

where x_i is the training vector, x is the testing vector with $a_i > 0$ and y_i represents Lagrange multipliers of dual optimisation problem [40]. This model produced an accuracy rate of 95.58% on the testing set after training using a c value of 1000 for the SVM. We also tested on larger and smaller c values, but obtained better results with 1000.

Taking into account the popularity of traditional MLPs for classification problems, we decided to compare the performance of the SVM against that of a MLP network. After considering a number of different network topologies, we obtained best results with the following MLP network configuration: one input layer with 8400 neurons taking ten 28 × 30 filtered images as input, one hidden layer with 93 neurons and one output layer with 7 neurons. The target values contained a one in the place of the target class and zero for the rest. The MLP uses a sigmoid activation function and was trained using resilient backpropagation (Rprop) to reduce the issues of falling into a local minimum caused by sigmoid activation functions. Figure 3 provides an illustration of this model. The Gabor+SVM approach is the same as shown in Fig. 3, but we replace the MLP in the figure with a SVM.

The initial weights of the MLP were randomly initialised and this model achieved its best performance after training for 175 epochs, with 100% per cent accuracy on the training set and 93.5% on the testing set. Learning rate was set to 0.0001 and remained constant during training. Both the SVM and MLP models are trained on 70% of the KDEF dataset and tested on the remaining 30%. Note that since all the classes contain the same number of samples, 140, 70% of the images from each class are randomly selected for training and the remaining 30% are used for testing.

3.4 Convolutional neural networks

Although Gabor filters offer good image representations and allow for the detection of edges and, thus, facial features that are important for emotion recognition, Gabor filters need to be carefully designed to produce such representations. In contrast, convolutional neural networks offer an alternative to prescribed methods and possess the ability to learn to extract features necessary for emotion recognition automatically. CNNs are a special type of neural networks that take advantage of spatial information in images. Similarly to Gabor filters, CNNs are biologically inspired and resemble the animal vision cortex [24]. Moreover, the convolutional layers in a CNN provide some degree of shift and deformation invariance [24]. The filters

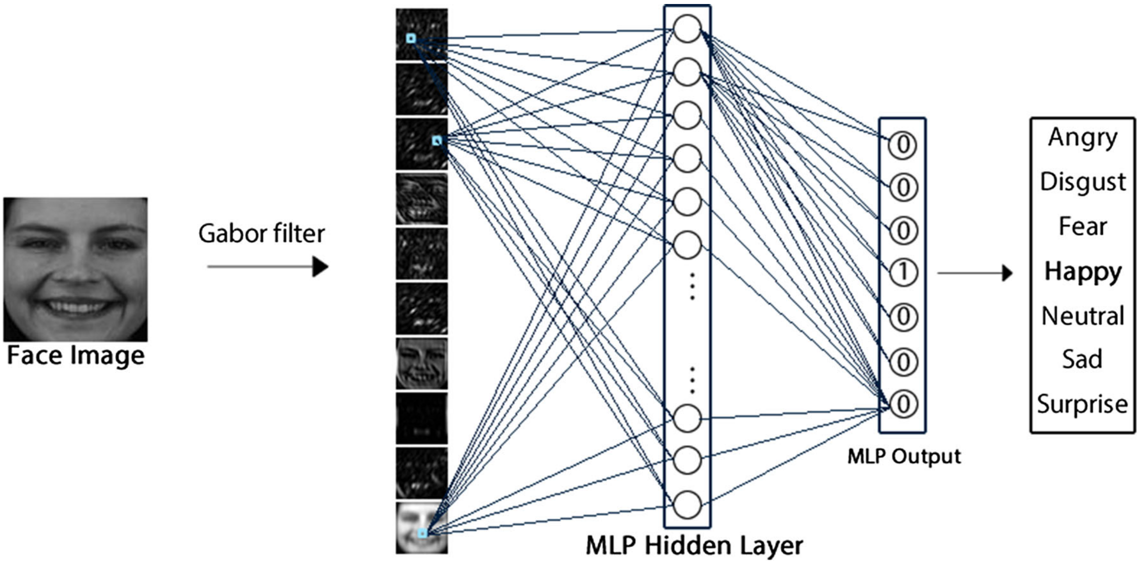

Fig. 3 Multilayer perceptron emotion recognition model taking an image convolved with ten Gabor filters as input. Hidden layer with 93 neurons and output layer with seven. Face image extracted from the KDEF dataset [7]

learnt by convolutional layers are similar to those produced by a bank of Gabor filters as illustrated in Fig. 5.

Our CNN emotion recognition model is composed of blocks of convolutional, ReLU, MaxPooling and BN layers [6]. We use BN since it allows for larger learning rates and faster convergence by normalising the distribution of each input feature at every layer [6]. BN has been shown to speed up convergence by reducing internal covariate shift, that is the change in distribution of network activations caused by change in the network's parameters during training, by using the mean and variance of each minibatch to normalise activations [6]. In the case of convolutional layers, this normalisation is done for each individual feature map.

The first two convolutional layers have filters of size 5 × 5, and the last two layers have filters of size 3 × 3. All convolutional layers have a sliding window of size one and zero padding of size two. Max pooling is done with a stride of size two, kernels of size 2 × 2 and zero padding of size one. Using larger filters seemed unnecessary given that the size of the images is only 100 × 100 and bigger filters often misrelevant information. Moreover, we did not observe any improvements in classification performance using more layers, while reducing the number of layers decreases performance by three per cent on average.

The output of the last block is connected to a fully connected layer, which in effect is a MLP with 100 neurons and a ReLU activation function. We also tried using a larger MLP, but saw no significant improvements. The output of this layer is also normalised with a BN layer before being classified as one of the seven classes. Figure 4 illustrates a visualisation of this model when using either a MLP and or a SVM as the classifier.

The output of convolutional layers in our model can be summarised as:

$$C(x_{u,v}) = (x+a)^n = \sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{j=-\frac{m}{2}}^{\frac{m}{2}} f_k(i,j) x_{u-i,v-j} \tag{4}$$

where f_k is the filter with a kernel size $n \times m$, applied to the input x. In our models, n is always the same as m. The convolutional layers in the first network use 20, 40, 60 and 30 filters, respectively. Every output of a convolutional layer in our models is shaped by a ReLU function. The feature vector is further reduced with pooling layers using the max operator. Furthermore, our models use a fully connected layer which in effect is a MLP. Let σ represent a ReLU activation function, then the output of the hidden layer is computed by:

$$F(x) = \sigma(W \times x) \tag{5}$$

The CNN is trained using stochastic gradient decent for 500 epochs as follows: the learning rate is set to 0.1 and dynamically adjusted down with a decay of 0.01. Let λ represent the initial learning rate, θ represent the learning rate decay, and ω the current epoch, and the learning rate LR is adjusted according to:

$$\mathrm{LR} = \frac{\lambda}{1 + (\omega \times \theta)} \tag{6}$$

During training, the output of the network is shaped by a

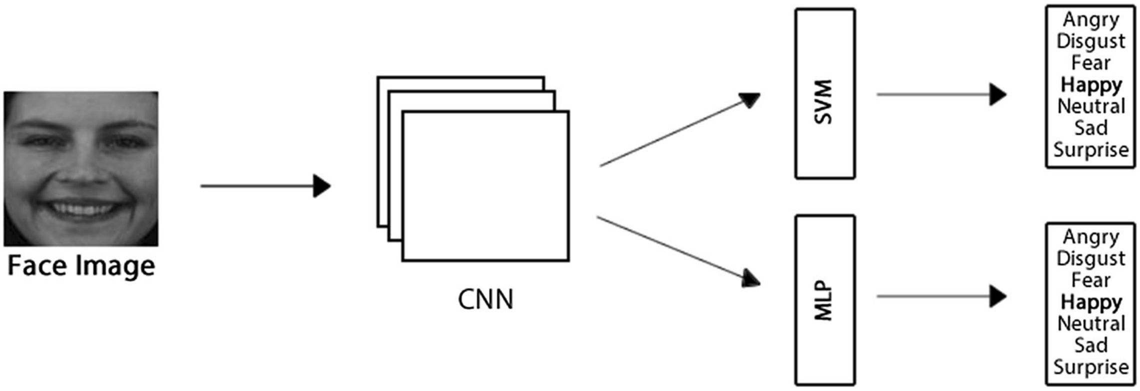

Fig. 4 Illustration of CNN+SVM and CNN+MLP emotion recognition models. CNN topology: Conv, BN, ReLU, MaxPool; Conv, BN, ReLU, MaxPool; Conv, BN, ReLU, MaxPool; Conv, BN, ReLU. MLP topology: Hidden, BN, ReLU. Face image extracted from the KDEF dataset [7]. Only one classifier, SVM or MLP, is applied at a time

SoftMax operator and the cross-entropy loss y is defined by:

$$y = -x_c + \log\left(\sum_j \exp(x_j)\right) \tag{7}$$

where c is the class ground truth. The classification obtained with the network is 91.16% on the testing set. We also tested this network with a SVM as a classifier instead of a MLP. Once the CNN is trained, we remove the classification layer together with the last BN and ReLU activation layers. Classification is then performed using a SVM as follows: the entire dataset is passed through the network, and the resulting downsampled feature vector is used to train the SVM. The accuracy obtained with this approach was 96.26%. Both models are trained and tested on the same splits of the KDEF used to train and test the models using Gabor filters for feature extraction and MLP and SVM for classification.

4 Emotional recognition models, results and discussion

The aim of this work is the development of a new architecture for an emotion recognition model that can allow a robotic companion to recognise human emotions in real-time and in an unconstrained environment. As an initial step, we have developed a set of emotion recognition models that employ SVMs and MLPs as classifiers using Gabor filters for feature representation.

The emotion recognition model that employed Gabor filters and a MLP achieves an overall accuracy of 93.5% on the KDEF dataset. The model that applied a multi-class SVM, also using a bank of Gabor filters for feature extraction, produced a state-of-the-art accuracy rate of 95.58% also on the KDEF dataset.

Our new architecture that only includes four convolutional layers is trained as a single unit using mini batch stochastic gradient decent and takes advantage of BN for a faster convergence. When the classifier is a MLP, the accuracy produced by this network is 91.16%. However, when the classifier is a SVM, it produces a state-of-the-art accuracy rate of 96.26% on the KDEF dataset. Table 1 illustrates the performance of all the models on each class of the dataset and the overall accuracy on the test set. Table 2 illustrates the confusion matrix produced by both models using SVM as a classifier.

As shown in Table 1, we obtained best results when using SVM as a classifier regardless of the feature extraction method. However, our CNN model slightly outperforms Gabor filters as a feature extraction method, at least for this particular dataset. One of the main differences that is shown in Table 2 between CNN+SVM and Gabor+SVM models is that CNN+SVM classified all the *surprise* images correctly obtaining a recall of 100%,

Table 1 Overall accuracy on the KDEF dataset per class per emotion recognition models

	Gabor+MLP	Gabor+SVM	CNN+MLP	CNN+SVM
Angry	96.81	90.48	88.10	95.24
Disgust	91.05	97.62	88.10	95.24
Fear	93.90	97.62	78.57	90.48
Happy	96.34	95.24	95.24	97.62
Neutral	92.41	100	97.62	100
Sad	94.60	95.24	92.86	95.24
Surprise	88.00	92.85	97.62	100
Average	93.5%	95.58%	91.16%	96.26%

Left to right: model using Gabor filters and MLP, model using Gabor filters and SVM, model using CNN and MLP, model using CNN and SVM

Table 2 (a): Gabor + SVM emotion recognition model confusion matrix on the test split of KDEF dataset; (b): CNN + SVM confusion matrix on test split of KDEF dataset

	A	*D*	*F*	*H*	*N*	Sa	Su	Rec
(a)								
A	38	0	0	0	3	1	0	90.48
D	0	41	0	0	0	1	0	97.62
F	1	0	41	0	0	0	0	97.62
H	2	0	0	40	0	0	0	95.24
N	0	0	0	0	42	0	0	100
Sa	0	0	1	0	1	40	0	95.24
Su	2	0	1	0	0	0	39	92.85
Prec	88.37	100	95.35	100	91.3	95.24	100	95.58
(b)								
A	40	0	1	0	0	1	0	95.24
D	1	40	0	0	0	1	0	95.24
F	1	0	38	0	0	2	1	90.48
H	0	0	0	41	1	0	0	97.62
N	0	0	0	0	42	0	0	100
Sa	0	0	1	0	1	40	0	95.24
Su	0	0	0	0	0	0	42	100
Prec	95.24	100	95.0	100	95.46	90.90	97.67	96.26

A angry, *D* disgust, *F* fear, *H* happy, *N* neutral, *Sa* sad, *Su* surprised. *Rec* recall, *Prec* precision. Vertical axis: true labels. Horizontal axis: predicted labels

whereas the Gabor+SVM model obtained a recall of 92.85%. Another discrepancy happened with *angry* emotions, in which the CNN+SVM model classified 95.24% of them correctly compared to 90.48% by the Gabor+SVM model. The Gabor+SVM model also produced a low precision score on *angry* due to a significant number of false positives. Both models produced the same performance, including misclassifications, on *sad* and *neutral*. The main advantage on classification performance from the CNN+SVM model over that of the Gabor+SVM model was on *angry* and *surprise*. And the main advantage of the Gabor+SVM model over the CNN+SVM model was on *Fear*, for which the CNN+SVM model obtained the lowest classification.

It is also evident that SVM produced higher classification rates compared to traditional MLP networks in the case of the dataset used for these experiments. However, we observed that the accuracy rate produced by SVM is largely dependent on the image pre-processing techniques applied to the data. We tried training the SVM with features obtained from the third convolutional layer, but observed a decrease of at least ten per cent in performance. Also slight changes in wavelength or orientation in the Gabor filters greatly influenced the classification performance of the SVM. Another advantage offered by CNN, compared to Gabor filters, is that they learn to extract a feature vector that represents the input image automatically, whereas Gabor filters need to be carefully crafted and optimised by means of trial and error.

Table 1 shows that although the CNN+SVM model produces a substantially higher accuracy rate than the CNN+MLP model, the ratio on misclassified labels remains almost the same. In both cases, fear is the most misclassified class. There is also a correlation on similar performance on particular emotions. Both models produced equal performances on angry and disgust, and neutral and surprise.

Another observation made was that although training a CNN often requires a lengthy process, it is a more automated process than hand-crafting Gabor filters. The bank of Gabor filters used in this work required a number of trial-and-error session given that slight changes in orientation or wavelength change the performance of the SVM significantly. These effects are not observable until a SVM is trained with the Gabor features. Therefore, with these observations, we conclude that convolutional neural networks offer a self-learning alternative approach to Gabor filters for feature selection in the domain of emotion recognition.

Figure 5 illustrates a side-by-side comparison of the image representations produced by our bank of Gabor filters and by the last layer of our convolutional network. As it can be observed, the features learnt by the CNN are much simpler than those produced by the bank of Gabor filters, this in effect can be explained by the loss of information caused by pooling layers in the CNN. However, even though the features extracted from the CNN contain much less information, they still retain important information necessary for emotion recognition. Furthermore, when looking at the filters learnt by the first convolutional layer these looked relatively similar to the ones produced by the bank of Gabor filters. Essentially, the first layer of the CNN learns Gabor-like filters, which are then improved further through deeper layers. Nonetheless, the representations created by both the CNN and Gabor filters highlight facial features such as the eyes, eyebrows and mouth, which allows us to confirm their key role in the recognition of emotions.

A similar hybrid model to the CNN+SVM model using five convolutional layers has been proposed by Ouellet [28]. The model proposed by the author uses a CNN with five convolutional layers and achieves a classification performance rate of 94.7% using a leave-one-subject-out cross-validation approach. In order to compare the CNN+SVM emotion recognition model proposed in this work against that of [28], we tested it on the CK+ dataset using the same leave-one-subject-out protocol. Our hybrid

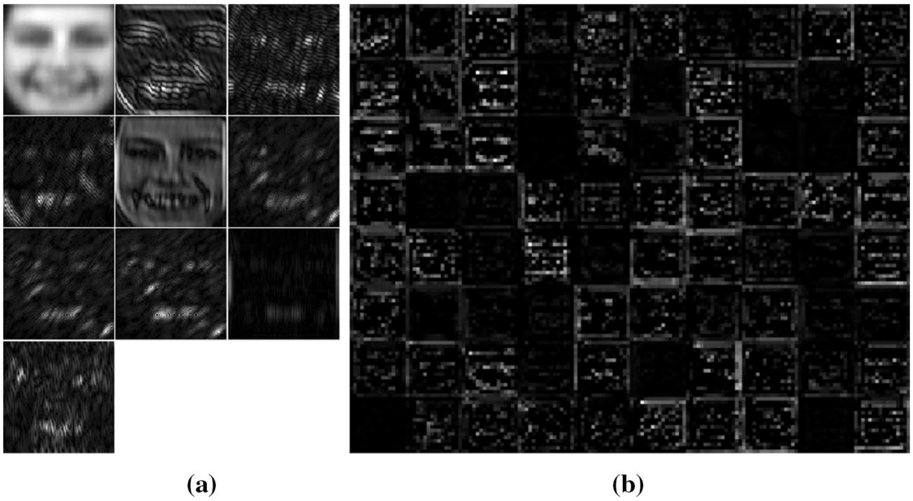

Fig. 5 **a** Magnitude response from Gabor filter and **b** output of the fourth convolutional layer of the CNN

model achieved a classification performance of 95.87% on the CK+ dataset, as illustrated in Table 3. Note that the CNN element has never seen images of the CK+ dataset and was trained on a subset of the KDEF dataset as discussed above.

Our new architecture approach slightly outperforms the model proposed by [28]. With the advantages over the model proposed by [29] being that the CNN component of our model was trained from scratch on a subset of 686 images extracted from the KDEF dataset, whereas the CNN component of the model proposed by the author was originally trained on 1.2 million images from ImageNet in the Large Scale Visual Recognition Challenge 2012, before being tested on the CK+ dataset. Moreover, our model could be more suitable for real-time emotion recognition given that it uses images of smaller size, 100 × 100 compared to 227 × 227, and only has four convolutional layers compared to five, making it faster to classify facial expression images. The fast convergence of the network, an average of 60 min on a quadcore processor for 500 epochs, gave us the opportunity to test different topologies and find the optimal parameters. When adding and removing layers to our model, we observed either a decrease or no significant increase in performance. We attribute the fast convergence rate to the CNN's simplified configuration and the use of BN which allows for larger learning rates [6]. BN also eliminates the need for other heuristics, such us Dropout, that may result in a loss of information.

Table 3 CNN + SVM emotion recognition model confusion matrix on the CK+ dataset after leave-one-subject-out cross-validation

	A	*C*	*D*	*F*	*H*	Sa	Su	Total
A	42	1	1	0	0	1	0	93.33
C	1	16	0	1	0	0	0	88.88
D	0	0	59	0	0	0	0	100
F	0	0	0	22	2	1	0	88.0
H	0	0	0	0	69	0	0	100
Sa	3	1	0	0	0	22	2	78.57
Su	0	1	0	0	0	0	82	98.79
								95.87%

A angry, *C* contempt, *D* disgust, *F* fear, *H* happy, *Sa* sad, *Su* surprised. Note that none of the images were used to train the CNN component

5 Initial real-time robot emotion recognition

Taking into account the key role played by human empathy in HHI, we believe that endowing social robots with the ability to recognise a set of emotions in an unconstrained environment will allow us to move on to the overall goal of creating an empathic robot. In this work, we have presented a set of emotion recognition models that provide state-of-the-art classification performance on the KDEF dataset. However, since the goal of our research is to move towards empathic robots, it is imperative to test the performance of these models in unconstrained environments in which an empathic robot may be used. We do not address empathy in robots in this work, but rather focus on the first stage of it, emotion recognition, and attempt to highlight the issues that the robot might find when performing real-time emotion recognition. Therefore, we have integrated our hybrid model, which produces the best accuracy rate on the KDEF dataset, within a humanoid robot. This section presents our initial experiments on real-time emotion recognition performed by a NAO robot and discusses observations made. As this is an initial experiment, the results presented are provisional and our goal is to establish how the model performs in a real-world application and what issues we will need to consider when making the shift to emotion recognition using a robot.

5.1 Experimental set-up and methodology

This experiment was carried out using our humanoid NAO robot, a 58 cm in height robot with a number of sensors and abilities such as moving, feeling, seeing, speaking, hearing and thinking [41]. Refer to Fig. 6 for an illustration. The NAO robot was set up in a cluttered room with average lighting conditions: lights were kept on and room has windows which allows for ever-changing natural lighting. The robot was placed on a one metre high surface in order to be able to track participant's faces. NAO performed real-time emotion recognition on four participants: two males and two females aged between 20 and 25 years old. Participants were randomly selected and had different ethnic backgrounds, as opposed to participants in the dataset used to train the emotion recognition model who are of white ethnic background. When interacting with the robot, participants were asked to express four different emotions of their choice.

NAO was programmed to track participant's faces and take an image of the participant once a face is detected. Participants were asked to stand within a distance of 1.5 metres away from the robot and express an emotion once NAO confirmed it detected their face: NAO fixes its head in the direction of the person standing in front and turns its eyes green to confirm it detected the user's face. Participants were asked to perform the emotional expression in a natural way.

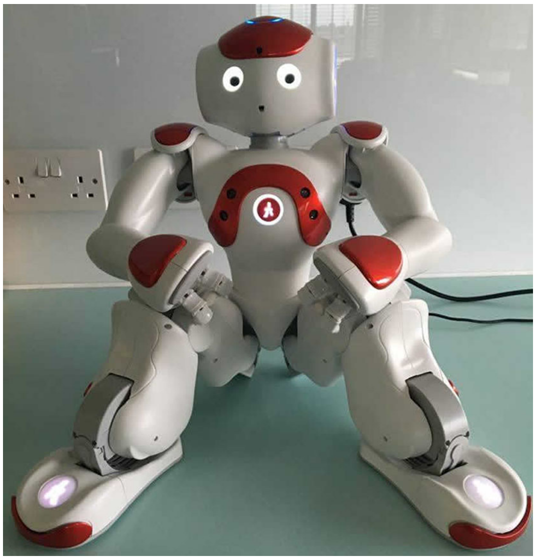

Fig. 6 NAO robot used in our experiments

Since the robot does not have enough processing power, classification of the image is done off-board. The image is sent to an external laptop for processing. The participant's face is then extracted and grayscaled. Once a face image is obtained, this is fed to our CNN+SVM hybrid emotion recognition model which returns a predicted value for the given image. When classification of the image is completed, NAO receives a command to express an appropriate response to the user's emotional state that could in the future be replaced with learned empathic behaviour: for example, express excitement through speech and body language if the classification result was *happy* or express sympathy and support if the classification was *sad*. Given that the purpose of this experiment is to highlight the issues that we will need to address in future work, we did not focus on having the robot perform the right action that would best improve the interaction process with the user. Moreover, we focused only on the recognition of emotions in the user and, thus, had the robot perform hard-coded actions. We performed four trials and asked participants to change the emotion they express each trial. Figure 7 illustrates sample images obtained by the robot and fed to our CNN+SVM emotion recognition model for classification.

5.2 Real-time emotion recognition results and discussion

Performing real-time emotion recognition in unconstrained environments is a challenge difficult to overcome for social robots. One of the main obstacles faced by these machines is the ever-changing environments, which make it difficult to obtain facial expression images of similar quality to ones used to train the emotion recognition models employed by the robot. We have conducted preliminary real-time emotion recognition experiments with our humanoid robot in an attempt to obtain an understanding of the models

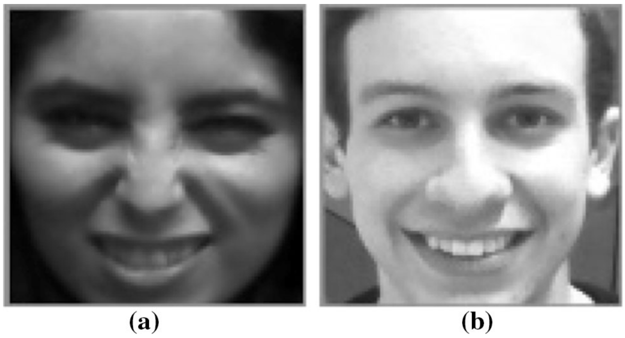

Fig. 7 **a** Sample image from subject F1 illustrating an angry expression, image misclassified with disgust and **b** correctly classified sample image from subject M2 illustrating a happy expression. Actual labels assigned according to participant's feedback performance on unseen, and marginally different, data. The robot used our hybrid CNN+SVM emotion recognition model which produces an accuracy rate of 96.26% on the KDEF dataset. When integrated with the robot, the CNN+SVM model produces a significantly lower average accuracy rate of 68.75%. Although testing of the model on the robot is preliminary, these findings do indicate the impact of the placing the model on a robot in an unconstrained environment and will direct our research in terms of adapting our emotion system so it achieves comparable results to testing using corpora images collected in controlled environments.

We performed four trials on four participants. Each trial consisted of each participant illustrating an emotion, not performed in previous trials, in front of the robot. Table 4 illustrates the emotions illustrated by each participant and the predicted label by the robot using our hybrid CNN+SVM emotion recognition model. The decrease in performance is attributed to the significant difference between the images obtained by the robot and those in the KDEF dataset. Moreover, the emotion recognition model has never seen images of any of the subjects before. In addition to this, in the KDEF dataset all faces are positioned in fixed coordinates. Whereas in our experiment, participants stood at varying distances from NAO, and given the difference in height the position of each participant's face is located at different coordinates in the image space. Furthermore, the lighting conditions varied during each trial. Nonetheless, these are preliminary results intended to shift our focus onto issues that will impact the robot's performance. As a consequence, future work will focus on performing emotion recognition with different light conditions and for faces at different angles.

One of the observations made was that NAO confused happy with disgust and sad with neutral. In the case of sad and neutral, it can be explained by the similarity between the two expressions; however, in the case of happy with disgust it is difficult to come to a conclusion given that happy is not very similar with disgust. However, as shown in Fig. 7a, the *angry* expression illustrated by participant F1 can easily be confused with *disgust*, as done by our emotion recognition model. In this particular instance, *disgust* was the class with the highest value and *angry* the one with the lowest value. Another observation made was that although participants were asked to illustrate emotions in a natural way, they reported emphasising their facial expressions more than they normally would since they were conscious of the difficulty of recognising emotions.

Another issue that needs to be taken into account is the technological limitations of our robot and most other robots. NAO has a pair of built-in cameras that provide a maximum resolution of 1280 × 960 at 30 frames per second. Better performance may be achieved with higher resolutions which may capture facial features better, such as wrinkle lines, which play a key role in the classification of emotions. Although not part of the experiments, we have also observed that people react positively to the empathic behaviours illustrated by NAO after identifying an emotion. This highlights the importance of not only being able to identify an emotion in the user, but also being able to respond with appropriate behaviour according to the identified emotional state.

The main purpose of this experiment was to highlight the issues that social robots will encounter when recognising emotions in real time. When analysing the scenarios where the robot failed and succeeded, we observed two main differences between the images correctly classified and those misclassified. First, as shown in Fig. 7, the lighting on image 7a is much different than that on image 7b. This was also observed on most of the images where the robot failed to detect the right emotion. Second, two of the misclassified images show the participant's faces with a slight angle and tilt. Both of these observations can be justified by the fact that the model was trained on a dataset that contains images of very similar quality. The authors of the KDEF dataset also explain that they centred the faces with a grid. We hypothesise that the robot would greatly benefit from an emotion recognition model trained with much larger datasets containing a wide variety of images taken in unconstrained environments.

Table 4 Real-time emotion recognition results: *F and M* denote participant's gender; *emotion* represents emotions expressed by participant and *label* the label predicted by NAO

Participant	Trial 1		Trial 2		Trial 3		Trial 4	
	Emotion	Label	Emotion	Label	Emotion	Label	Emotion	Label
Subject F1	Happy	Disgust	Disgust	Disgust	Sad	Sad	Neutral	Neutral
Subject F2	Angry	Disgust	Disgust	Disgust	Fear	Fear	Happy	Happy
Subject M1	Angry	Angry	Neutral	Sad	Surprise	Surprise	Sad	Neutral
Subject M2	Happy	Happy	Neutral	Neutral	Sad	Neutral	Surprised	Surprised
Accuracy	2/4		3/4		3/4		3/4	
	11/16							

Last row shows the overall classification accuracy

In this work, we have presented preliminary results on real-time emotion recognition performed by our humanoid robot in an unconstrained environment. The purpose of this experiment is to obtain an understanding of the performance of emotion recognition models in such environments and highlight the issues affecting performance to be addressed in future work. Our real-time emotion recognition results showed that NAO successfully recognised participant's emotional states 11 out of 16 times. The main advantage offered over other systems that perform real-emotion recognition is that our model recognises seven different emotions compared to the traditional positive, neutral and negative emotional states [14]. Moreover, these results allowed us to conclude that emotion recognition in unconstrained environments is possible, though there are some issues that need to be considered in future work, such as illumination and face angle invariance. Furthermore, these results allow us to progress towards automated emotion recognition in social robots, which will, in the near future, allow us to create a system that does not only allow our robot to identify seven different emotional states in the user, but also empathise with them.

6 Conclusions and future work

In this work, we have presented a set of emotion recognition models trained and tested on the KDEF dataset. Our best model was a combination of a CNN and a SVM, and it produced a state-of-the-art performance rate on the KDEF dataset and comparable results to larger models on the CK+ dataset. When tested on the CK+ dataset, note that none of the CK+ images were used to train the CNN component of the model; this hybrid CNN+SVM architecture produced slightly better results than a larger model proposed by [28] and falls 3.73% short from the state of the art [38]. Nonetheless, our hybrid model uses less data to train, converges relatively quickly and has a smaller number of model parameters. This work has also showed the advantages of CNN over Gabor filters in terms of feature extraction and the advantages of SVM over MLP for feature classification, at least in terms of emotion recognition.

This hybrid architecture offers novelty over similar approaches, also employing a CNN, in its simplified configuration that requires less hyperparameters, is relatively faster to train compared to the current state of the art and is able to learn on smaller amounts of data. Moreover, this hybrid architecture offers comparable, and in some cases better, classification performance rates than larger and more complex architectures. This hybrid architecture also takes advantage of BN for a faster convergence. Note that one of the main limitations of this work is that it does not perform as well in real-life scenarios, as discussed in section five, though this will be explored in future work.

In future work, we will shift our focus to the Deep CNN model, given the advantage that it offers over Gabor filters for automatic feature extraction and representation. We will explore the performance of the same model when the convolution layers are pre-trained as stacked Auto-Encoders; each layer is trained to encode the input and a layer is added to decode the downsampled representation obtained by the first layer. Once all layers are trained individually, they are combined into a single model and the classification layer is added. The entire model is then trained, for the classification layer, and fine-tuned, for the convolutional layers. This method has proved to be successful in the past [42]. In addition to this, we will explore the possibility of reducing the number of layers or using a random set of patches as a representation of the image, in order to further improve training time by taking advantage of the hybrid model's simplified architecture and its ability to learn on a smaller number of training samples.

We have also provided preliminary findings on real-time emotion recognition performed by our NAO robot using the new hybrid CNN+SVM model. This experiment was carried out in an unconstrained environment and was intended to highlight the issues that a social robot may face when performing real-time emotion recognition. Letting our robot to perform real-time emotion recognition allowed us to realise the importance of training the emotion recognition models with realistic data, taken in unconstrained environments where the robot will be used. This has been pointed out by Castellano et al. [11] in the past. As a result, we will look into the development of a dataset with images taken in real-life scenarios in which a social robot could potentially be used. This will also allow us to train the robot with images that are not always perfect, do not have the same illumination conditions or have the perfect angle. This is because the robot will not always be at the same height of the user or be able to obtain an image with the face centred. Although there exist methods to correct this, they would only add complexity and delay the response of the robot. In an ideal scenario, an emotion recognition model should produce similar performance on the dataset used for training and in the environment in which it will be used, regardless of the environment conditions.

Once an emotion recognition model produces promising performance for real-time emotion recognition on the NAO robot, the next step will be to develop a model that allows a robot to learn the same emotional state within its very own system: composed of sensor and motor values. We hypothesise that applying the properties of mirror neurons, i.e. simultaneous action perception and execution, to a neural architecture would allow us to create a system that allows the robot empathise with users in real-time and with self-learnt actions rather than hard-coded ones as done in this work. This model will have to take into account that the robot is constrained by its ability to demonstrate emotional states using body language and facial expressions. This will naturally be subject to the robot being used and its anthropomorphic characteristics. Finally, we will explore the inclusion of reinforcement learning for continuous learning within the robot and allow it to adjust its behaviours according to the user responses. We believe that this is a viable path to a more personalised and possibly intimate interaction between robots and humans.

In this work, we made progress towards the development of an empathic robot, a robot with the ability to (i) recognise human emotions through facial expressions, (ii) illustrate emotional states itself, and (iii) automatically and autonomously produce and associate responses to specific emotional states. We have targeted the first and perhaps most essential skill that an empathic robot must possess: recognising emotions. The work presented here was based on the hypothesis that in order to develop machines that can express human-level intelligence, it is imperative to interpret existing knowledge of how the human body works and apply it to the computational models designed to provide robots with intelligence. Our work intends to reduce the gap between artificial and natural mechanisms by incorporating existing knowledge in the field of neuroscience into the development of artificial neural networks for emotion recognition and empathy imitation in a social robot. This will be achieved by making use of approaches such as features extraction, recognition and empathy that associates recognition and production in the manner found in the biological systems. In this paper, we addressed what is considered to be the first step towards empathic robots, emotion recognition.

Acknowledgements We would like to thank the School of Computing, Electronics and Mathematics, at Coventry University for funding this research. We would like to also acknowledge the contribution of the Barry Gidden fund which partially funded this work. The authors would also like to acknowledge the invaluable contribution of Maria Charalambous, Kenny Ruiz, Ibrahim Alamakky and Danielle Labhardt in creating this paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Scassellati B, Admoni H, Mataric M (2012) Robots for use in autism research. Annu Rev Biomed Eng 14:275–294
2. Fasola J, Mataric M (2013) A socially assistive robot exercise coach for the elderly. J Hum Robot Interact 2(2):3–32
3. Chang W-L, Šabanovic S, Huber L (2013) Use of seal-like robot PARO in sensory group therapy for older adults with dementia. In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction. IEEE Press
4. Soler MV, Agüera-Ortiz L, Rodrguez JO, Rebolledo CM, Muoz AP, Prez IR, Ruiz EO, Snchez AB, Cano VH, Chilln LC, Ruiz SF (2015) Social robots in advanced dementia. Front Aging Neurosci 3:133. https://doi.org/10.3389/fnagi.2015.00133
5. Ruiz-Garcia A, Elshaw M, Altahhan A, Palade V (2016) Emotion recognition using facial expression images for a robotic companion. In: Engineering applications of neural networks: 17th international conference, EANN 2016, Aberdeen, UK, September 2–5, 2016, proceedings, pp 79–93. doi:https://doi.org/10.1007/978-3-319-44188-7_6
6. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. http://arxiv.org/abs/1502.03167
7. Lundqvist D, Flykt A, Öhman A (1998) The Karolinska Directed Emotional Faces—KDEF. CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet. ISBN 91-630-7164-9
8. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn–Kanade dataset (CK+): a complete expression dataset for action unit and emotion-specified expression. In: Proceedings of the third international workshop on CVPR for human communicative behavior analysis (CVPR4HB 2010), San Francisco, USA, pp 94–101
9. Robertson J (2014) Human rights vs. robot rights: forecasts from Japan. Crit Asian Stud 46(4):571–598. doi:https://doi.org/10.1080/14672715.2014.960707
10. Dahl T, Boulos M (2013) Robots in health and social care: a complementary technology to home care and telehealthcare? Robotics 3:1–21
11. Castellano G, Paiva A, Kappas A, Nabais F, Aylett R, Barendregt W, Hastie H, Bull S (2013) Towards empathic virtual and robotic tutors. In: Lecture notes in computer science, pp 733–736
12. Toh LPE, Causo A, Tzuo PW, Chen IM, Yeo SH, Link S (2016) A review on the use of robots in education and young children. Sci Direct 19:148–163
13. Kory Westlund J, Gordon G, Spaulding S, Lee J, Plummer L, Martinez M, Das M, Breazeal C (2015) Learning a second language with a socially assistive robot. In: The 1st international conference on social robots in therapy and education, 2015, Almere, The Netherlands. https://www.media.mit.edu/publications/learning-a-second-language-with-a-socially-assistive-robot/
14. Affectiva (2016). http://www.affectiva.com/
15. Kim ES, Berkovits LD, Bernier EP, Leyzberg D, Shic F, Paul R, Scassellati B (2013) Social robots as embedded reinforcers of social behavior in children with autism. J Autism Dev Disord 43(5):1038–1049
16. Rabbitt SM, Kazdin AE, Hong JH (2015) Acceptability of robot-assisted therapy for disruptive behavior problems in children. Arch Sci Psychol 3(1):101–110. https://doi.org/10.1037/arc0000017
17. KSERA (2016) Knowledgeable service robots for aging. http://www.aat.tuwien.ac.at/ksera/index_en.html
18. GeriJoy (2016) Care and companionship for seniors—GeriJoy. http://www.gerijoy.com/
19. Leite I, Pereira A, Mascarenhas S, Martinho C, Prada R, Paiva A (2013) The influence of empathy in humanrobot relations. Int J Hum Comput Stud 71(3):250–260
20. Graaf MMA, Ben Allouch S, Dijk JAGM (2016) Long-term acceptance of social robots in domestic environments: insights from a user’s perspective. In: AAAI

21. Duffy BR (2006) Fundamental issues in social robotics. Int Rev Inf Eth 6:31–36
22. Boughrara H, Chtourou M, Ben Amar C, Chen L (2014) Facial expression recognition based on a MLP neural network using constructive training algorithm. Multimed Tools Appl 75:709–731
23. Kahou S, Michalski V, Konda K, Memisevic R, Pal C (2015) Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on international conference on multimodal interaction (ICMI '15), pp 467–474
24. Lawrence S, Giles CL, Tsoi AC, Back AD (1997) Face recognition: a convolutional neural-network approach. IEEE Trans Neural Netw 8(1):98–113. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=554195
25. Gupta A, Garg M (2016) A human emotion recognition system using supervised self-organising maps. In: 2014 International conference on computing for sustainable global development (INDIACom), pp 654–659
26. Sarnarawickrame K, Mindya S (2013) Facial expression recognition using active shape models and support vector machines. In: 2013 International conference on advances in ICT for emerging regions (ICTer), pp 51–55
27. Sohail ASM, Bhattacharya P (2011) Classifying facial expressions using level set method based lip contour detection and multi-class support vector machines. Int J Pattern Recognit Artif Intell 25(6):835–862. doi:https://doi.org/10.1142/S0218001411008762
28. Ouellet S (2014) Real-time emotion recognition for gaming using deep convolutional network features. CoRR abs/1408.3, 6. http://arxiv.org/abs/1408.3750
29. Ahsan T, Jabid T, Chong U-P (2013) Facial expression recognition using local transitional pattern on gabor filtered facial images. IETE Tech Rev 30(12):47. http://tr.ietejournals.org/article.asp?issn=0256-4602;year=2013;volume=30;issue=1;spage=47;epage=52;aulast=Ahsan l slightly outperforms Gabor filters as a feature extraction method, at least for this particular dataset
30. Chelali FZ, Djeradi A (2015) Face recognition using MLP and RBF neural network with Gabor and discrete wavelet transform characterization: a comparative study. Math Probl Eng 2015:1–16. http://www.hindawi.com/journals/mpe/2015/523603/
31. Mehta N, Jadhav S (2016) Facial emotion recognition using log Gabor filter and PCA. In: Proceedings of international conference on computing communication control and automation (ICCUBEA), pp 1–5
32. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings of the fourth IEEE international conference on automatic face and gesture recognition (FG'00), Grenoble, France, pp 46–53
33. Paul PP, Monwar MM, Gavrilova ML, Wang PSP (2010) Rotation invariant multivew face detection using skin color regressive model and support vector regression. Int J Pattern Recognit Artif Intell 24(8):1261–1280. https://doi.org/10.1142/S0218001410008391
34. Khan SA, Hussain A, Usman M, Nazir M, Riaz N, Mirza AM (2014) Robust face recognition using computationally efficient features. J Intell Fuzzy Syst 27(6):3131–3143.<GotoISI>://WOS:000345981600037
35. Hassner T, Harel S, Paz E, Enbar R (2015) Effective face frontalization in unconstrained images. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 4295–4304. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7299058
36. Hewahi NM, Baraka ARM (2012) Impact of ethnic group on human emotion recognition using backpropagation neural network. BRAIN. Broad Res Artif 20–27. http://brain.edusoft.ro/index.php/brain/article/view/284
37. Khashman A (2009) Application of an emotional neural network to facial recognition. Neural Comput Appl 18(4):309–320
38. Burkert P, Trier F, Afzal MZ, Dengel A, Liwicki M (2015) DeXpression: Deep Convolutional Neural Network for expression recognition. arXiv preprint 1–8. http://arxiv.org/abs/1509.05371
39. Beaudry O, Roy-Charland A, Perron M, Cormier I, Tapp R (2014) Featural processing in recognition of emotional facial expressions. Cognit Emot 28(3):416–432. http://www.ncbi.nlm.nih.gov/pubmed/24047413
40. Kuhn H, Tucker A (1951) Nonlinear programming. In: Proceedings of the second Berkeley symposium on mathematical statistics and probability, pp 481–492
41. Aldebaran (2016) Who is NAO? https://www.aldebaran.com/en/cool-robots/nao
42. Tan CC, Eswaran C (2010) Reconstruction and recognition of face and digit images using autoencoders. Neural Comput Appl 19(7):1069–1079