# Emotion Recognition from Face Images in an Unconstrained Environment for usage on Social Robots

1st Nicola Webb
*Bristol Robotics Laboratory*
*University of the West of England*
Coventry, UK
email: nicola.webb@brl.ac.uk

2nd Ariel Ruiz-Garcia
*Computing, Electronics and Mathematics*
*Coventry University*
Coventry, UK
ariel.9arcia@gmail.com

3th Mark Elshaw
*Computing. Electronics and Mathematics*
*Coventry University*
Coventry, UK

4th Vasile Palade
*Centre for Data Science*
*Coventry University*
Coventry, UK

***Abstract*—Deep neural networks have proven to be efficient systems for learning complex data representations. However, one of their main constraints is their inability to deal with changes in the data distribution. For instance, in real-time facial expression recognition, the data used to evaluate a model commonly differs in quality compared to that used to train the model, leading to poor generalization performance. In this work we propose a novel Deep Convolutional Neural Network (CNN) architecture pre-trained as a Stacked Convolutional Autoencoder (SCAE) to address emotion recognition in unconstrained environments. The SCAE is trained in a greedy layer-wise unsupervised fashion, and combines convolutional and fully connected layers and learns to encode facial expression images as an illumination and facial pose invariant feature vector. The CNN offers state-of-the-art classification rate of 99.52% on a combined corpus of gamma corrected version of the CK+, JAFFE, FEEDTUM and KDEF datasets. When evaluated on unseen data obtained in unconstrained environments, our approach achieves 79.75%, an increase of over 28% compared to a CNN without our pre-training approach, supporting the methodology proposed in this work.**

***Index Terms*—Stacked Convolutional Autoencoders, Greedy Layer-Wise Training, Deep Learning, Emotion Recognition, Social Robotics**

## I. Introduction

Emotion recognition is essential for human-robot interaction. Social robots are finding places in everyday activities industries such as specialized education [1] and care [2]. An increase in demand for social robots means an increasing importance in the improvement in human-robot interaction (HRI), with the development of more effective communication. One way to improve this communication is through refining a robot's ability to recognize human emotions.

Although deep learning (DL) has set many benchmarks in the domains of computer vision and signal processing, one limitation of deep models is their inability to generalize on novel data with non-uniform conditions. For instance, in a real-life scenario where a social robot is to be used, obtaining good quality images with a full-frontal view of the user's face can be very difficult. This and other factors such as constant changes in natural and artificial lighting, lead to drastic changes in the data distribution and poor generalization performance.

In this paper, we propose a novel deep CNN architecture to address emotion recognition in unconstrained environments with specific attention to facial pose and illumination invariance. The CNN is initially pre-trained as a SCAE that learns to map facial expression images to a hidden pose and illumination invariant feature vector, and to a representation of the faces with zero-degree pose. Unlike traditional autoencoder models that employ the input image as the target image, the model presented here employs an image of frontal view and with good level of image luminance as the target reconstruction. The encoder element of the SCAE, which produces a translation, illumination and facial pose invariant feature vector, is then used to initialize a CNN. The CNN is fine-tuned for classification and evaluated on a corpus collected by a Nao robot in uncontrolled environments.

In addition, we propose training the SCAE model using an improved version of the Greedy Layer-Wise (GLW) algorithm that: improves training times, reduces error accumulation in early layers, and improves overall generalization performance.

Although our proposed approach relies on the availability of multi-illumination images to learn to encode images as an illumination invariant feature vector, we demonstrate that by employing gamma correction [3], we can obtain equal or better results.

## II. State of the Art

We are starting to see social robots being implemented into real world applications, as we are experiencing an increasingly aging society that would benefit from the assistance of such. However, for this transition to be as seamless as possible, these robots will require the ability to produce empathetic responses

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

to the emotions of the humans they assist. Empathy is a key part in human-human communication, so replication of such for social robots is important. The development of real time emotion recognition will aid in the creation of empathetic robots and possibly an increase in acceptance of robots in society. This section discusses the recent advances in human robot interaction in regard to social robots, particularly on their uses within a care environment [2], [4], [1]. Additionally, the current state-of-the-art deep learning approaches to facial emotion recognition are discussed.

### A. Social Robots

There is an increasing demand for social robotics in the care industry. The EU is predicted to be facing a population crisis by the year 2060, where the percentage of those over 65 is set to double from 27.8% to 50.1% [5]. In response to this, EU governments are looking to implement assistive robots into homes in the upcoming years [6]. The authors of [2] have evaluated the use of social robots in care homes for those with mild dementia. Using Silbot3 as the robotic platform, participants were asked to play through 6 different scenarios with the robot, such as waking them up, and their responses were recorded. Most participants gave positive feedback on their interactions, giving praise to the robot's reminders and voice recognition functionality. The use of voice recognition is helpful as it is easier for participants to interact with the robot. However, the addition of emotion recognition would have been beneficial for some of the scenarios, such as if they are asked why they do not wish to take their medication, so the appropriate actions can be taken.

A study investigating the effects of social robots on children with Autism Spectrum Disorder (ASD) was conducted in [4]. The authors investigated the effectiveness of the inclusion of a socially assistive robot in an intervention for children with ASD. The experiment involved placing children with ASD and a supervisor in a room with a robot. The children were left to interact with the robot in conditions where the robot performed a simple action randomly, and an condition where the robot's actions were prompted by interaction from the child. The total amount of speech increased in the second condition, as well as interactions with the robot. The results show how meaningful interaction from a socially assistive robot can provoke social behavior from a child with ASD.

Another study into the effects of social robots on children with ASD was conducted [1]. The authors explored the impact of social robots on gestural usage, on 13 children with ASD. The participants were first shown a selection of 8 gestures by a Nao robot with associated meanings, for example, hands covering ears to represent 'noisy'. They were then asked to act out gesturally the correct response for a given scenario based on what they had previously learned from the Nao. The participants were able to better replicate the correct gestures per scenario, showing how social robots can be a useful tool in specialized education and development. These studies demonstrate that emotion perception and replication are essential for meaningful and engaging human-robot interaction.

### B. Emotion Recognition

Emotion recognition can be achieved using facial expressions, speech, body language, or a combination of these. We look at emotion recognition from facial expressions since in unconstrained environments it is easier to obtain facial images than other data, such as audio or full body images. Moreover, emotion recognition from facial expressions has proven to yield higher accuracy rates.

Most work into emotion recognition is carried out on images of a single person. [7] approached facial emotion recognition at the group level using two types of CNN. Firstly, the authors used two CNNs for aligned and non-aligned facial images for emotion recognition of a single face. Secondly, they used a CNN for the sentiment of the image as a whole, for example, an image taken from a wedding would denote happiness. They improve on the baseline for this dataset to achieve 83.9% and 80.9% classification accuracy on the validation and testing sets respectively.

Concerning emotion recognition on static images, high levels of accuracy have been achieved on facial expression corpora collected in controlled environments. For example, [8] proposed a new CNN architecture for emotion recognition, which includes two parallel feature extraction blocks. The authors evaluated their proposed method on the Extended Cohn-Kanade (CK+) Dataset [9] and the MMI Facial Expression Dataset [10]. On the CK+ dataset, they obtained a result of 99.6%, an increase from the previous bench-mark of 99.2% using a conventional CNN.

Work into emotion recognition has expanded from recognition on facial expression corpora collected in static conditions. [11] trained a deep CNN on a dataset of irregular facial expression images. The primary dataset used was the EmotiW dataset, comprised of a series of clips from movies, each labeled with the 7 basic emotions. In addition, they trained their CNN on the Toronto Face Database [12], containing typical frontal, well-lit facial expression images, and images taken from Google images. They achieved a final classification accuracy of 41.03%. The usage of a more realistic dataset is an important step towards true real time emotion recognition.

Another work regarding facial emotion recognition with irregular images was presented by [13]. The authors trained multiple deep CNNs on two datasets, resulting in a final accuracy of 61.29%. The FER dataset was used for pre-training, with some images undergoing randomized perturbation, subsequently providing more unseen images for the network. An accuracy of 55.69% was attained. The networks were then fine-tuned on the Static Facial Expressions in the Wild dataset, a dataset comprised of labeled movie stills, which provide more natural facial expressions than standard facial expression datasets.

[11] approached facial emotion recognition in video using both audio and visual facial expressions. They used the FER-2013 Face Database and the Toronto Face Dataset for training their model and used the Acted Facial Expression in the Wild (AFEW) dataset for testing. They implemented a deep

CNN for classifying facial emotions from the video footage and achieved an accuracy of 37.35%. In combination with a deep belief network on the extracted audio, they achieved an improved 44.71% rate of classification.

The authors of [14] propose a method to deal with illumination invariance using an adaptive filter based on temporal local scale normalization, and pre-training as SCAE on large amounts of unlabeled data. The authors achieve an accuracy of 90.52% on the CK+ corpus when performing emotion recognition. A similar approach was proposed by [15], in which the authors employ a SCAE model to deal with illumination invariance in emotion recognition. The authors used an image with good illumination as the target reconstruction for images with poor luminance. The authors report an accuracy rate of 99% on the CK+, Facial Expressions and Emotions (FEEDTUM) [16], Japanese Female Facial Expressions (JAFFE) [17] and the Karolinska Directed Emotional Faces (KDEF) [18] corpora.

The authors [19] propose a novel approach to face frontalization using a Generative Adversarial Network combined with a 3D Morphable Model. This approach requires a 3d scan of faces as reference for the GAN to create a frontal representation. Moreover, the approach proposed by the authors is able to produce landmarks to localize specific features in a face. However, the authors do not focus on emotion recognition, and as such, there is no emphasis on retaining facial features necessary for emotion recognition.

Although the works discussed in this section achieve remarkable results, they do not explicitly address some of the main issue in facial expression recognition: pose and illumination invariance. Some of the works either address one or the other, or focus on face frontalization but do not evaluate their approaches on facial expression recognition.

## III. Methodology

As discussed in the previous section, although many works in the literature have achieved state-of-the-art results, they do not address two important problems faced by emotion recognition: illumination and facial pose invariance. Although some of the works do address one or the other, they do not address both concurrently or only focus on face frontalization but not on emotion recognition. We explicitly address these two issues, which we have often observed to be the main cause of poor generalization performance when employing deep learning models in real life unconstrained environments, and focus on retaining facial features necessary for emotion recognition.

Deep CNNs are commonly the preferred choice for image processing tasks due to their ability to exploit spatial information and extract salient features in images. However, because in this experimental setup we aim to reduce facial pose to zero degrees, CNNs are not suitable for this task; some features need to be relocated within image space but convolutional kernels are unaware of global information and therefore are unable to shift features around. On the other hand, Mulltilayer Perceptrons (MLPs) take advantage of global features but fail to consider spatial information. We address pose and illumination invariance by exploiting the ability of CNNs to retain salient features and exploit spatial information, and the ability of MLPs to exploit global information as discussed in more detail in section III-B. The approach proposed reduces the need for very complex and deep architectures with high computational cost.

Our SCAE model is trained to improve image luminance and reduce facial pose. Accordingly, we employ the Multi-PIE dataset [20]. The Multi-PIE dataset captures facial expression images from a range of poses with 19 differing levels of illumination for every image. It contains 750,000 images from 337 participants. Pose variant images were taken from a front facing camera, 0 degrees, to angles of ±90 degrees. Images were taken at intervals of 15 degrees. Since the images at pose greater than ±60 do not contain many visual features, we discard images with a pose greater than this and only use a total of 580, 907 images covering all 19 illumination conditions and 13 viewpoints. This corpus is randomly split into 80% training and 20% validation and because it does not contain labels for emotions it is only used to train the SCAE model in an unsupervised manner.

Since the Multi-PIE corpus does not contain any labels for emotions, and to improve the generalization performance of our model, we combine four different datasets commonly used in the literature: the CK+ [9]; the KDEF dataset [18]; the JAFFE dataset [17]; and the FEEDTUM database [16]. Ekman's six universal emotion categories including neutral are the emotion categories included in each dataset. The emotion categories are therefore as follows: angry, disgust, fear, happy, neutral, sad, and surprise, plus neutral. We consider neutral states as all other emotions develop from a neutral state.

The CK+ contains 486 images from 97 participants. The KDEF dataset contains 4900 images from 70 participants with an even mix of males to females. The JAFFE dataset contains 213 images taken of 10 Japanese female participants. The FEEDTUM dataset consists of video clips of 18 participants reacting to stimuli to provoke an emotive response. Due to every sequence in the FEEDTUM corpus starting and ending with a neutral face, we discard the first 30% and the last 10% of every sequence to ensure that there are no neutral faces labeled as a given emotion.

Every individual corpus is split into 70% training and 30% testing. The resulting subsets are then combined into a single large corpus referred to as Combined Facial Expressions (CFE) hereafter.

To supplement the aforementioned dataset, we collected additional facial expression images in the uncontrolled environments faced by social robots. The pictures were captured with a 58-centimeters-tall humanoid robot Nao robot, which possesses a 1.22-megapixel camera with an output of 30fps. These images were taken in three different sessions with a differing number of participants in each and in two separate environments that both contained large windows. The images were taken over a period of a few hours in each session, which meant the lighting in the environment was varied throughout the images. Figure 1 illustrates sample images from this corpus.

21 male and 7 female participants took part in the sessions, and consisted of students and university lecturers, ranging from ages 20-55 and a mix of at least 5 cultural backgrounds. When participants entered the room, they were asked to seat facing the Nao robot, which was in the sitting position to best match the height of participants' eye line. This resulted in varying level of tilt in the resulting images. Moreover, participants were not asked to remove any accessories they were wearing, or at what distance to sit away from the Nao. They were given instructions to express seven emotions in a natural way. This resulted in a total of 196 images.

To validate the images and overcome participant bias, we asked three independent parties to label each image collected with the emotion they believed it represented. If an image was labeled as the same emotion by each person, the image was put into the final testing dataset. From this, only 121 images were kept. Note that these images are only used to evaluate our model.

### A. *Image Pre-processing*

In order to discard background noise, we employ a Histogram of Oriented Gradients (HOG) face detector [21] to crop faces on all corpora used in this work. Once the faces are extracted, and since color does not add any information necessary for emotion recognition, the images are gray-scaled for dimensionality reduction and to speed up training times. Because the resulting cropped images differ in size, we also scale the images to 224 × 224 using bipolar interpolation. In addition to this, since unlike the images in the Multi-PIE corpus the images in the CFE dataset only contain a single degree of illumination, we employ gamma correction to alter the training subset of this corpus. Gamma correction changes an image's luminance with a non-linear alteration of the input and output values. Given the input image $i$, the altered image $x$ is defined by:

$$x = \left(\frac{i}{255}\right)^{\frac{1}{\gamma}} \times 255 \tag{1}$$

where $\gamma \in \{0.4, 0.6, 0.8, ..., 3.4\}$. However, when $\gamma = 1.0$ the input image remains unchanged and since these images were taken in controlled environments we assume they have a good level of relative luminance. As a result, the unchanged image becomes the target reconstruction image, $x_\mu$, for every input image $x$, including itself, in the SCAE model. Figure 2 shows a sample image after $\gamma$ correction.

For the Multi-PIE corpus, because for every image there exist 19 corresponding copies with varying relative luminance levels and zero facial pose, we estimate the relative luminance for each one of these corresponding images and pick the image with luminance level closest to the mean as the desired reconstruction target. In effect, this means that the image with facial pose at zero degrees and with good luminance level is used as reconstruction target for itself, and all other images with facial pose at $\pm\{0, 15, 30, 45, 60\}$ and 19 different degrees of illumination. Relative luminance is defined by $Y = 0.2126R + 0.7152G + 0.0722B$ where RGB are the color channels.

### B. *Unsupervised Feature Learning*

Learning to classify facial expression images with facial pose as a given emotion can be a challenge difficult to overcome. This is due to many of the features necessary for emotion recognition missing from the facial expression image. As a result, contemporary work in the domain of facial expression recognition is commonly done on frontal-view images. However, such scenarios are unrealistic for emotion recognition in unconstrained environments. Changes in image luminance also increase the complexity of classifying facial expression images and often results in poor generalization performance. In this work we facilitate the task of classifying facial expression images with varying luminance and facial pose by reducing the search space for the classifier. This is achieved by using a SCAE to reduce facial pose to zero degrees and improve relative image luminance. This results in a significantly smaller data distribution, and thus, an exponentially smaller search space for the classifier proposed in section III-C.

An autoencoder is composed of an encoder and a decoder. The encoder is a function $f$ that maps an input $x$ to a hidden representation $h(x)$ in such a way that:

$$h = f(x) = s_f(Wx + b) \tag{2}$$

where $s_f$ is an activation function that provides the encoder network with non-linearity. $W$ is a weight matrix and $b$ a bias. The decoder is a function $g$ that maps $h$ to a reconstruction $y$ that is an approximation of the input $x$. It has the form:

$$y = g(x) = s_f(Wx + b) \tag{3}$$

However, in our experimental setup we want to map an input image $x$ to a target reconstruction $x_\mu$ that lies in a different distribution. Therefore, the proposed autoencoder maps the input $x$ to a an approximation of $x_\mu$. This is accomplished by finding the parameters $\theta$ that minimize the reconstruction error between the reconstruction $y$ and the target image $x_\mu$:

$$J(\theta) = \sum_{x \in D_n} L\Big(x_\mu, g\big(f(x)\big)\Big) \tag{4}$$

where $L$ is the reconstruction error and $D_n$ is the set of training samples form the training subsets of the Multi-PIE and CFE corpora.

Although convolutional networks exploit spatial information in images, they also restrict the freedom of the reconstructions in our set up. Recall that a convolutional layer is defined as:

$$C(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{5}$$

where $I$ is the input image and $K$ is the filter kernel with

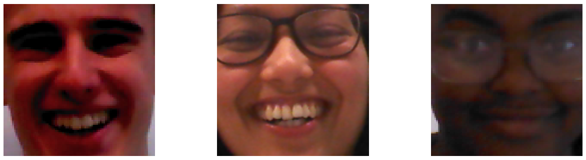

Fig. 1. Sample images collected using our Nao robot. Three subjects from three different ethnic backgrounds illustrating a happy emotional states.

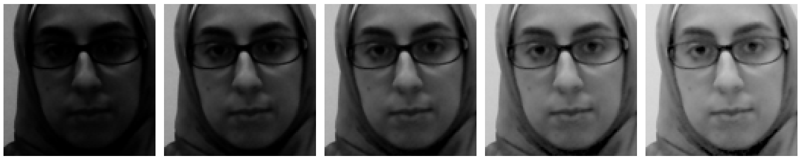

Fig. 2. Sample image after $\gamma$ correction. From left to right: $\gamma = \{0.4, 0.6, 1, 1.4, 1.8\}$.

dimensions $m \times n$. Because the size of the filter kernel only sees a slice of the input image, also with the dimensions $m \times n$, the output is constrained to remain within these boundaries. In our experimental set-up pixel values often need to be shifted outside of this view. For instance, a facial expression image with an estimated pose of $-60$ degrees will contain most of the salient facial features within the left half, or first 100-pixel values, of the image space. A representation of the same image at zero degrees requires these features be placed in the center of the image, therefore many of the values are shifted over 50 places.

Naturally, convolutional layers can be replaced with fully connected ones since they do not take spatial information into account. However, our experiments show that these are prone to overfitting and fail to generalize on unseen data, i.e. data from completely different datasets as those used during training. As a result, we propose a SCAE model that exploits both convolutional and fully connected layers. The encoder model in the SCAE employs two convolutional layers with 20 and 40 $5 \times 5$ filters, followed by a fully connected layer with 1000 hidden units, one more convolutional layer with 80 $3 \times 3$ filters, and another fully connected layer also with 1000 hidden units. This setup ensures that existing facial features are retained through convolutional kernels and repositioned through fully connected layers.

The decoder element is only made of up four fractional-strided convolutional layers, often referred to as deconvolutional layers. Two of the layers have 80 $3 \times 3$ filters and the other two have 40 and 20 $5 \times 5$ filters. Every layer in the SCAE is followed by Batch Normalization layers for faster learning and ReLU layers for non-linearity. The SCAE model is trained for 20 epochs on the training subsets using Adam [22]. Higher number of epochs or other network configurations did not improve the performance of the model. The reconstruction loss $L$ between $y$ and $x_\mu$ is measured using the mean absolute value of the element-wise difference between the desired target and the reconstruction:

$$C = \frac{\sum_{i=1}^{n} |x_{\mu_i} - y_i|}{n} \tag{6}$$

where $x_\mu$ and $y$ are both vectors with a total of $n$ elements. Learning rate is set to 0.1 and remains constant throughout training.

Since training deep models can be a challenging task, we employ GLW [23], which has proven to improve generalization performance, as our training algorithm. In GLW unsupervised training, each individual layer is treated as an individual shallow network and trained individually as an autoencoder. The trained layers are used to extract features and train the next layer. Once all the layers are trained, these are stacked together and trained further as a single network. However, in previous works [24] we have observed this method to be prone to error accumulation in early layers. This often leads to the network learning to learn features that are far from an optimal solution. As a result, instead of only fine-tuning the final stack as a deep autoencoder, we fine-tune at every step: once a shallow autoencoder is trained, we add it to the stack of trained autoencoders and fine-tune the current stack We refer to this approach as Gradual Greedy Layer-Wise algorithm.

The objective of the SCAE is to learn a pose and illumination invariant feature vector. Because it would be computationally expensive to estimate the luminance level of every reconstruction, we only evaluate this on a random batch from the validation subset every two epochs. If the difference between the relative luminance of the reconstructed images and that of the input batch is within a certain threshold, then we stop the training.

However, this does not tell us much about the estimated pose of the reconstructed image. This is because it would be very computationally expensive to estimate it for the deeper layers since we do not have an actual target representation. Nonetheless, our experiments show that by the time the training process is halted according to the relative luminance condition, the facial pose is close enough to zero degrees for the purpose of this research. Furthermore, the nature of this experimental setup does not warrant a perfect reconstruction since we only care about reducing the search space for the classifier.

### C. Facial Expression Classification

Once the SCAE model is trained, the decoder element is removed and replaced with a classification layer to form a CNN classifier. This model is fine-tuned for classification on the training subset of the CFE corpus and evaluated on the testing subset and the NaoFaces corpus. Note that none of the images in the NaoFaces are used during the training process and are only used for testing. The classification loss for this model is measure using a cross-entropy criterion defined as:

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

where $c$ is the class ground-truth for input sample x. Fine-tuning is done for 50 epochs using Stochastic Gradient Descent (SGD) and Nesterov momentum. Learning rate is initialized to 0.001 and remains constant.

For comparison purposes, we also train a ResNet-34 on the CFE corpus to compare the performance of our proposed method against a state-of-the-art architecture. We specifically chose the 34-parameterized layer network since deeper architectures did not produce a significant improvement in performance. Moreover, models with different topologies such as VGG or Inception did not produce better results. This can be justified by the complexity of our datasets, which do not require more layers to learn to extract salient features.In addition, we compare against the results presented in [15] which only deal with illumination invariance, whereas here we deal with illumination and facial pose invariance at once. To the best of the author's knowledge no other works in the literature explicitly deal with both issues at once.

The ResNet-34 model is also trained using SGD with Nesterov momentum for 100 epochs, further training did not yield any better performance. We follow the same training process as done by [25], which includes several preprocessing steps such as resizing of images, luminance and color adjustment and flipping the images. Evaluation was done using five crops: all corners and center crop, as done by the authors. Without these pre-processing steps the accuracy dropped marginally.

## IV. Results and Discussion

### A. Unsupervised Feature Learning

In this work we have introduced a novel SCAE architecture that exploits convolutional layers to retain spatial information and salient facial features through filter kernels, and fully connected layers to relocate these features and reduce facial pose to zero degrees. The SCAE model also learns to reconstruct all images with a similar level of image luminance and can compensate for missing information, e.g. when some of the facial features are not visible it predicts what they look like. In addition, the SCAE model is trained using Gradual-GLW, an improved version of GLW.

As it can be observed in Figure 3 the reconstructions produced by the SCAE model retain all the facial features necessary for emotion recognition intact. In cases where the images have a facial pose and very poor illumination, the SCAE model can still predict what the frontal view of the face looks like. Moreover, all reconstructions have virtually the same level of image luminance. In effect, by reducing facial pose and improving image luminance, our SCAE model reduces the complexity and distribution of the input data, greatly reducing the search space of the CNN classifier, which only has to deal with frontal view images with good illumination. The success of the SCAE model in reducing facial pose and improving image luminance also indicate that Autoencoders can in fact learn to map a given distribution to a different distribution.

The proposed training method produces image reconstructions with significantly less error than when using GLW over the proposed Gradual-GLW. Gradual-GLW also resulted in faster convergence of the SCAE. Moreover, much of the success of the SCAE model in reconstructing facial expression images with a facial pose as images with zero degree pose is due to our novel SCAE architecture. The SCAE model proposed employs convolutional layers to exploit salient features and spatial information, and fully connected layers to reposition these features.

### B. Facial Expression Classification

In [15] we proposed a similar approach to address illumination invariance and reported a state-of-the-art accuracy rate of 99.14% on the test subset of the CFE corpus. The pre-training approach was similar to the one proposed in this paper; a SCAE that maps an input image to a representation of the same with better luminance. However, in this work we have also addressed pose invariance and trained the SCAE using our proposed Gradual-GLW training method. The classification results obtained using the methodology introduced in this work when evaluated on the CFE corpus are of 99.52%, marginally improving the state-of-the-art reported in [15] and significantly outperforming the approach done by [14], who obtain 90.52% on the CK+ corpus using their illumination invariant approach which also includes a SCAE. Although not significantly better, the results achieved in this work support the proposed methodology.

As it can be observed in Table I, when tested on our data collected from the Nao robot our accuracy rate is 79.75%. The proposed method in this research also outperforms the proposed method by [15] on unseen data. Nonetheless, the method presented here tries to deal with pose and illumination invariance, compared to just illumination as done in [15]. Additionally, when neither pose nor illumination are considered, we obtain a significantly lower performance of 50.80% on unseen data. These results demonstrate the potential of our pre-training approach which intends to reduce the search space of the classifier by reducing the data distribution.

In regards to specific classes, $Disgust$, $Fear$, $Neutral$, were some of the most misclassified emotions by the other two approaches and for which there was more room for improvement. Moreover, because in previous work, these classes along with $Surprise$ were mostly confused with one another, by improving the recognition of one of them, the model is

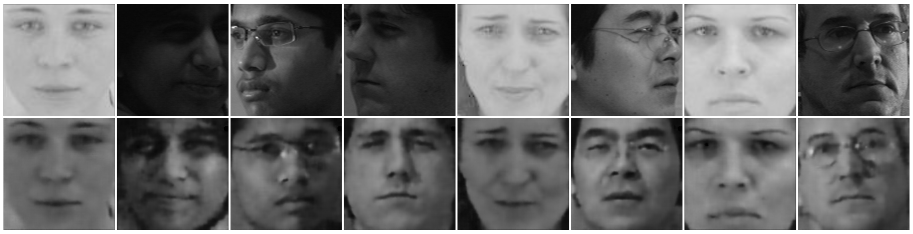

Fig. 3. Sample image reconstructions to zero degree facial pose.

TABLE I

Classification performance comparison on our NaoFaces corpus: ResNet-34 —state-of-the-art classifier; CNN —illumination invariant classifier from [15]; SCAE+CNN(ours) pose and illumination invariant classifier proposed in this work. Note that this corpus is only used for testing.

| | $Resnet34$ | $CNN$ [15] | SCAE+CNN(ours) |
|---|---|---|---|
| $Angry$ | 50.00% | 85.14% | 85.14% |
| $Disgust$ | 41.16% | 66.66% | 75.00% |
| $Fear$ | 54.54% | 72.72% | 81.81% |
| $Happy$ | 64.28% | 96.43% | 96.43% |
| $Neutral$ | 38.46% | 42.3% | 57.69% |
| $Sad$ | 54.54% | 72.72% | 72.72% |
| $Surprise$ | 52.63% | 78.95% | 89.47% |
| $Total$ | 50.80% | 73.55% | 79.75% |

able to better tell the difference between them. $Sad$ has always been one of the most difficult emotions to classify as we have observed in previous works and in the literature. This has mostly been confused with $Neutral$, which also explains the low performance on that particular class. Moreover, $Neutral$ is the only class to have been confused with most others. We hypothesize that this is in part due to employing the FEEDTUM corpus in our training data. This is composed of transitions from neutral states to peak of a given emotion, and back to a neutral state. We discarded the first $30\%$ and last $10\%$ of every sequence but every sequence is different and this did not guarantee that all the neutral faces were removed completely. As a result, many images with a neutral state were labelled as a given emotion during training.

Finally, happy is the emotion best classified by the other two architectures and as such the one with less room for improvement. The few remaining misclassifications may be due to the way people express emotions, since our dataset contains images from subjects with different ethnic backgrounds. For instance, in Figure 1, two of the subjects express happiness with a big smile and illustrating their teeth, whereas the last one does not. These subtle differences are likely to be seen as different by the network, given that the CNN will highlight different salient features to represent each image.

The difference in performance on seen and unseen data can be attributed to the difference in the distribution of both. In the NaoFaces corpus, participants were allowed to wear scarves, glasses, hats, and other accessories. As opposed to the images in the CFE corpus in which facial features of all participants are clearly visible. Moreover, other factors such as facial tilt may have an effect on the performance of the model. Nonetheless, the proposed methodology does significantly outperform state-of-the-art classifiers when no pre-training is done.

These observations demonstrate the robustness of our method when dealing with nonuniform data or changes in the data distribution. Moreover, our model is more suitable and effective for the problem of real time emotion recognition, as it able to classify images of those who are not looking directly at the camera. Our model is able to reconstruct the parts of an image that are not visible, such as when only part of a face is visible due to high degrees of pose. This implies a reduced need for training with such large amounts of data, as the model could compensate for when information is missing from an image.

To the best knowledge of the authors, this is the first work in the literature that attempts to jointly address pose and illumination invariance in the domain of facial expression recognition. Although other works have focused on pose invariance, they do not apply it to the domain of emotion recognition. Similarly, other contemporary works addressing illumination invariance employ more complex or hard-coded methods such as noise injection [26], blurring images with Gaussian filters [27], a combination of histograms, principal component analysis (PCA) and discrete cosine transforms [28], or complex and very deep CNN architectures [29]. Our approach greatly improves on such methods by proposing a method that learns to adjust both, illumination and facial pose, and significantly reducing the search space of the CNN classifier.

Although our method heavily relies on the availability of multi-illumination data, we have also demonstrated that gamma correction can be employed when there exists a lack of data. Moreover, in theory, our approach should reduce the need for more complex image pre-processing approaches often employed when training deep networks such as the use of histograms or adjustment of color [25].

## V. Conclusions and Future Work

In this work we have introduced a novel architecture for pose and illumination towards real time emotion recognition. Our method exploits convolutional and fully connected layers to improve image luminance and reduce facial pose to zero degrees using our training approach Gradual-GLW, which overcomes some of the limitations of GLW, mainly error accumulation, and produces remarkable facial pose and illumination invariant reconstructions. We also demonstrated that our proposed method offers state-of-the-art classification performance on unseen data collected in uncontrolled environments with a Nao robot.

Our method achieves 99.52% classification performance on a combined dataset of standard facial emotion images. When evaluated on novel data with nonuniform conditions taken by a Nao robot we achieve an accuracy of 79.75%. This is an improvement on previous works on emotion recognition in uncontrolled environments by 28%. Our method reconstructs faces with a facial pose and varying illumination as faces with zero facial pose and good illumination in order to get the most accurate classification. By training to deal with varying poses, we improve the accuracy of emotion classification, which aids in the progress towards real-time emotion recognition in unconstrained environments for social robots.

To extend on this work in future, considerations could be made to facial tilt, such as a participant facing upwards. Similar to pose invariance, facial tilt also results in some of a participant's face not being visible, leading to a loss in information. Future work will also explore using a Generative Adversarial Autoencoder [30] as these have demonstrated to produce remarkable reconstructions.

## References

[1] W. So, and M. Wong, C. Lam, W. Lam, A. Chui, T. Lee, H. Ng, C. Chan, and D. Fok “Using a social robot to teach gestural recognition and production in children with autism spectrum disorders,” Disability and Rehabilitation: Assistive Technology, vol.13, pp. 527–539, 2018.

[2] B. Ahn, H. Ahn, C. Sutherland, J. Lim and B. Macdonald, Development and Evaluation for Human-Care Scenario using Social Robots, 2018.

[3] H. Farid, “Blind inverse gamma correction,” IEEE Transactions on Image Processing, vol. 10, pp 1428–1433, 2001.

[4] K. Elissa, Toward Socially Assistive Robotics for Augmenting Interventions for Children with Autism Spectrum Disorders, Springer, Berlin, Heidelberg, 2009.

[5] B. Rechel, E. Grundy, J. Robine, J. Cylus, J. Mackenbach, C. Knai and M. McKee, “Ageing in the European Union,” Lancet, vol. 381, pp. 1312–1322, 2013.

[6] M.Marsella, European Union funded research into robotics for ageing well - Digital Single Market.

[7] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng and Y. Qiao, Group emotion recognition with individual facial emotion CNNs and global image based CNNs, vol.17, pp. 549–552, 2017.

[8] P. Burkert, F. Trier, M. Afzal, A. Dengel and M. Liwicki, DeXpression: Deep Convolutional Neural Network for Expression Recognition, 2015.

[9] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I, Matthews, “The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010.

[10] M. Valstar and M. Pantic, Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database, 2017.

[11] S. Kahou, C. Pal, X. Bouthillier, P. Froumenty, E. Glar Gülçehre, Ça R. Memisevic, P. Vincent, A. Courville, Y. and Bengio, Combining Modality Specific Deep Neural Networks for Emotion Recognition in Video, 2013.

[12] J. M. Susskind, A. K. Anderson and G. E. Hinton, The Toronto face database, Technical Report, 2010.

[13] Z. Yu, and C. Zhang, “Image based Static Facial Expression Recognition with Multiple Deep Network Learning”, Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15, 2015.

[14] O. Gupta, D. Raviv and R. Raskar, “Deep video gesture recognition using illumination invariants,” Neurocomputing, vol. Mar, 2016.

[15] A. Ruiz-Garcia, N. Webb, V. Palade, M. Eastwood, and M. Elshaw, ”Deep Learning for Real Time Facial Expression Recognition in Social Robots”, Proceedings of the International Conference on Neural Information Processing, pp. 392–402, 2018.

[16] F. Wallhoff, B. Schuller, M. Hawellek and G. Rigoll, ”Efficient Recognition of Authentic Dynamic Facial Expressions on the Feedtum Database”, 2006 IEEE International Conference on Multimedia and Expo, pp. 493–496, 2006.

[17] M. Lyons, J. Budynek, and S. Akamatsu, ”Automatic classification of single facial images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, pp. 1357–1362, 1999.

[18] D. Lundqvist, A. Flykt and A. Öhman, Arne ”The Karolinska Directed Emotional Faces - KDEF CD ROM from Department of Clinical Neuroscience, Psycology section”, Karolinska Institutet, pp. 3–5, 1998.

[19] X, Yin, X, Yu, K. Sohn, X, Liu, Xiaoming and M, Chandraker, Towards Large-Pose Face Frontalization in the Wild, 2017.

[20] R, Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker “Multi-PIE,” 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–8, 2008.

[21] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, ”Face recognition using Histograms of Oriented Gradients”,Pattern Recognition Letters, Vol. 32 pp. 1598–1603, 2011.

[22] D. Kingma, and J, Ba, Adam: A Method for Stochastic Optimization, 2014.

[23] Y. Bengio, ”Learning Deep Architectures for AI”, Foundations and Trends in Machine Learning, Vol. 2 pp. 1–127, 2009.

[24] A. Ruiz-Garcia, M, Elshaw, A. Altahhan and V, Palade, ”Stacked deep convolutional auto-encoders for emotion recognition from facial expressions”, 2017 International Joint Conference on Neural Networks (IJCNN), pp. 1586–1593, 2017.

[25] K.Hen, X, Zhang, S, Ren, and J. Sun, Adam: Deep Residual Learning for Image Recognition, 2015.

[26] P. Ca, L, Edu, I. Lajoie, Y. Ca, and P. Ca, ”Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion”, Journal of Machine Learning Research, Vol. 11, pp. 3371–3408, 2010.

[27] D.-U. Liu, K.-M. Lam and L.-S. Shen, ”Adaptive filter,Edge extraction,Lambertian reflectance,Optimal threshold segmentation,Principal component analysis,Ratio image”, Pattern Recognition, Vol. 38, pp. 1705–1716, 2005.

[28] C. Tosik, A. Eleyan and M. S. Salman, ”Illumination invariant face recognition system”, 2013 21st Signal Processing and Communications Applications Conference (SIU), pp. 1–4, 2013.

[29] X. Chen, X. Lan, G. Liang, J. Liu and N. Zheng, ”Pose-and-illumination-invariant face representation via a triplet-loss trained deep reconstruction model”, Multimedia Tools and Applications, Vol. 76, pp. 22043–22058, 2017.

[30] A. Makhzani, J. Shlens, N. Jaitly and I. Goodfellow, Adversarial Autoencoders, 2016.