Deep Learning for Emotion Recognition in Faces

Ariel Ruiz-Garcia(✉), Mark Elshaw, Abdulrahman Altahhan, and Vasile Palade

Faculty of Engineering, Environment and Computing, School of Computing, Electronics and Mathematics, Coventry University, Priory Street, Coventry CV1 5FB, UK
ariel.ruizgarcia@coventry.ac.uk

Abstract. Deep Learning (DL) has shown real promise for the classification efficiency for emotion recognition problems. In this paper we present experimental results for a deeply-trained model for emotion recognition through the use of facial expression images. We explore two Convolutional Neural Network (CNN) architectures that offer automatic feature extraction and representation, followed by fully connected softmax layers to classify images into seven emotions. The first architecture explores the impact of reducing the number of deep learning layers and the second splits the input images horizontally into two streams based on eye and mouth positions. The first proposed architecture produces state of the art results with an accuracy rate of 96.93 % and the second architecture with split input produces an average accuracy rate of 86.73 %, respectively.

Keywords: Deep learning · Convolution neural networks · Emotion recognition · Empathic robots

1 Introduction

It has long been suggested that emotions are an important aspect of everyday life and essential for effective human-to-human interactions [1]. There has been a growing focus on improving interaction between humans and machines by allowing this to happen in a natural manner [2]. One way to enable this natural interaction is to allow the machine to recognise the emotional state of the user, empathise with them and create appropriate responses [3]. For example, a social robot would be able to encourage a cancer patient to take their medication in a more efficient manner if it could understand the emotional state of the patient. In this work we present experimental results on emotion recognition through the use of facial expressive images, a first step towards the development of an emphatic robot.

Humans express emotions through facial expressions, therefore automated emotional recognition systems have relied on these to recognise emotions. Various intelligent techniques have been used to perform emotion recognition from

A.E.P. Villa et al. (Eds.): ICANN 2016, Part II, LNCS 9887, pp. 38–46, 2016.
DOI: 10.1007/978-3-319-44781-0_5

faces such as Hidden Markov Models [4], State Vector Machines [5] and neural networks [6,7]. We have recently seen in the development of the use of deep learning (DL) for neural networks to perform classification [8–14]. This paper will explore two architectures for Convolutional Neural Networks (CNN) to achieve deep learning classification of emotional states: Happy, Sad, Angry, Surprise, Fear, Disgust, and Neutral, from facial expressive images. These architectures will firstly, explore the impact on reducing the number of deep learning layers and secondly, the use of a novel image representation approach that splits the input images and makes use of two deep learning streams. The structure of the paper is as follows: Section 2 gives a brief description of the background on existing work; Sect. 3 describes the experimental methodology employed; Sect. 4 reports the results obtained. The succeeding section provides conclusions to the paper including future work.

2 Human Emotion Recognition - Previous Approaches

Human emotion recognition mechanisms, whether psychological or neurological, often rely on facial features to detect or recognize a specific emotion. However, the creation of a robot that can recognise emotions from images raises a number of difficulties. For example, using good quality images with enough relevant emotion-related information is often difficult due to the high computational costs imposed by big data processing and imminent changes in the environment [15]. One efficient way to overcome the former is by surveying the environment in an explorative stage and then quickly extract important features for this environment through Deep Learning, which can then be used to train a controller to achieve a specific task [15]. In order to overcome the high computational costs imposed by big visual sensory data, Altahhan [15] introduced a model that utilizes double deep learning for feature representation and action learning.

In this paper we focus on facial expression images due to the greater amount of emotion related information they contain. This approach exploits facial features such as the mouth, eyes, eyebrows and nose to classify people images as having a specific emotion. Khashman [16] proposed a neural network architecture which includes a pair of emotional neurons to account for anxiety levels. Additionally, global pattern averaging is applied in order to reduce the size of the input image over a tenfold. Khashman [16] reports an accuracy rate of 87.78 % for the proposed architecture. Another common approach to emotion recognition is making use of facial feature point localization. Sohail, and Bhattacharya [17] presented a method which includes identifying eleven different points and measuring the distances between these. This method requires reconstructing a representation of a neutral face to use as reference. Once a feature vector is obtained, this is inputted to a neural network which produces an average recognition rate of 92 %. A similar feature extraction method has been introduced by Hewahi and Baraka [18] in which they extract 28 features which describe the distances between certain points. They also consider ethnic group as an input factor while building the recognition model; a backpropagation neural network,

and have reported an accuracy rate of 83.3 %. Gabor filter is also one of the most popular methods in image-processing due to its ability to detect edges and remark salient features, and due to its resemblance to the perception in the human visual system [19]. Ahsan et al. [19] have used a combination of Gabor filter with Local Transitional Pattern together with an SVM to successfully classify facial expression images, obtaining an average accuracy rate of 95 %. Chelali and Djeradi [20] have proposed a similar approach which relies on the magnitude vector produced by Gabor filter.

Most of the approaches described above produce state of the art results for the first stage of the problem we aim to solve: emotion recognition. However, they lack the capacity to create an approach that represents and selects the salient features in an autonomous manner. This issue can conveniently be solved by employing Deep Learning (DL) techniques as done by Altahhan [15]. DL offers an outstanding alternative to prescribed feature extraction and representation. More precisely, Convolution Neural Networks (CNN) have the ability to autonomously create a vector of salient features while at the same time reducing dimensionality space by having fewer parameters than fully connected networks with the same number of layers. Levi and Hassner [8] use different image representations, including Local Binary Pattern features, as input to a number of CNN ensembles in order to boost recognition performance. Ouellet [9] presented a deep CNN to extract relevant features from still images and then classify them as seven different emotions using a Support Vector Machine. The author reports a recognition rate of 94.4 % after training with 1.2 million images. Researchers at Google Inc. have proposed a 22 layer network, omitting five pooling layers, architecture called GoogLeNet [10]. This architecture has set a state of the art benchmark for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 and has inspired a number of other architectures [10]. Another architecture for large scale recognition is proposed by Krizhevsky et al. [11]. Burkert et al. [12] presented an architecture with a pair of parallel feature extraction blocks consisting of Convolutional, Pooling, and rectified linear unit (ReLU) Layers. The authors achieved an average 99.6 % accuracy rate on the CKP dataset.

3 Methodology

3.1 Emotional Face Corpus

The emotion recognition from faces using CNN in this paper used the Karolinska directed Emotional faces database (KDEF) [21]. It contains a set with 70 individuals: 35 males and 35 females, all between 20 and 30 years old, each displaying seven different emotional expressions in five different angles. In our experiments we only use front angle images; a subset of 980 images. All images were taken under a con-trolled environment: subjects wore uniform T-Shirt colours, faces were centred with a grid, and eyes and mouths were positioned in fixed image coordinates [21]. To speed up training, face images were extracted, grey-scaled and resized to 100 by 100 as shown in Fig. 1 below. Our training set contained

Fig. 1. Sample extracted face images from the KDEF database [21]. Subject 07 displaying seven emotions: sad, surprised, neutral, happy, fear, disgust, angry.

98 randomly selected front angle images per emotion, giving us a total of 686 input samples. Our testing set contained 42 images per emotion and thus a total of 294 training samples.

3.2 Architectures

Since our aim is to explore biologically inspired neural architectures we decided to employ CNN for feature extraction and representation given that they are inspired by animal vision cortex [13]. These models are well known for their ability to extract salient features and for being faster than traditional models such as Multilayer Perceptron (MLP) networks due to a smaller number of parameters required for training. This paper explores two main architectures for the CNN to identify the number of deep learning layers that best represent the images and the impact of a split input stream representation for the architecture structure. Figure 2 illustrates a detailed description.

The architectures we propose are made up of convolution, rectified linear unit (ReLU), max pooling, and local response normalization (LRN) layers followed by one fully connected layer and one softmaxloss layer for classification. The convolutional layers incorporate constraints and achieve some degree of shift and deformation invariance using local receptive fields, shared weights, and spatial subsampling [13]. Their output can be summarized as:

$$C(x_{u,v}) = (x+a)^n = \sum_{i=-\frac{n}{2}}^{\frac{n}{2}} \sum_{j=-\frac{m}{2}}^{\frac{m}{2}} f_k(i,j) x_{u-i,v-j}. \tag{1}$$

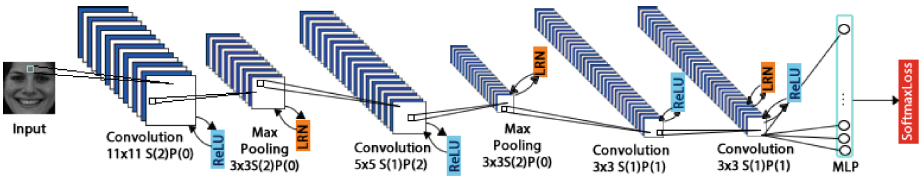
where f_k is the filter with a kernel size $n \times m$, applied to the input x. In our models n is always the same as m. The convolutional layers in the first network use 60, 90, 120, and 240 filters respectively. Whereas the split input model learns 60, 90, and 120 filters. Given that ReLU functions marginally reduce training times in deep convolutional networks [11], every output of a convolutional layer in our models is shaped by a ReLU function. Given an input value x, ReLU output is given by:

$$f(x) = \max\{0, x\}. \tag{2}$$

The input is further reduced with max pooling layers. Let x_i be the input and m be the size of the filter, then the output of the max pooling layers is calculated as:

$$M(x_i) = \max\left\{x_{i+k,i+l} \middle| |k| \leq \frac{m}{2}, |l| \leq \frac{m}{2} k, l \in \mathbb{N}\right\}. \tag{3}$$

a)



b)

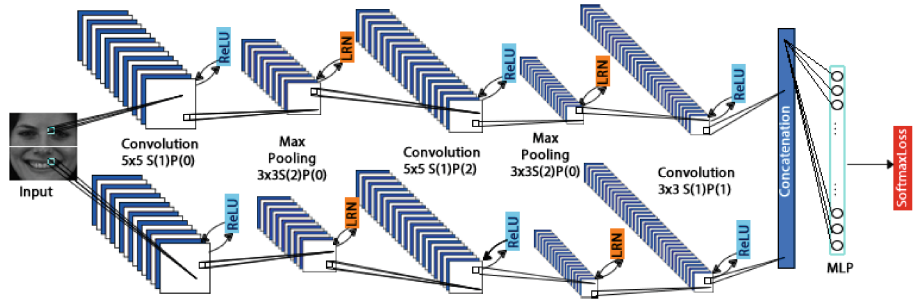


Fig. 2. (a) CNN with reduced deep learning layer, (b) Network with split input; S defines the stride size and P the padding. Face images from subject 07 in KDEF database [21].

Moreover, all spatial locations of the output of some of the pooling layers and ReLU layers are uniformly normalized using the Local Response Normalization (LRN) operator. Let k be the output channel, and $G(k) \subset \{1, 2, , D\}$ represent a corresponding subset of input channels, the output of LRN is calculated as follows:

$$y_{ijk} = x_{ijkz} \left(k + \alpha \sum_{t \in G(k)} x_{ijt}^2 \right)^{-\beta} . \tag{4}$$

Furthermore, our models use a fully connected layer which in term is an MLP. Let σ represent a sigmoid activation function, then the output of the hidden layer is computed by:

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

Finally, the last layer in our models employs a *softmaxloss* operator which in turn is a combination of the softmax operator followed by the log-loss operator. Given the class ground-truth c, softmaxloss output is computed by:

$$y = -\sum_{ij} \left(x_{ijc} - \log \sum_{d=1}^{D} e^{x_{ijd}} \right). \tag{6}$$

The training process for both architectures was the same: the learning rate for filters and biases was initially set to 1.0 and dynamically adjusted down to 0.00001

over 1000 epochs, whereas the momentum was set to 0.9. The input vector was down-sampled by convolution and pooling layers using a sliding window with stride of 2.

4 Results and Discussion

After training for 15,500, epochs the first model achieved its best performance producing an accuracy rate of 96.93 % on the testing set, not far from the results obtained by Burkert et al. [12]. Further training with the same parameters seems to cause overfitting. The second architecture proposed splits the image horizontally in half and feeds each half to a corresponding sub architecture to be processed in parallel. Each sub-architecture learns a representation of different face parts: in the case of the first half the salient features highlighted are the areas around the eyes whereas the second half highlights the area surrounding the mouth. The translation invariant features obtained from each subnetwork are then recombined for classification. This model with split input has been training for just 5,280 epochs and has already achieved state of the art performance with an accuracy rate of 86.73 %. Table 1 illustrates the confusion matrices for both models; as it can be observed both networks achieved a higher performance rate when classifying facial images illustrating happy emotions. It is evident that both of our models misclassify neutral faces the most. This might be due to the similarity of this emotion with all the others, especially with sadness. As it can be observed in Fig. 1 above there is not a big difference between these two expressions and neutral has previously been defined as the basic human emotion [17] which implies that all other emotions are developed from this.

Table 1. Left: first network confusion matrix, right: split input network confusion matrix. A: angry; D: disgust; F: fear; H: happy; N: neutral; Sa: sad; Su: surprised.

	A	*D*	*F*	*H*	*N*	*Sa*	*Su*
A	42	0	0	0	0	0	0
D	0	40	0	1	0	1	0
F	0	1	39	1	0	0	1
H	0	0	0	42	0	0	0
N	0	1	0	0	39	2	0
Sa	0	1	0	0	0	41	0
Su	0	0	1	0	0	0	41

	A	*D*	*F*	*H*	*N*	*Sa*	*Su*
A	40	1	1	0	0	0	0
D	1	32	1	3	0	3	1
F	1	2	37	0	0	0	2
H	2	0	0	42	0	0	0
N	1	2	1	0	32	5	1
Sa	0	3	2	1	0	36	0
Su	0	0	3	0	0	0	39

We have explored in this paper two new CNN architectures that create state of the art results. Our first architecture has achieved such performance with a reduced number of layers as opposed to the model proposed by [12]. Although second model, which uses two deep learning streams, has produced lower performance it has only been training for a fraction of the time that the first one was trained for. Moreover, this network has already outperformed the performance of the first model at 5000 epochs. We attribute this increase in performance to the

split input around the mouth and eye areas; since these two are determining key factors for emotion recognition, each network ensemble learns to extract only of these salient features, thus having to do lesser weight modifications.

Beyond the neuroscientific and biological aspect, human emotions allow us to connect and share experiences with other people regardless of background. This cognitive process is vital for human-human interactions and could improve human-robot interactions. Our research aims to contribute to solving this issue by proposing a neural architecture that can allow a robotic machine to recognise a user's emotional state. To this day, empirical models such as Support Vector Machines seem to be the dominant classifiers in emotion recognition through facial expression images due to their high performance rate. However, the performance of these classifiers heavily relies on the image preprocessing techniques applied on the images. CNN, on the other hand, have the ability to extract and learn features autonomously. Our second architecture contains similar properties to that proposed by [12], however our model uses less parameters and less layers, being marginally faster and therefore more suitable for online learning.

5 Conclusion

To the best of our knowledge we are the first to propose an architecture for emotion recognition which splits the image into two sections in order to extract features with different parameters. This approach uses two network ensembles to extract salient features from around the mouth and eye areas. The model seems to take advantage of the most salient features, which are essential for emotion recognition. This approach has produced promising results and will therefore be improved in future work.

We hope that our research brings social robots a step closer to been fully accepted by society. The results reported above illustrate a fundamental initial step towards achieving this goal by providing us with a method for self-organised or autonomous feature extraction and representation learnt explicitly for emotion recognition. However, despite the high recognition performance achieved in the experiments we conducted, we have to take into consideration the fact that the training and testing datasets contain images of similar quality and taken under controlled environments. Future work will address the ability of the architectures developed to compensate for light and angle variations. In this manner, the number of layers and parameters will be adjusted accordingly in order to achieve the same performance results in real environments. Given the promise shown by the architecture with a split input, future work will look at its performance with the input being split into more sections or with random patches.

Future work will look at the development of an associative architecture to be combined with the proposed CNN. Additionally, future work will explore the possibility of using a multimodal approach and incorporate other inputs such as: speech signals, body language, heart rate readings, etc. into our model in order to obtain a comprehensive representation of the emotions and better recognition rates. Reinforcement learning techniques will be explored to allow the robot to learn which responses improve the interaction process with the user.

References

1. Lewis, M., Haviland-Jones, J., Barrett, L.: Handbook of Emotions. Guilford Press, New York (2008)
2. Chavhan, A., Chavan, S., Dahe, S., Chibhade, S.: A neural network approach for real time emotion recognition. IJARCCE **4**(3), 259–263 (2015)
3. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Interspeech, pp. 223–227 (2014)
4. Cohen, I., Garg, A., Huang, T.: Emotion recognition from facial expressions using multi-level HMM. In: Neural Information Processing Systems, vol. 2 (2000)
5. Sarnarawickrame, K., Mindya, S.: 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 (2013)
6. Boughrara, H., Chtourou, M., Ben Amar, C., Chen, L.: Facial expression recognition based on a mlp neural network using constructive training algorithm. Multimed. Tools Appl. **75**, 709–731 (2014)
7. Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI 2015), pp. 467–474 (2015)
8. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI 2015), pp. 503–510 (2015)
9. Ouellet, S.: Realtime emotion recognition for gaming using deep convolutional network features. CoRR. abs/1408.3750 (2014)
10. Szegedy, C., Lui, W., Jia, Y., Sermanet, P., Reed, S., Auguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–19 (2014)
11. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1106–1114 (2012)
12. Burkert, P., Trier, F., Afzal, M.Z., Dengel, A., Liwicki, M.: DeXpression: Deep Convolutional Neural Network for Expression Recognition. CoRR. abs/1509.05371 (2015)
13. Lawrence, S., Giles, C., Tsoi, A.C., Back, A.: Face recognition: a convolutional neural network approach. IEEE Trans. Neural Netw. **8**, 98–113 (1997)
14. Brosch, T., Tam, R.: Efficient training of convolutional deep belief networks in the frequency domain for application to high-resolution 2D and 3D images. Neural Computation. **27**, 211–227 (2015)
15. Altahhan, A.: Navigating a robot through big visual sensory data. Procedia Comput. Sci. **53**, 478–485 (2015)
16. Khashman, A.: Application of an emotional neural network to facial recognition. Neural Comput. Appl. **18**, 309–320 (2008)
17. Sohail, A., Bhattacharya, P.: Classifying facial expressions using level set method based lip contour detection and multi-class support vector machines. Int. J. Pattern Recogn. Artif. Intell. **25**, 835–862 (2011)
18. Hewahi, N., Baraka, A.: Impact of ethnic group on human emotion recognition using backpropagation neural network. Broad Res. Artif. Intell. Neurosci. **2**, 20–27 (2011)

19. Ahsan, T., Jabid, T., Chong, U.: Facial expression recognition using local transitional pattern on gabor filtered facial images. IETE Tech Rev. **30**, 47 (2013)
20. Chelali, F., Djeradi, A.: Face recognition using MLP and RBF neural network with Gabor and discrete wavelet transform characterization: a comparative study. Math. Prob. Eng. **2015**, 116 (2015)
21. Lundqvist, D., Flykt, A., Ahman, A.: The Karolinska Directed Emotional Faces - KDEF. CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet (1998). ISBN 91-630-7164-9