Emotion Recognition using Deep Convolutional Neural Networks

Enrique Correa Arnoud Jonker Michaël Ozo Rob Stolk

June 30, 2016

A key step in the humanization of robotics is the ability to classify the emotion of the human operator. In this paper we present the design of an artificially intelligent system capable of emotion recognition trough facial expressions. Three promising neural network architectures are customized, trained, and subjected to various classification tasks, after which the best performing network is further optimized. The applicability of the final model is portrayed in a live video application that can instantaneously return the emotion of the user.

1 Introduction

Ever since computers were developed, scientists and engineers thought of artificially intelligent systems that that are mentally and/or physically equivalent to humans. In the past decades, the increase of generally available computational power provided a helping hand for developing fast learning machines, whereas the internet supplied an enormous amount of data for training. These two developments boosted the research on smart self-learning systems, with neural networks among the most promising techniques.

1.1 Backgrounds

One of the current top applications of artificial intelligence using neural networks is the **recognition of faces** in photos and videos. Most techniques process visual data and search for general patterns present in human faces. Face recognition can be used for surveillance purposes by law enforcers as well as in crowd management. Other present-day applications involve automatic blurring of faces on Google Streetview footage and automatic recognition of Facebook friends in photos.

An even more advanced development in this field is **emotion recognition**. In addition to only identifying faces, the computer uses the arrangement and shape of e.g. eyebrows and lips to determine the facial expression and hence the emotion of a person. One possible application for this lies in the area of surveillance and behavioural analysis by law enforcement. Furthermore such techniques are used in digital cameras to automatically take pictures when the user smiles. However, the most promising applications involve the humanization of artificial intelligent systems. If computers are able to keep track of the mental state of the user, robots can react upon this and behave appropriately. Emotion recognition therefore plays a key-role in improving human-machine interaction.

1.2 Research objective

In this research we mainly focus on neural network based artificially intelligent systems capable of deriving the emotion of a person through pictures of his or her face. Different approaches from existing literature will be experimented with and the results of various choices in the design process will be evaluated. The main research question therefore reads as follows: **How can an artificial neural network be used for interpreting the facial expression of a human?**

The remainder of this article describes the several steps taken to answer the main research question, i.e. the sub-questions. In section 2, a literature survey will clarify *what the role of facial expressions is*

in emotion recognition and *what types of networks are suitable for automated image classification.* The third section explains *how the neural networks under consideration are structured* and *how the networks are trained.* Section 4 describes *how the final model performs* after which a conclusion and some recommendations follow in the last section. It may be noted that the aim of our work is not to design an emotion recognizer from scratch but rather to review design choices and enhance existing techniques with some new ideas.

2 Literature review

For the development of a system that is able to recognize emotions through facial expressions, previous research on the way humans reveal emotions as well as the theory of automatic image categorization is reviewed. In the first part of this section, the role of interpreting facial expressions in emotion recognition will be discussed. The latter part surveys previous studies on automatic image classification.

2.1 Human emotions

A key feature in human interaction is the universality of facial expressions and body language. Already in the nineteenth century, Charles Darwin published upon globally shared facial expressions that play an important role in non-verbal communication [3]. In 1971, Ekman & Friesen declared that *facial behaviors are universally associated with particular emotions* [5]. Apparently humans, but also animals, develop similar muscular movements belonging to a certain mental state, despite their place of birth, race, education, etcetera. Hence, if properly modelled, this universality can be a very convenient feature in human-machine interaction: a well trained system can understand emotions, independent of who the subject is.

One should keep in mind that facial expressions are not necessarily directly translatable into emotions, nor vice versa. Facial expression is additionally a function of e.g. mental state, while emotions are also expressed via body language and voice [6]. More elaborate emotion recognition systems should therefore also include these latter two contributions. However, this is out of the scope of this research and will remain a recommendation for future work. Readers interested in research on emotion classification via speech recognition are referred to Nicholson et al. [14]. As a final point of attention, emotions should not be confused with mood, since mood is considered to be a long-term mental state. Accordingly, mood recognition often involves longstanding analysis of someone's behaviour and expressions, and will therefore be omitted in this work.

2.2 Image classification techniques

The growth of available computational power on consumer computers in the beginning of the twenty-first century gave a boost to the development of algorithms used for interpreting pictures. In the field of image classification, two starting points can be distinguished. On the one hand pre-programmed feature extractors can be used to analytically break down several elements in the picture in order to categorize the object shown. Directly opposed to this approach, self-learning neural networks provide a form of 'black-box' identification technique. In the latter concept, the system itself develops rules for object classification by training upon labelled sample data.

An extensive overview of analytical feature extractors and neural network approaches for facial expression recognition is given by Fasel and Luettin [6]. It can be concluded that by the time of writing, at the beginning of the twenty-first century, both approaches work approximately equally well. However, given the current availability of training data and computational power it is the expectation that the performance of neural network based models can be significantly improved by now. Some recent achievements will be listed below.

(i) A breakthrough publication on automatic image classification in general is given by Krizhevsky and Hinton [9]. This work shows a deep neural network that resembles the functionality of the human visual cortex. Using a self-developed labelled collection of 60000 images over 10 classes, called the CIFAR-10 dataset, a model to categorize objects from pictures is obtained. Another important outcome of the research is the visualization of the filters in the network, such that it can be assessed how the model breaks down the

pictures.

(ii) In another work which adopts the CIFAR-10 dataset [2], a very wide and deep network architecture is developed, combined with GPU support to decrease training time. On popular datasets, such as the MNIST handwritten digits, Chinese characters, and the CIFAR-10 images, near-human performance is achieved. The extremely low error rates beat prior state-of-the-art results significantly. However it has to be mentioned that the network used for the CIFAR-10 dataset consists of 4 convolutional layers with 300 maps each, 3 max pooling layers, and 3 fully connected output layers. As a result, although a GPU was used, the training time was several days.

(iii) In 2010, the introduction of the yearly Imagenet challenge [4] boosted the research on image classification and the belonging gigantic set of labelled data is often used in publications ever since. In a later work of Krizhevsky et al. [10], a network with 5 convolutional, 3 max pooling, and 3 fully connected layers is trained with 1.2 million high resolution images from the ImageNet LSVRC-2010 contest. After implementing techniques to reduce overfitting, the results are promising compared to previous state-of-the-are models. Furthermore, experiments are done with lowering the network size, stating that the number of layers can be significantly reduced while the performance drops only a little.

(iv) With respect to facial expression recognition in particular, Lv et al. [13] present a deep belief network specifically for use with the Japanese Female Facial Expression (JAFFE) and extended Cohn-Kanade (CK+) databases. The most notable feature of the network is the hierarchical face parsing concept, i.e. the image is passed through the network several times to first detect the face, thereafter the eyes, nose, and mouth, and finally the belonging emotion. The results are comparable with the accuracy obtained by other methods on the same database, such as Support Vector Machine (SVM) and Learning Vector Quantization (LVQ).

(v) Another work on the Cohn-Kanade database [1] makes use of Gabor filtering for image processing and Support Vector Machine (SVM) for classification. A Gabor filter is particularly suitable for pattern recognition in images and is claimed to mimic the function of the human visual system. The emotion recognition accuracies are high, varying from 88% on anger to 100% on surprised. A big disadvantage of the approach however is that very precise pre-processing of the data is required, such that every image complies to a strict format before feeding it into the classifier.

(vi) One of the most recent studies on emotion recognition describes a neural network able to recognize race, age, gender, and emotion from pictures of faces [7]. The dataset used for the latter category is originating from the Facial Expression Recognition Challenge (FERC-2013). A clearly organized deep network consisting of 3 convolutional layers, 1 fully connected layer, and some small layers in between obtained an average accuracy of 67% on emotion classification, which is equal to previous state-of-the-art publications on the same dataset. Furthermore this thesis lays down a valuable analysis of the effect of adjusting the network size, pooling, and dropout.

Underlined by some other literature, the most promising concept for facial expression analysis is the use of deep convolutional neural networks. However, the network from [2] (ii) is considered to be too heavy for our limited amount of available processing resources. The original network from [10] (iii) is large as well, but smaller versions are claimed to be equally suitable. Furthermore, due to their somewhat analytical and unconventional approaches, we will not evaluate [13] (iv) and [1] (v). Hence, in the next section, three deep architectures in total will be subjected to an emotion classification problem. These architectures are derived from, but not necessarily equal to, the networks described at items i, iii, and vi.

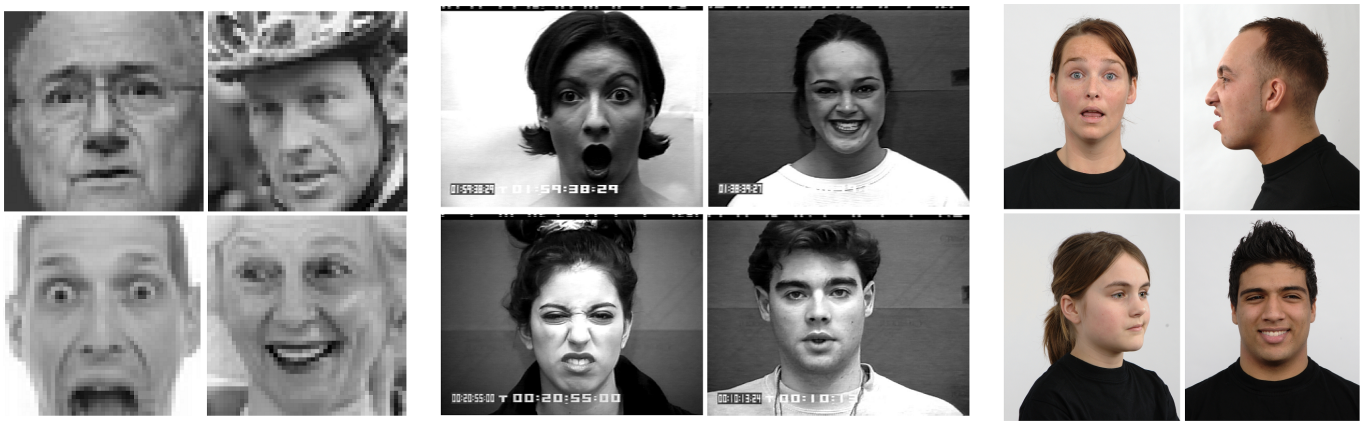

Figure 1: Samples from the FERC-2013 (left), CK+ (center), and RaFD (right) datasets

3 Experiment setup

To assess the three approaches mentioned previously on their capability of emotion recognition, we developed three networks based on the concepts from [9], [10], and [7]. This section describes the data used for training and testing, explains the details of each network, and evaluates the results obtained with all three models.

3.1 Dataset

Neural networks, and deep networks in particular, are known for their need for large amounts of training data. Moreover, the choice of images used for training are responsible for a big part of the performance of the eventual model. This implies the need for a both high qualitative and quantitative dataset. For emotion recognition, several datasets are available for research, varying from a few hundred high resolution photos to tens of thousands smaller images. The three we will discuss are the Facial Expression Recognition Challenge (FERC-2013) [8], Extended Cohn-Kanade (CK+) [12], and Radboud Faces Database (RaFD) [11], all shown in figure 1.

The datasets differ mainly on quantity, quality, and 'cleanness' of the images. The FERC-2013 set for example has about 32000 low resolution images, where the RaFD provides 8000 high resolution photos. Furthermore it can be noticed that the facial expressions in the CK+ and RaFD are posed (i.e. 'clean'), while the FERC-2013 set shows emotions 'in the wild'. This makes the pictures from the FERC-2013 set harder to interpret, but given the large size of the dataset, the diversity can be beneficial for the robustness of a model.

We reason that, once trained upon the FERC-2013 set, images from 'clean' datasets can easily be classified, but not vice versa. Hence for the three networks under consideration, training will be done using 9000 samples from the FER-2013 data (see figure 2) with another 1000 new samples for validation. Subsequently testing will be done with 1000 images from the RaFD set to get an indication of performance on clean high quality data. This latter set has an even distribution over all emotions.

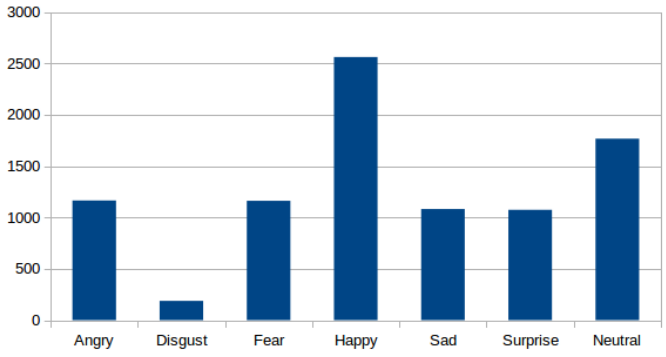

Figure 2: Number of images per emotion in the training set

Please note that non-frontal faces and pictures with the label contemptuous are taken out of the RaFD data, since these are not represented in the FERC-2013 training set. Furthermore, with use of the Haar Feature-Based Cascaded Classifier inside

the OpenCV framework [15], all data is preprocessed. For every image, only the square part containing the face is taken, rescaled, and converted to an array with 48x48 grey-scale values.

3.2 Networks

The networks are programmed with use of the TFLearn library on top of TensorFlow, running on Python. This environment lowers the complexity of the code, since only the neuron layers have to be created, instead of every neuron. The program also provides real-time feedback on training progress and accuracy, and makes it easy to save and reuse the model after training. More details on this framework can be found in reference [16].

(A) The first network to test is based on the previously described research by Krizhevsky and Hinton [9]. This is the smallest network of the three, which means that it has the lowest computational demands. Since one of the future applications might be in the form of live emotion recognition in embedded systems, fast working algorithms are benificial.

The network consists of three convolutional layers and two fully connected layers, combined with maxpooling layers for reducing the image size and a dropout layer to reduce the chance of over fitting. The hyper parameters are chosen such that the number of calculations in each convolutional layer remains roughly the same. This ensures that information is preserved throughout the network. Training is performed using different numbers of convolutional filters to evaluate their effect on the performance.

(B) In 2012, the AlexNet convolutional network was developed for classifying images in more than 1000 different classes, using 1.2 million sample pictures from the ImageNet dataset. Due to the fact that in this research the model only has to distinguish seven emotions, and due to our limited computing resources, the size of the original network is considered to be too large.

Hence, instead of 5 convolutional layers we applied 3, and in the subsequent 3 fully connected layers the number of nodes of each fully connected was reduced from 4096 to 1024. While the original network was divided for parallel training, it was observed that was not necessary for the smaller version. The network also makes use of local normalization to speed up the training and dropout layers in order to reduce the overfitting.

(C) The last experiments are performed on a network based by the work of Gudi [7]. Since this research also aimed on recognizing 7 emotions using the FERC-2013 dataset, the architecture should be a good starting point for our research.

The original network starts with an input layer of 48 by 48, matching the size of the input data. This layer is followed by one convolutional layer, a local contrast normalization layer, and a maxpooling layer respectively. The network is finished with two more convolutional layers and one fully connected layer, connected to a softmax output layer. Dropout was applied to the fully connected layer and all layer contain ReLu units.

For our research, a second maxpooling layer is applied to reduce the number of parameters. This lowers the computational intensity of the network, while the reduction in performance is claimed to be only 1-2%. Furthermore the learning rate is adjusted. Instead of linearly decreasing the learning rate as done by Gudi [7], we believe a learning rate which makes use of momentum would converge faster, as the momentum increases the learning rate when the gradient keeps going in the same direction.

3.3 Evaluation

All networks are trained for 60 epochs with the data mentioned in section 3.1. Figure 3 and table 1 show various details of the training process and the final model. For network A, the final accuracy on the validation data is around 63%. Already after 10 epochs, the accuracy raised above 60%, indicating quick learning capabilities. Furthermore it is noteworthy that adjusting the filter dimension did not have a big influence on the accuracy, though it has

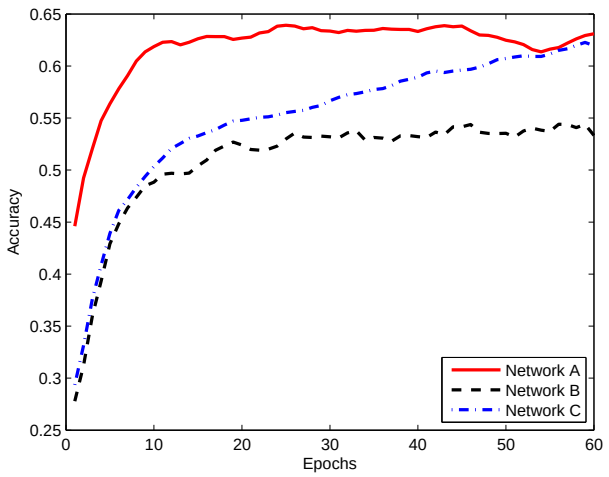

Figure 3: Accuracy on the validation set during training epochs

on the processing time. This means that fast models can be made with very reasonable performance.

Table 1: Details of the trained networks

Network	Accuracy		Size
	Validation	RaFD	
A	63%	50%	Small
B	53%	46%	Large
C	63%	60%	Medium

Surprisingly, the second, much larger, network learns quickly as well, but converges to an accuracy of about 54%. Apparently reducing the network size breaks down the promising performance of the original network more than expected. Together with the much higher computational intensity, and therefore slower live performance, this model is not a worthy challenger of the other two architectures.

Network C shows a somewhat slower learning curve, but the final accuracy on the validation set is similar to that of network A. The processing demands are in between that of the other networks, so based on this fact, network A seems to be the most promising approach for our emotion recognition task. However, the performance of network C on the extra RaFD testset is significantly better (60%) than that of network A (50%). This indicates better generalizing capabilities, which is very important for future applications. Hence, in the next chapter, the model from network C will be further investigated and tested.

4 Final model

The last described network from section 3.2 was observed to have the most promising performance for practical applications. An overview of its architecture is shown in figure 4. The source files for this network, as well as other scripts used for this project can be found on *https://github.com/isseu/emotion-recognition-neural-networks*.

As can be seen from figure 3, the accuracy seems to still increase in the last epochs. We therefore will train the network for 100 epochs in the final run, to make sure the accuracy converges to the optimum. In an attempt to improve the final model even more, the network will be trained on a larger set than the one described previously. Instead of 9000 pictures, training will be done with 20000 pictures from the FERC-2013 dataset. The ratios of the emotions present in this set are given in figure 5. Newly composed validation (2000 images) and test sets (1000 images) from the FERC-2013 dataset are used as well, together with the well-balanced RaFD test set from the previous experiment.

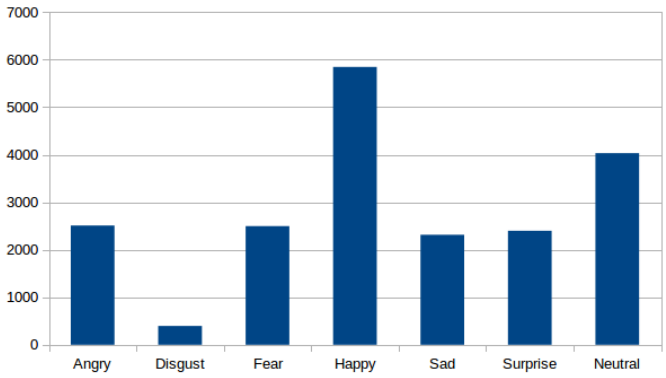

Figure 5: Number of images per emotion in the final training set

4.1 Test results

The accuracy rates of the final model are given in table 2. On all validation and test sets the accuracy was higher than during previous runs, underlining that

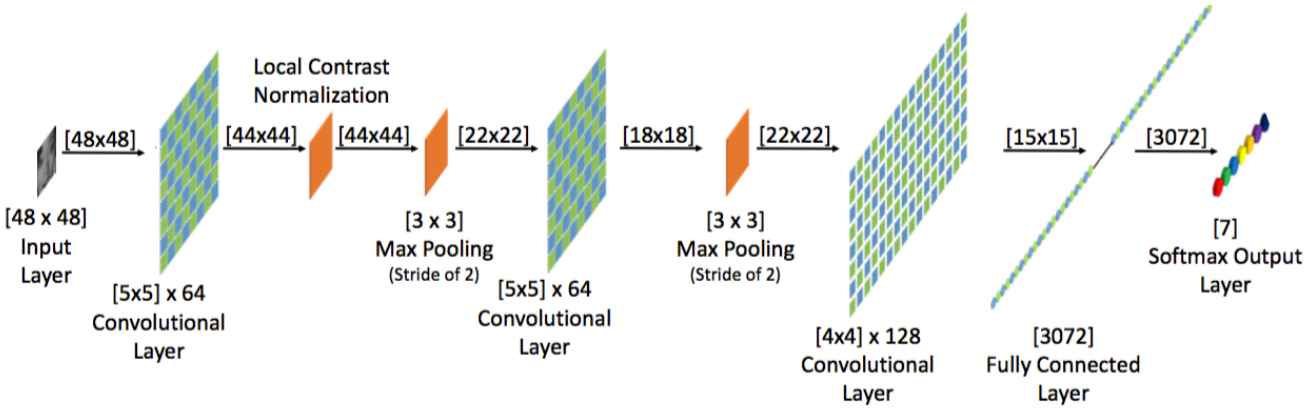

Figure 4: Overview of the network architecture of the final model

more data and longer training can improve the performance of a network. Given that the state-of-the-art networks from previous research obtained about 67% on test sets, and keeping in mind our limited resources, the results are in fact pretty good. Notable is the accuracy on the RaFD test set, which contains completely different pictures than the training data. This illustrates the powerful generalizing capabilities of this final model.

Table 2: Accuracy of the networks

Network	FERC-2013		RaFD
	Validation	Test	
A	63%		50%
B	53%		46%
C	63%		60%
Final	66%	63%	71%

To see how the model performs per emotion, a table is generated, depicted in figure 6. Very high accuracy rates are obtained on happy (90%), neutral (80%), and surprised (77%). These are in fact the most distinguishable facial expressions according to humans as well. Sad, fearful, and angry are often misclassified as neutral too though. Apparently these emotions very look alike. The lowest accuracy is obtained on sad (28%) and fearful (37%). Finally it is noteworthy that even though the percentage of data with label disgusted in the training set is low, the classification rate is very reasonable. In general the main diagonal, showing the right classification, can be distinguished clearly.

4.2 Live application

As is already mentioned, live emotion recognition through video is one of the most important key-points in human-machine interaction. To show the capabilities of the obtained network, an application is developed that can directly process webcam footage through the final model.

With use of the aforementioned OpenCV face recognition program [15], the biggest appearing face from real-time video is tracked, extracted, and scaled to usable 48x48 input. This data is then fed to the input of the neural network model, which in its turn returns the values of the output layer. These values represent the likelihood that the each emotion is depicted by the user. The output with the highest value is assumed to be the current emotion of the user, and is depicted by an emoticon on the left of the screen. Figures 7 to 13 shows the live application and the interaction with the authors of this research and with live television.

Although it is hard to objectively assess, the live application shows promising performance. Though, it encounters problems when shadows are present on the face of the subject. All emotions are easily recognized when acted by the user, and when pointing the camera on the television, most emotions in the wild can be classified. This once again emphasizes the power of using neural network based models for future applications in emotion recognition.

5 Conclusion

The conclusion and recommendations of this research will be delivered by each group member individually.

References

[1] T. Ahsan, T. Jabid, and U.-P. Chong. Facial expression recognition using local transitional pattern on gabor filtered facial images. *IETE Technical Review*, 30(1):47–52, 2013.

[2] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

[3] C. R. Darwin. *The expression of the emotions in man and animals*. John Murray, London, 1872.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[5] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[6] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[7] A. Gudi. Recognizing semantic features in faces using deep learning. *arXiv preprint arXiv:1512.00743*, 2015.

[8] Kaggle. Challenges in representation learning: Facial expression recognition challenge, 2013.

[9] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, 2009.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[12] P. Lucey, J. F. 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. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[13] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In *Smart Computing (SMARTCOMP), 2014 International Conference on*, pages 303–308. IEEE, 2014.

[14] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. *Neural computing & applications*, 9(4): 290–296, 2000.

[15] OpenSourceComputerVision. Face detection using haar cascades. URL `http://docs.opencv.org/master/d7/d8b/tutorial_py_face_detection.html`.

[16] TFlearn. Tflearn: Deep learning library featuring a higher-level api for tensorflow. URL `http://tflearn.org/`.

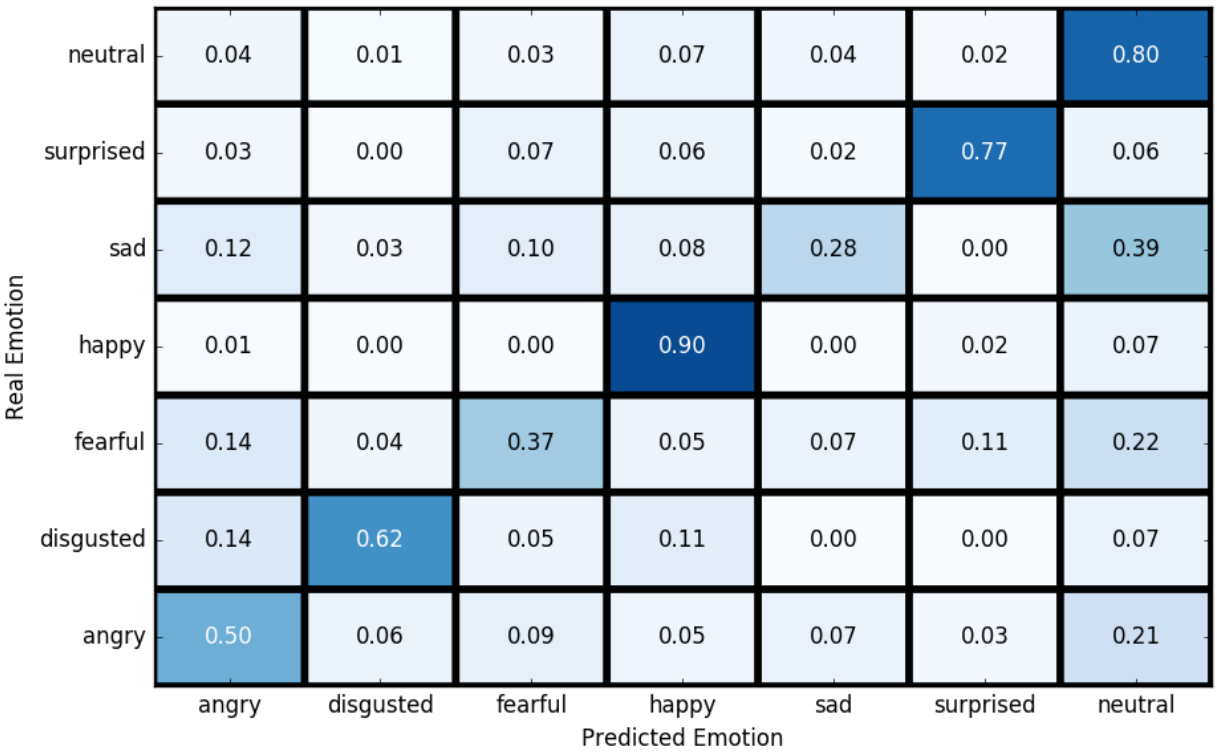

Figure 6: Performance matrix of the final model. Vertically the input, horizontally the output.

Figure 7: Screenshot of the live app

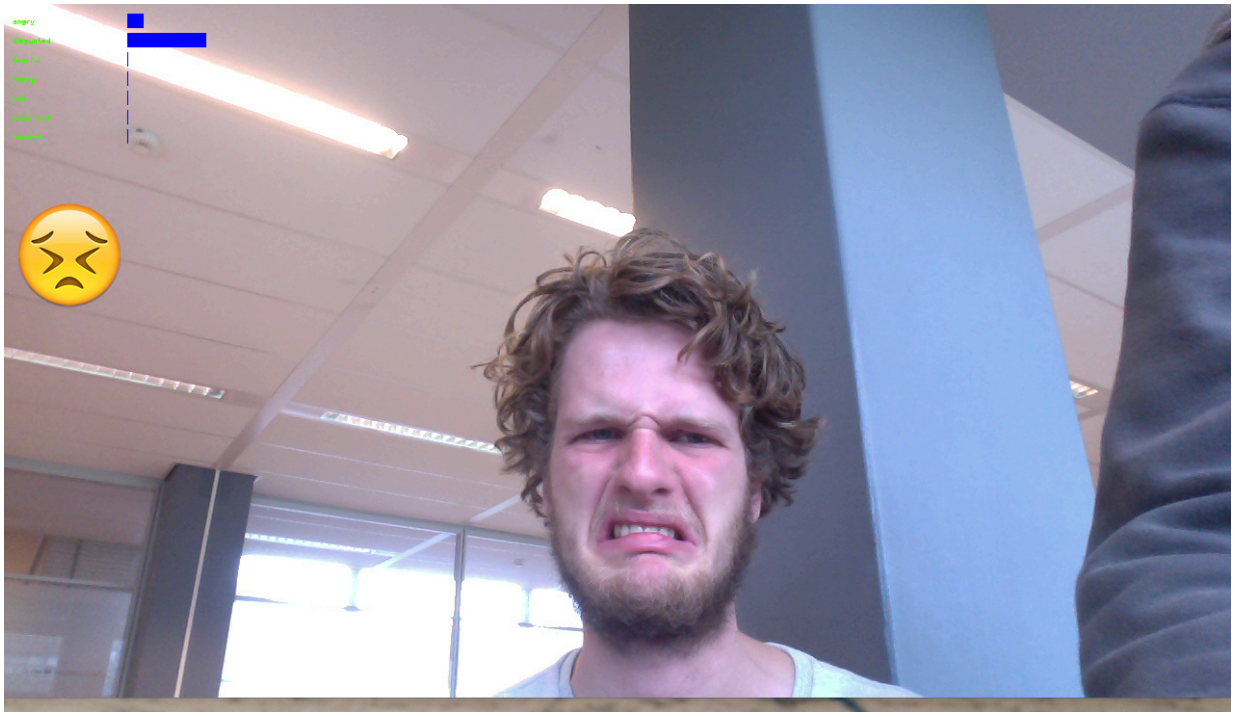

Figure 8: Screenshot of the live app

Figure 9: Screenshot of the live app

Figure 10: Screenshot of the live app

Figure 11: Screenshot of the live app

Figure 12: Screenshot of the live app

Figure 13: Screenshot of the live app