Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-15 (2021) 299–303

# Image Pre-processing Significance on Regions of Impact in a Trained Network for Facial Emotion Recognition

**H. Arabian*. V. Wagner-Hartl.****
**J. Geoffrey Chase***. K. Möller***

**Institute for Technical Medicine (ITeM), VS-Schwenningen 78054, Germany (Tel: 7720-307-4390; e-mail: H.Arabian@hs-furtwangen.de).*
***Department of Industrial Technologies, Hochschule Furtwangen University, 78532 Tuttlingen, Germany.*
**** Center for Bioengineering, Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand}*

**Abstract:** Facial emotion recognition (FER) has gained interest and focus over the years. It can be useful in many different applications and could offer significant benefit as part of feedback systems to help train children with Autism Spectrum Disorder (ASD) who struggle to recognize facial expressions and emotions. This paper explores the effectiveness and significance of image pre-processing in Neural Networks on developing suitable models for classification. Transfer Learning using the popular "AlexNet" architecture was used in the development of the model with three different approaches for image inputs. Model performance was compared using accuracy of randomly selected validation set after training on a different random training set from the Oulu-CASIA database and visualizations of predicted areas of importance analyzed. Image classes were distributed evenly, and accuracies of up to 99.90% were observed with small variation between approaches but significant difference in regions of impact. The visualization process highlighted the importance of image pre-processing prior to network training to improve accuracy and eventual efficacy for this application in ASD.



*Keywords:* Autism Spectrum Disorder (ASD), Facial Emotion Recognition (FER), AlexNet, Oulu-CASIA, Network Prediction Visualization.

## 1. INTRODUCTION

Emotion recognition has gained significant interest and growth over the past few years. Facial emotion recognition (FER) can be used for treatment of Autism Spectrum Disorder (ASD) in children, a developmental brain disorder impairing social interaction, communication, behaviours, and interests of individuals (American Academy of Pediatrics, 2001). Estimates reveal 1 out of 59 individuals are affected by ASD (Rylaarsdam, and Guemez-Gamboa, 2019).

Children suffering from ASD are accustomed to a certain routine and any deviation from normalcy can cause psychological and emotional challenges to the child, as well as increased stress levels for the caregiver (Lugo-Marín et al, 2021). All these issues limit inter-personal and educational opportunities, as well as having long term economic consequences for the individual and society. An individually adjusted virtual world combined with a reward system in the form of a gaming platform, and the technical affinity of most ASD children creates a suitable atmosphere for a gamified approach to treatment, which is typically defined by intensive individual and small group interaction sessions with trained facilitators and educators.

When delivering a certain message, 55% of its efficacy is based on the facial component (Mehrabian, 2017). Machine Learning techniques are required to transform an image of a face into a corresponding emotion. In previous work (Arabian et al, 2021) the importance of choosing local regions of interest within an image where discussed using traditional machine learning techniques of feature extraction and classification models. In this paper the use of a Convolution Neural Network (CNN) for FER modelling is studied for its effectiveness and to analyse its important regions for decision making. The popular pretrained Neural Network "Alex Net" (Krizhevsky et al, 2017) is used for this study.

This study takes three different image input approaches for training. The Oulu-CASIA (Zhao et al, 2011) database is used to test the performance of the different models. The trained networks are analysed to assess the accuracy of its classifications and its predictions are visualized to study the regions of relative importance chosen by the model for FER.

The aim of this study is to highlight the significance of image pre-processing in Deep Neural Network models for FER to improve training, overall accuracy and efficacy.

## 2. SYSTEM DESCRIPTION

### *2.1 Methodology*

The three approaches for image inputs to the network are: 1) Approach 1, the original image as is in the wild with no image pre-processing performed; Approach 2, segmentation


10.1016/j.ifacol.2021.10.272

of the Face using Viola-Jones (Viola and Jones, 2001) object detection algorithm; and Approach 3 using the red to green color intensity ratio performed on the Face region extracted from Approach 2. Different image pre-processing techniques for Approach 3 were studied at the preliminary stages of this study. It was noticed that the red to green color ratio yielded better image results, on a random sample set from the database, than other techniques.

After performing pre-processing techniques, the data was trained using a modified version of the pre-trained "Alex Net" (Krizhevsky et al, 2017) architecture. Models were tested using data from the Oulu-CASIA database with a 70% training and 30% validation set split ratio.

After training, the Gradient-weighted Class Activation Mapping (Grad-CAM) visualization method was implemented to visualize the regions of impact on classification output determined by the model.

### *2.2 Region Detection and Enhancement*

To eliminate background noise and highlight the face of the subject, the cascade detection algorithm of Viola-Jones (Viola and Jones, 2001) was implemented in Approach 2 and Approach 3. For the Face region the "FrontalFaceLBP" (Mathworks, 2021) model, which detects upright and forward facing faces using Local Binary Patterns to encode facial features, with Merge Threshold of 5 and no Region of Interest (ROI) were set. This function is located in the vision Cascade Object Detector toolbox of MATLAB (Mathworks, 2021). The image was then resized to 227x227 pixels and referred to as Approach 2 hereafter.

When utilizing detection algorithms some background noise still remains which may affect the decision making process. Therefore, to ensure a more robust model an enhancement following the Face detection output was implemented. This enhancement is part of Approach 3, where the ratio of Red to Green from the RGB matrix was computed to highlight facial motion (Becouze et al, 2007). After analysing the outcomes from the different RGB ratio combinations, the Red to Green was chosen since it effectively highlighted the complexions of the individuals in the dataset. A limit of 1.0 was selected, so any ratio below this limit is set to 0 converting the original input image into a binary image.


Fig. 1. Sample of image inputs (left to right) for: a) Approach 1, raw image as given, b) Approach 2; c) Approach 3. The image is of a subject taken from the Oulu-CASIA database.

The output then passes through the Open morphological operation to eliminate some outlier points and then the holes filled via flood-fill. The binary data is then converted to unit 8 data type and multiplied with the face image from Approach 2. It is further cropped according to the limits of the facial component remaining. Figure 1 shows the three different approaches (Approach 1, Approach 2, Approach 3) for image input.

### *2.3 Model Parameters*

A Convolutional Neural Network architecture was developed through transfer learning by utilizing the "Alex Net" (Krizhevsky et al, 2017) architecture with its initial pre-trained parameters and modified for FER classification. The network is composed of 25 layers consisting of 3 convolutional layers and 3 fully connected layers. The bottom 3 layers of the network were altered to fit the number of emotion classes. The fully connected layer "fc8" was reduced from 1000 classes to 7 and its Bias learning rate and Weight learning rate factor were set to 20. A SoftMax activation and a classification layer completed the architecture.

### *2.4 Performance Criteria*

The network was trained using the training options described in Table 1. A random selection of images from the dataset was chosen to be distributed into the training and validation sets according to the 70% training and 30% validation split. The average true positive accuracies predicted from the validation set were used as the performance criteria of the model and study approach.

**Table 1. Network Training Options**

| Parameter | Method / Value Selected |
|---|---|
| Solver | Stochastic Gradient Descent with Momentum (SGDM) |
| Mini Batch Size | 150 |
| # Epochs | 50 |
| Shuffle | Every Epoch |
| Initial Learning Rate | 0.0001 |
| Momentum | 0.9 |

### *2.5 Grad-CAM Visualization*

The Grad-CAM visualization technique is a generalization of the class activation mapping (Selvaraju et al, 2017). It uses the gradients of the classification scores with respect to a convolution features map to determine which part of an image has a strong impact on classification (Mathworks, 2021). The layers chosen for this study were the SoftMax activation and the Rectified Linear Unit (ReLU) activation "relu5" in order to calculate and generate the maps for each image in the validation dataset.

Each image produces $n$ number of maps, where $n$ is the number of classes. The mean of the map data from each class of each image in the validation set was calculated and then analysed. A final independent validation was also performed

using images form the Japanese Female Facial Expression (JAFFE) database (Lyons et al, 2020).

*2.6 Database Description*

The Oulu-CASIA database consists of videos or image sequences from 80 different subjects expressing 6 different emotions of Anger (Ang), Disgust (Disg), Fear, Happiness (Happ), Sadness (Sad) and Surprise (Surp). Each of the image sequences starts with a neutral expression and ends with a strong expression of the particular emotion (Zhao et al, 2011). The image sequences of the original RGB, of visible light with strong illumination lighting were selected for this study. The dataset selected consists of 10,379 images in total, each representing facial portraits, as seen in Fig.1.

## 3. EXPERIMENTAL RESULTS & DISCUSSIONS

*3.1 Data Selection and Distribution*

Implementing the face detection algorithm of Approach 2, on the Oulu-CASIA database proved effective. In 2.97 % of the images in the chosen database, detection of the face failed. To assure equal opportunity for all approaches those images were removed from the dataset and further analysis. The new dataset was composed of 10,074 images distributed near equally between the emotion classes in Table 2. This even distribution provides an equal opportunity among the classes without giving a bias to a particular emotion class and helped in the splitting of the data into training and validation sets.

**Table 2. Class Distribution of Dataset**

| Emotion Class | Database | Dataset | Training Set | Validation Set | % Class Distrib. |
|---|---|---|---|---|---|
| Anger | 1,790 | 1,728 | 1,210 | 518 | 17.15 |
| Disgust | 1,633 | 1,571 | 1,100 | 471 | 15.59 |
| Fear | 1,796 | 1,770 | 1,239 | 531 | 17.57 |
| Happiness | 1,791 | 1,770 | 1,239 | 531 | 17.57 |
| Sadness | 1,668 | 1,617 | 1,132 | 485 | 16.06 |
| Surprise | 1,701 | 1,618 | 1,133 | 485 | 16.06 |
| Total | 10,379 | 10,074 | 7,053 | 3,021 | 100.00 |

*3.2 Performance Analysis*

Table 3 summarizes prediction testing results from the validation and training sets with the three different approaches (Approach 1, Approach 2, Approach 3), where Approach 1 is the raw image alone or a baseline comparator. As observed, there are slight differences in accuracy and loss values, which suggest any approach is suitable for FER. Approach 1 showed the best performance with Approach 3 the lowest. Looking at the difference between the training and the validation accuracies, they are in a range of less than 1.5% difference which suggest that the model is not being over-fitted.

**Table 3. Statistical Data from Modelling**

| | Approach 1 | Approach 2 | Approach 3 |
|---|---|---|---|
| Valid. Accuracy. | 99.90 % | 98.78 % | 98.58 % |
| Valid. Loss | 0.0066 | 0.0471 | 0.0517 |
| Train. Accuracy | 99.33 % | 100.00 % | 98.67 % |
| Train. Loss | 0.0253 | 0.0342 | 0.0453 |

Tables 4 to 6 represent the confusion matrix charts for the different models. The x-axis represents the predicted class while the y-axis represents the true class.

**Table 4. Confusion Matrix of Approach 1**

| | Ang | Disg | Fear | Happ | Sad | Surp |
|---|---|---|---|---|---|---|
| Ang | 100.0 | 0 | 0 | 0 | 0 | 0 |
| Disg | 0 | 100.0 | 0 | 0 | 0 | 0 |
| Fear | 0 | 0 | 100.0 | 0 | 0 | 0 |
| Happ | 0 | 0 | 0 | 100.0 | 0 | 0 |
| Sad | 0.21 | 0 | 0 | 0 | 99.58 | 0.21 |
| Surp | 0 | 0 | 0 | 0 | 0.21 | 99.79 |

Looking at the confusion matrices, it is clear Approach 1 showed an over fitting potential of the data reaching 100% accuracy on the validation set. The misclassifications in Approaches 2 and 3 were distributed evenly across the classes. The classes of Sad and Happiness showed the best performance while the class of Surprise and Fear the least for Approaches 2 and 3 respectively.

The reason Sad class showed the least performance in Approach 1 while best and second best for Approaches 2 and 3 respectively was clarified in Figs. 2-4. The Sad class was less over-fit because of the region focused by the model, where other classes of Approach 1 focused on the background more, the Sad class had a higher focus region around the mouth and chin region.

**Table 5. Confusion Matrix of Approach 2**

| % | Ang | Disg | Fear | Happ | Sad | Surp |
|---|---|---|---|---|---|---|
| Ang | 98.84 | 0.58 | 0 | 0.19 | 0.39 | 0 |
| Disg | 0.21 | 99.36 | 0 | 0 | 0.43 | 0 |
| Fear | 0.18 | 1.32 | 98.12 | 0 | 0.19 | 0.19 |
| Happ | 0.56 | 0 | 0 | 99.25 | 0 | 0.19 |
| Sad | 0.41 | 0 | 0 | 0 | 99.38 | 0.21 |
| Surp | 0 | 0.21 | 0.21 | 0.41 | 1.44 | 97.73 |

**Table 6. Confusion Matrix of Approach 3**

| % | Ang | Disg | Fear | Happ | Sad | Surp |
|---|---|---|---|---|---|---|
| Ang | 98.65 | 0.19 | 0 | 0 | 0.77 | 0.39 |
| Disg | 1.06 | 97.24 | 1.06 | 0.21 | 0.43 | 0 |
| Fear | 0.19 | 1.51 | 97.93 | 0 | 0 | 0.37 |
| Happ | 0.19 | 0 | 0 | 99.44 | 0 | 0.37 |
| Sad | 0.41 | 0 | 0 | 0 | 99.38 | 0.21 |
| Surp | 0 | 0.41 | 0 | 0.41 | 0.41 | 98.77 |

*3.3 Prediction Visualization Analysis*

Figures 2 to 4 represent the mean of the Grad-Cam map intensities for each class on the validation set images. The

difference can easily be noticed between the different approaches. In Fig. 2 in particular, the focus region varied between classes and areas posing no significance to emotion recognition, such as the strong intensities, in yellow, at the white board above the hair and at the side of the table. Fig. 2 strengthens the theory derived from Table 1 that Approach 1 was over-fit. This was evident in the highlighted region in the figure which is part of the background, and the background is the same throughout the images of the database.

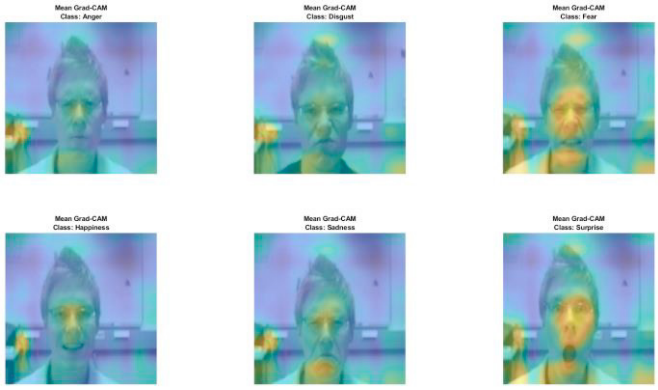


Fig. 2. Grad-CAM Prediction Visualizations of Approach 1 for each Emotion Class.

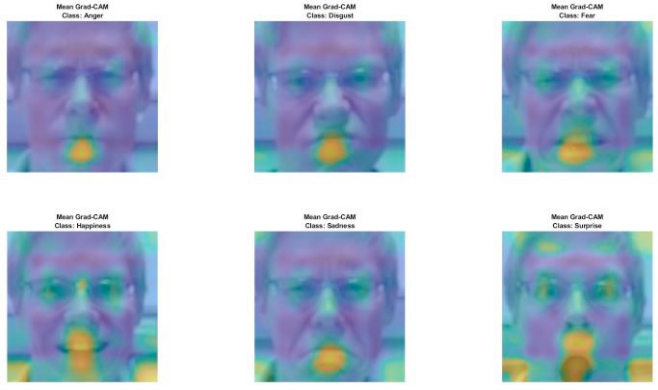


Fig. 3. Grad-CAM Prediction Visualizations of Approach 2 for each Emotion Class.

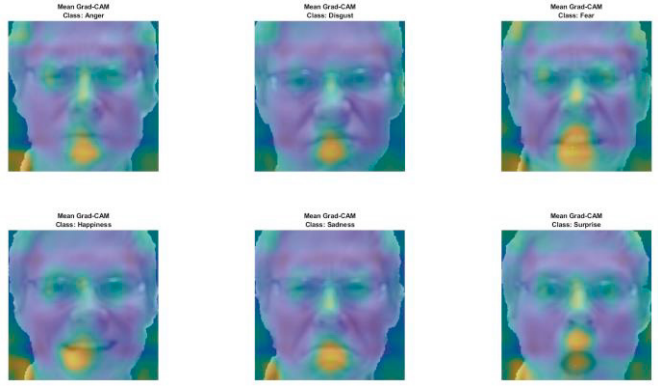


Fig. 4. Grad-CAM Prediction Visualizations of Approach 3 for each Emotion Class.

In Fig. 3 the region of impact is more consistent and highlighted at the areas of the mouth and eyes. However, there exist some areas, such as the corners of the images, which are outside the facial region area and thus do not provide any information for emotion classification. Fig. 4 representing Approach 3 performed better than Approaches 1 and 2, by keeping the focus concise to the mouth and eye regions and thus diminishing the large influence of areas outside the curve of the face, as segmented and shown in Fig. 1. Approach 3 showed promise, but still requires more enhancement to alleviate the influential impact of some areas, such as the left corner of the image and clothes from classification.

### *3.4 Cross Database Validation*

The models were tested against images from the JAFFE database (Lyons et al, 2020). The results showed accuracy levels of 27.87%, 31.69%, and 33.33% with the classes of Happiness and Surprise being robust with minimum 75.40% and 57.00% correct classifications, respectively. This independent validation result strengthens the conclusion that Approach 3, with more image pre-processing and facial focus, performed the best and showed better robustness to different images.

The accuracy results from the cross database validation showed that the model did not generalize well to images not seen in its database. This outcome suggests the fine tuning of model parameters and consideration of different network architectures to improve on generalization, which will be addressed in future work.

When testing the models on a real time app developed for FER it was noted the classes of Happiness and Surprise had better robustness in all approaches. Fig. 5 represents the real time application developed for FER testing in this work.

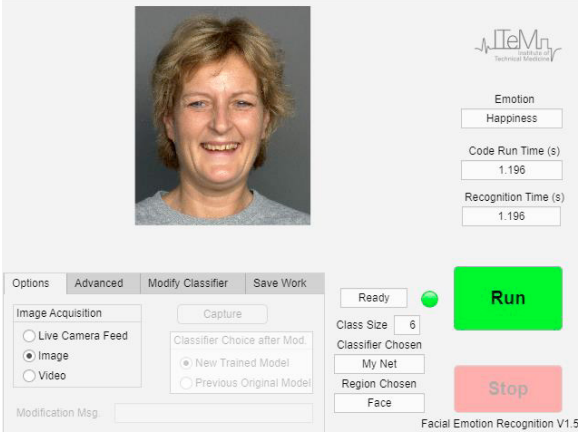


Fig. 5. Real-Time App developed for FER. Image of Subject from the FACES database (Ebener et al, 2010).

## 4. CONCLUSION

In this study the effects of image pre-processing were analysed for their significance in network decision making.

The results show, that the models were quite effective in achieving high accuracy levels if images were taken from the trainings database reaching values of up to 99.90% validation accuracy. However, a substantial difference in the areas of strong impact on classification of FER was found between the approaches. The visualizations helped to prove the hypothesis that proper image pre-processing is essential in assuring the selection of descriptive features related to FER in network modelling.

Further research is required to enhance the pre-processing technique to eliminate non facial components and fine tuning of the network parameters to better incorporate the problem of FER modelling. Different network architectures and attention modules are also being studied to find a model that is more robust and generalizes well.

## ACKNOWLEDGMENT

Partial support by a grant from the German Federal Ministry of Research and Education (BMBF) under project No. 13FH5I06IA – PersonaMed is gratefully acknowledged.

## AUTHOR'S STATEMENT

Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

## REFERENCES

(2001). American Academy of Pediatrics: The pediatrician's role in the diagnosis and management of autistic spectrum disorder in children. Pediatrics 107 (5), 1221–1226. https://doi.org/10.1542/peds.107.5.1221.

Arabian, H., Wagner-Hartl, V., Möller, K. (2021). Facial emotion recognition based on localized region segmentation. https://doi.org/10.5281/zenodo.4922791.

Becouze, P., Hann, C. E., Chase, J. G., and Shaw, G. M. (2007). Measuring facial grimacing for quantifying patient agitation in critical care. Computer methods and programs in biomedicine 87 (2), 138–147. https://doi.org/10.1016/j.cmpb.2007.05.005.

Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). FACES--a database of facial expressions in young, middle-aged, and older women and men: development and validation. Behavior research methods 42 (1), 351–362. https://doi.org/10.3758/BRM.42.1.351.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM 60 (6), 84–90. https://doi.org/10.1145/3065386.

Lugo-Marín, J., Gisbert-Gustemps, L., Setien-Ramos, I., Español-Martín, G., Ibañez-Jimenez, P., Forner-Puntonet, M., Arteaga-Henríquez, G., Soriano-Día, A., Duque-Yemail, J. D., and Ramos-Quiroga, J. A. (2021). COVID-19 pandemic effects in people with autism spectrum disorder and their caregivers: evaluation of social distancing and lockdown impact on mental health and general status. Research in autism spectrum disorders 83, 101757. https://doi.org/10.1016/j.rasd.2021.101757.

Lyons, M. J., Kamachi, M., and Gyoba, J. (2020). Coding Facial Expressions with Gabor Wavelets (IVC Special Issue). https://doi.org/10.5281/zenodo.4029679.

MathWorks Inc. (2021). *MATLAB* (9.10.0 (R2021a)). [Software].

Mehrabian, A. (2017). Communication without words. In C. David Mortensen (ed.), *Communication Theory*. Routledge, New York, NY.

Rylaarsdam, L., and Guemez-Gamboa, A., (2019). Genetic causes and modifiers of autism spectrum disorder. Frontiers in cellular neuroscience 13, 385. https://doi.org/10.3389/fncel.2019.00385.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017 - 2017). Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 10/22/2017 - 10/29/2017. IEEE, 618–626.

Viola, P., and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8-14 Dec. 2001. IEEE Comput. Soc, I-511-I-518.

Zhao, G., and Huang, X., Taini, M., Li, Stan Z., and Pietikäinen, M. (2011). Facial expression recognition from near-infrared videos. Image and Vision Computing 29 (9), 607–619. https://doi.org/10.1016/j.imavis.2011.07.002.