Facial expression recognition using deep learning

Cite as: AIP Conference Proceedings **2334**, 070003 (2021); https://doi.org/10.1063/5.0042221
Published Online: 02 March 2021

Harisu Abdullahi Shehu, Md. Haidar Sharif and Sahin Uyaver

Facial Expression Recognition Using Deep Learning

Harisu Abdullahi Shehu[1,a)], Md. Haidar Sharif[2,b)] and Sahin Uyaver[3,c)]

[1]*Victoria University of Wellington, New Zealand*
[2]*University of Hail, Kingdom of Saudi Arabia*
[3] *Turkish-German University, Faculty of Science, Department of Energy Science and Technologies, Istanbul, Turkey*

a)Corresponding author: harisu.abdullahi.shehu@ecs.vuw.ac.nz
b)md.sharif@uoh.edu.sa
c)uyaver@tau.edu.tr

Abstract. Facial expression recognition has become an increasingly important area of research in recent years. Neural network-based methods have made amazing progress in performing recognition-based tasks, winning competitions set up by various data science communities, and achieving high performance on many datasets. Miscellaneous regularization methods have been utilized by various researchers to help combat over-fitting, to reduce training time, and to generalize their models. In this paper, by applying the Haar Cascade classifier to crop faces and focus on the region of interest, we hypothesize that we would attain a fast convergence without using the whole image to analyze facial expressions. We also apply label smoothing and analyze its effect on the databases of CK+, KDEF, and RAF. The ResNet model has been employed as an example of a neural network model. Label smoothing has demonstrated an improvement of the recognition accuracy up to 0.5% considering CK+ and the KDEF databases. While the application of Haar Cascade has shown to decrease the achieved accuracy on KDEF and RAF databases with a small margin, fast convergence of the model has been observed.

Keywords: Deep learning, Emotion, Facial expression, Haar cascade, Label smoothing, Recognition.

INTRODUCTION

Facial expression recognition is a key area of research that has gained attention in the past few decades due to the increasing number of surveillance cameras. It helps us to communicate our own internal emotional states as well as identify human emotions and behaviors [1]. The increasing demand for robots to assist humans in a shared workspace has motivated researchers to start thinking as to whether to introduce robots in different workspaces. But these robots need to understand human emotions for interacting in an intuitive way.

Deep Learning (DL) models are very powerful and have performed really well in by achieving state-of-the-art performance on various databases (e.g., ImageNet [2]). Not long after Rumelhart et al. [3] derived back-propagation for the quadratic loss function, many efforts have been made by researchers to explore additional methods to achieve a better classification accuracy and attain fast convergence since the DL models require a longer time to be trained. Until recently, Szegedy et al. [4] introduced label smoothing which helps to improve accuracy by computing cross-entropy with a soft target. Since then, label smoothing has been used for improving the accuracy of deep learning models in various classification, object, speech, and image recognition tasks. Nevertheless, label smoothing has not always been effective on all kinds of datasets [5].

In this paper, we aim to address the issues of fast convergence on the facial expression-based datasets by applying the Haar Cascade (HC) classifier to crop faces. We focus on the region of interest (ROI) rather than analyzing the whole image. We also apply label smoothing and analyze its resulting effect on the databases of CK+ [8], Karolinska Directed Emotional Faces (KDEF) [9], and Real-world Affective Faces (RAF) [10].

DATASET

Fig. 2 (a) demonstrates samples of expressions from RAF, CK+, and KDEF databases. From left to right of Fig. **??**(a) shows anger, disgust, fear, happy, neutral, sad, and surprise expressions.

Fourth International Conference of Mathematical Sciences (ICMS 2020)
AIP Conf. Proc. 2334, 070003-1–070003-5; https://doi.org/10.1063/5.0042221
Published by AIP Publishing. 978-0-7354-4078-4/$30.00

The CK+ [7] is a database of mainly posed facial expressions. The database has seven peak expressions; six basic (anger, disgust, fear, happy, sad, surprise) expression defined by Ekman [8] and the contempt expressions. The CK+ database has a varying sequence starting from 10 to 60 frames. In this paper, a total of 1236 images consisting of the last three frames of each sequence of the six basic expressions are used as the peak expressions and the first frame of each sequence is used as a neutral expression.
The KDEF [9] uses a set of 4900 images of human facial expression. The database consists of seven different facial expressions which include the six basic + the neutral expression. All the 4900 images in the KDEF dataset are used in this research.
The RAF [10] uses images consisting of the six basic expression plus neutral expression downloaded from the internet. In this paper, we have used the aligned images of the RAF dataset, which holds separate training and test images. The test set consists of 3068 images, whereas the training set comprises up to 12271 images.

METHOD

Fig. 1(a) represents the facial expression recognition building block. Expressions are analyzed in two different ways; one way is analyzing the expression by directly feeding raw images to the ResNet model after pre-processing operation. Another way is by applying the Haar Cascade (HC) classifier to detect faces. Detected faces are cropped, pre-processed, and then fed to the model to perform the analysis. In situations, whereby the HC classifier fails to detect a face, the full image is fed to the model.

Pre-processing

Images have been resized to 128×128 pixels and converted to an array. Labels are one-hot encoded. Normalization has been applied on the pixels of raw images to increase the computation speed.

Label Smoothing (LS)

Sometimes, the LS [4] helps to increase the accuracy of a model by changing hard, binary label assignments to soft label assignments. We eliminate binary assignment and allow the actual class of each image to have 0.9 probability, whereas the remaining 10% is distributed among all other classes just so as to avoid over-fitting our model and to reduce the confidence level of the model. Fig. 1(b) shows how label smoothing affects the label of our data after it has been applied.

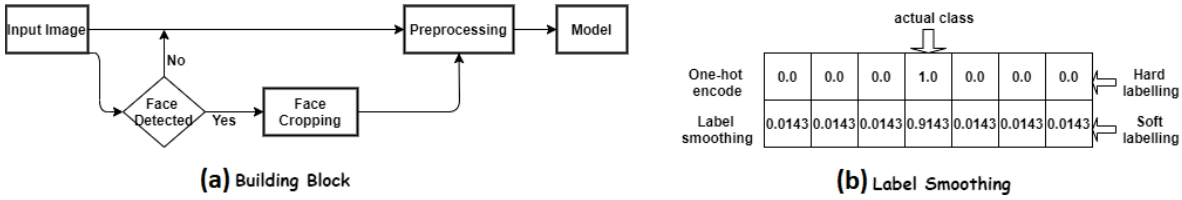

FIGURE 1. (a) shows flow diagram of emotion recognition. (b) hints the technique to use label smoothing to seven emotion labels.

Haar Cascade (HC)

The HC [6] has been used to detect faces on all images from the CK+, RAF, and KDEF databases. Detected faces are cropped and resized to 128×128 pixels to ensure a focus on ROI. Fig. 2 (b) depicts an example of a cropped emotional face from the CK+ database. Cropped faces are then fed to the model for training and emotion analysis.

ResNet Model

The ResNet50 is a type of neural network model with 50 layers. The same architecture of the model has been used as introduced in [11]. The model is trained from scratch over 200 epochs with a scheduled reduction in the learning rate after 80, 120, 160, and 180 epochs. Data augmentation technique has also been applied to increase the diversity of the data applied for training and prevent the model from over-fitting.

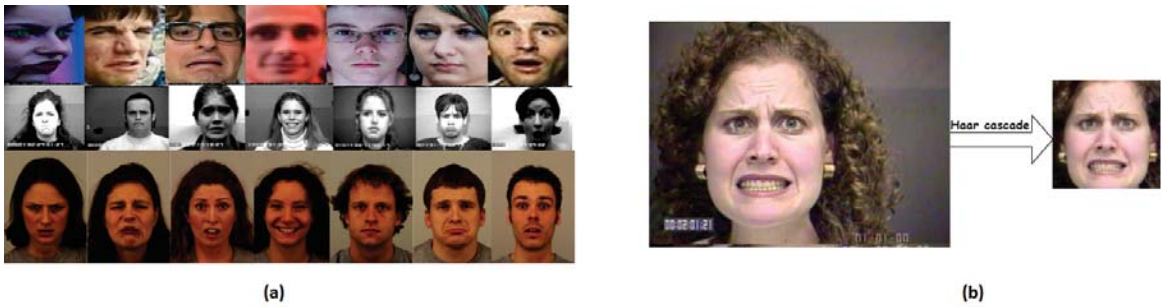

FIGURE 2. (a) shows samples of various expressions from RAF, CK+, and KDEF databases. Images in the first-row, mid-row, and last row represent expressions from RAF, CK+, and KDEF, respectively. (b) points to a sample cropped image using the HC.

RESULTS AND FINDINGS

Fig. 3(a) belongs to a bar chart showing the accuracy [12] obtained by the CK+, KDEF, and RAF databases on raw images and with the application of LS and HC. The LS increases the achieved accuracy by the ResNet model on raw data of the CK+ (from 98.17% to 98.78%) and the KDEF (from 96.48% to 96.67%) databases.
However, the achieved accuracy on the RAF database has been decreased from 81.68 to 80.89%. Although the LS does not seem to have any effect on the RAF database in terms of accuracy, applying K-means clustering on the extracted features of an arbitrarily selected two expressions (anger and disgust) from the test data, we found that features extracted by a model trained with label smoothing tend to have a better cluster with high intra-class similarity (see Fig. 3(b)) than features extracted by a model trained with no the LS (see Fig. 3(c)) on the same set of images. In addition, the presence of outliers among features of the same cluster can be observed in the cluster generated by a model trained without label smoothing; whereas the model trained with label smoothing tend to have better normality. Applying HC to focus on the ROI has increased the performance on the CK+ from 98.17% to 98.25%. While the application of HC has not shown an improvement in accuracy on the KDEF and RAF databases, with the method achieving 92.44%, and 80.57% on the KDEF and RAF databases. Fig. 4 has shown that fast convergence is achieved with the application of the HC method.

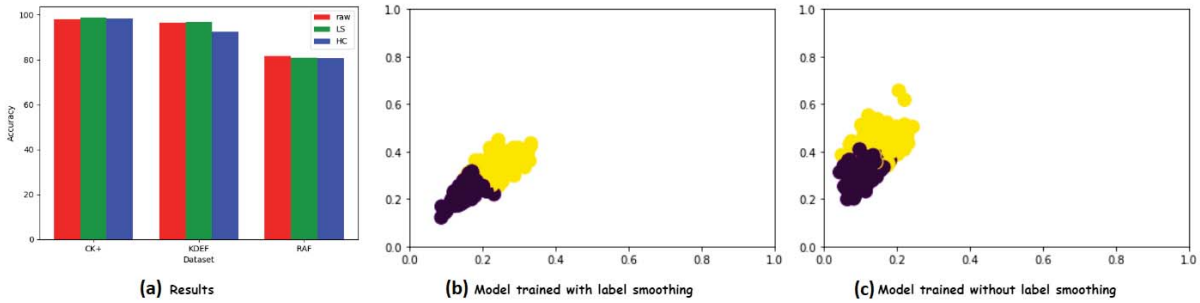

FIGURE 3. (a) indicates accuracy considering CK+, KDEF, and RAF. (b) hints feature cluster obtained by a model trained with LS. (c) demonstrates feature cluster obtained by a model trained without LS.

Fig. 4 depicts the training and validation accuracy achieved by the CK+, KDEF, and RAF databases. It is noticeable that both CK+ and KDEF trained with no the HC classifier (as shown in Fig. 4(a) and (b)) take longer time to converge as compared to the trained with HC (as shown in Fig. 4(d) and (e)). Conversely, almost no difference can be observed when the same technique has been applied to the RAF database (see Fig. 4(c) and (f)). This is due to the fact that the aligned images of the RAF database consists of already cropped face images (see Fig. **??**(a)). As such, the application of HC does not make any significant changes to the images.

CONCLUSION

We used the ResNet50 architecture to detect facial expression on images from the databases of CK+, KDEF, and RAF. Images from the KDEF were taken in five different formats: full left profile, half left profile, straight, half right

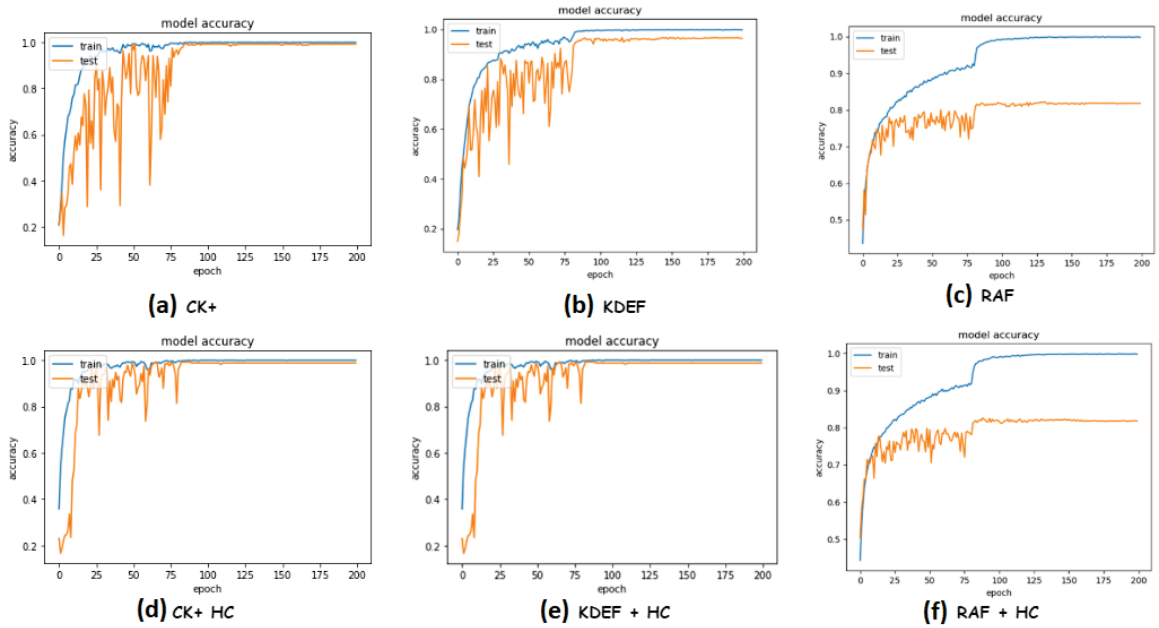

FIGURE 4. Above graphs demonstrate the training and validation accuracy obtained by the ResNet model on (a) CK+, (b) KDEF, and (c) RAF without using HC as well as (d) CK+, (e) KDEF, and (f) RAF with using HC to the images.

profile, and full right profile. The fast convergence on the CK+ and KDEF databases was obtained by the HC. Label smoothing did not increase the accuracy of certain databases, but it can achieve a better clustering result. Non-frontal face images from the KDEF were very difficult to detect, thus it was not possible to crop them properly. The proposed expression-recognition model showed an accuracy of up to 98.78%, 96.67%, and 81.68% on the CK+, KDEF, and RAF, respectively. Future work would overcome the existing problem by applying a more robust method.

REFERENCES

[1] H. Feng and J. Shao, "Facial Expression Recognition Based on Local Features of Transfer Learning," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2020, pp. 71-76, doi: 10.1109/ITNEC48623.2020.9084794.

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[3] Rumelhart, D. E., Hinton, G. E., Ronald J Williams R., J. (1986). Learning representations by back-propagating errors. Nature, 323:19.

[4] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[5] Müller, R., Kornblith, S., & Hinton, G. (2019). When Does Label Smoothing Help? (NeurIPS). Retrieved from http://arxiv.org/abs/1906.02629

[6] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1. https://doi.org/10.1109/cvpr.2001.990517

[7] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010, (May 2014), 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

[8] P. Ekman and W. V. Friesen, “Constants across cultures in the face andemotion.” Journal of personality and social psychology, vol. 17, no. 2, p. 124, 1971.

[9] Lundqvist, D., Flykt, A., & Ohman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

[10] Li, S., Deng, W., & Du, J. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing, 28(1), 356-370. https://doi.org/10.1109/TIP.2018.2868382

[11] Tran, E., Mayhew, M. B., Kim, H., Karande, P. & Kaplan A. D., Facial Expression Recognition Using a Large Out-of-Context Dataset. 2018 IEEE Winter Applications of Computer Vision Workshops (WACVW), Lake Tahoe, NV, 2018, pp. 52-59. http://doi: 10.1109/WACVW.2018.00012

[12] Sharif, M.H., An eigenvalue approach to detect flows and events in crowd videos. Journal of Circuits, Systems and Computers, vol. 26, no. 7, p. 1750110, 2017.