Proceedings on Automation in Medical Engineering
Proc AUTOMED, Vol 2, No 1 (2023): Article Ref 751
www.journals.infinite-science.de/automed/article/view/751

Infinite Science Publishing

Dynamic emotion recognition using histogram of oriented gradients

H. Arabian[1]*, and K. Möller[1]

[1] Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany
[*] Corresponding author, email: Herag.Arabian@hs-furtwangen.de

Abstract: A dynamic emotion recognition model is being developed for the use in a closed loop feedback system to support the emotional development of children with autism spectrum disorder. A neural network model is designed to learn the patterns in the acquired data. The model takes in input the sequence of features extracted from the traditional feature extraction technique of histogram of oriented gradients. The model performance is assessed by the classification accuracies of the OULU-CASIA database validation set. Accuracies of up to 43.64% were achieved. This preliminary model marks the first step in developing a robust dynamic emotion recognition platform.

I. Introduction

With advancements in technology and peoples' reliance on them, the domain of health apps combined with artificially intelligent systems has become a new topic of interest with a wide range of possibilities [1]. One application in the field of medicine is the use of emotion recognition models to help assess and monitor the patient more efficiently. A novel application of emotion recognition is planned, for the use in a closed loop feedback system, to help support the emotional development of children with autism spectrum disorder (ASD) [2]. ASD is a neuro-developmental disorder that impairs social interaction, communication, behaviors and interests of individuals [2], [3]. It is estimated that nearly 1 out of every 59 individuals is affected by ASD [4].

A dynamic emotion recognition model is considered for a more robust interpretation of facial expressions. It is observed that 55% of emotional understanding is portrayed by the facial expressions [5]. The facial expression image sequences captured through a camera are then passed through an image pre-processing algorithm to focus on the facial component and reduce background noise. The feature extraction method of histogram of oriented gradients (HOG) [6] is then applied to capture the relevant relations between the pixels, i.e. shape and texture.

A temporal neural network model is used to classify the emotions from the extracted features. The OULU-CASIA [7] facial emotions database is selected for the model analysis. The performance of the model is based on the accuracy, confusion matrix metrics, and receiver operating characteristics (ROC) of the validation set.

The aim of this study is to show the feasibility of using HOG features for dynamic emotion recognition and its suitable implementation in a closed loop real time system.

II. Material and methods

To improve the emotion recognition task, image pre-processing was performed to reduce background noise and focus on the face of the individual. The cascade object detection algorithm of Viola-Jones [8] was implemented. An image fusion technique was also adopted to highlight changes in the static image in order to improve the representation and transition tracking from one image frame to the other in the sequence.

After the image pre-processing the feature extraction method of HOG was implemented. The parameters were set as a cell size of 8x8, a block size of 2x2, 9 pixel neighbors, and 9 histogram bins and no signed orientation. The generated features were a vector of size 26244x1.

The HOG features were then set as input to the temporal neural network model. The model is 5 layers deep, with a 1-D convolution, 3 fully connected (FC), and 1 long short-term memory (LSTM) layer. Fig. 1 shows the design of the network model. The set parameters are a filter size of 9 with a stride of for the 1-D convolution layer, a filter size of 9 and stride of 9 for the average pooling operation. The first 2 FC layers and the LSTM layer have 128 hidden layers. The dropout layer has a drop probability of 0.6. The input is a sequence of feature vectors and the classification is performed according to a sequence to sequence classification using the cross-entropy loss function.

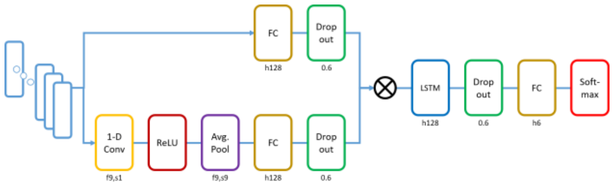

Figure 1: Temporal neural network model architecture.

The dataset was split into 2 sets, a training and a validation set with an 80/20 split. The model training was run 4 times, with each run selecting a random sample set for training and validation. The performance was based on true positive (TP) predictions of the validation set. The confusion matrix metrics were computed as well as the ROC analyzed.

The model was trained using the stochastic gradient descent with momentum (SGDM) optimization on 150 epochs, with an initial learning rate of 0.001 with a learn drop rate of 0.9 every 15 epochs. The platform used for training was MATLAB 2022a run on a Windows 10, Ryzen 7, 8 Core processor with 64G RAM.

II.I. Data Selection

The emotions database of OULU-CASIA was selected for the model training and validation. The database is composed of image sequences with a neutral expression at start and a strong expression of the particular emotion at the end. The 80 subjects in the database expressed the 6 basic emotions of Anger, Disgust, Fear, Happiness, Sadness and Surprise. The data was collected under 3 lighting conditions of dark, strong and weak, and 2 image acquisition methods of near infra-red (NIR) and visible light (VL) [7]. For this study the dataset of VL with strong illumination was chosen, with a total of 10,379 static images or 480 image sequences. The first 4 frames of each image sequence were relabeled as a neutral expression.

The implemented image pre-processing phase proved to be effective with the loss of just 10.41% of the static images in the chosen OULU-CASIA dataset. This failure rate was due to missed detection of some regions of interest during processing. Those images were then excluded from further analysis. The final dataset for model building was composed of 9,299 images distributed near equally between the emotion classes, this was equivalent to 478 image sequences.

III. Results and discussion

Table 1 represents the mean of the 4 different model prediction results on the validation set. The statistical data shows that the model did not perform well with accuracies reaching 43.64% with a standard deviation of 2.21.

Table 1: Model Performance.

Accuracy	43.64 ± 2.21
ROC	54.59 ± 3.55
F1-Score	39.43 ± 3.03
Recognition Rate	0.162s ± 0.06
Training Duration	~ 9 hrs.

This low performance is due to the limited availability of data as well as a short training duration, i.e. number of epochs. This is complemented by the weak performance shown by the area under the ROC curve with 54.59%. This weak outcome indicates that the classifier is not robust in distinguishing between the different classes. When comparing the performance of the different emotion classes, it was noticed that the 2 classes of happiness and neutral performed the best. Class accuracies of 72.7% ± 11.22 and 81.70% ± 9.3 were achieved for both happiness and neutral respectively. This achievement is related to the facial expression of happiness, as it has stronger differences in the expressiveness that can be identified more effectively through the feature extractor. The high performance outcome of the neutral class is attributed to the frame location within the image sequence. In contrast, the worse performing classes were that of anger and sadness as the distinguishing parameter for these emotion classes is prone to misclassifications as they require higher attention to detail.

Some limitations were considered for this study. The use of constant parameter for the feature extraction of HOG without fine tuning. The removal of some image samples from the dataset due to inability to detect regions of interest.

IV. Conclusions

In this study the use of the traditional feature extraction method of histogram of oriented gradients was evaluated for dynamic emotion recognition. The approach achieved accuracies of 43.63% for a small database of image sequences. The model was able to recognize the key features of importance for emotion classification and track their changes across the temporal domain. The dynamic modelling on a sequence to sequence strategy also helped address a key issue with real-time recognition, as this method can classify emotions of each frame and would not require the collection and storage of prior and post temporal information thereby achieving a detection rate of 6 Hz.

ACKNOWLEDGMENTS

Partial support by grants from the project KOMPASS funded by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) of Baden-Wuerttemberg and the German Federal Ministry of Research and Education (BMBF) under project No. 13FH5I06IA – PersonaMed are gratefully acknowledged.

AUTHOR'S STATEMENT

Conflict of interest: Authors state no conflict of interest.

REFERENCES

[1] H. Arabian, T. Abdulbaki Alshirbaji, N. A. Jalal, N. Ding, B. Laufer, and K. Moeller, 'Identifying User Adherence to Digital Health Apps', presented at the IUPESM World Congress on Medical Physics and Biomedical Engineering (IUPESM WC2022), Singapore, 2022, vol. in press.

[2] H. Arabian, V. Wagner-Hartl, and K. Möller, 'Facial emotion recognition based on localized region segmentation', Jun. 17, 2021. doi: 10.5281/zenodo.4922791.

[3] L. Tebartz van Elst et al., 'FASTER and SCOTT&EVA trainings for adults with high-functioning autism spectrum disorder (ASD): study protocol for a randomized controlled trial', *Trials*, vol. 22, no. 1, p. 261, Apr. 2021, doi: 10.1186/s13063-021-05205-9.

[4] L. Rylaarsdam and A. Guemez-Gamboa, 'Genetic Causes and Modifiers of Autism Spectrum Disorder', *Front. Cell. Neurosci.*, vol. 13, p. 385, 2019, doi: 10.3389/fncel.2019.00385.

[5] A. Mehrabian, 'Communication without words', in *Communication Theory*, C. D. Mortensen, Ed. Routledge, 2017.

[6] N. Dalal and B. Triggs, 'Histograms of oriented gradients for human detection', in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Jun. 2005, vol. 1, pp. 886–893 vol. 1. doi: 10.1109/CVPR.2005.177.

[7] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, 'Facial expression recognition from near-infrared videos', *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011, doi: 10.1016/j.imavis.2011.07.002.

[8] P. Viola and M. Jones, '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*, Dec. 2001, vol. 1, p. I–I. doi: 10.1109/CVPR.2001.990517.