2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Facial Emotion Recognition Focused on Descriptive Region Segmentation*

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

Abstract— **Facial emotion recognition (FER) is useful in many different applications and could offer significant benefit as part of feedback systems to train children with Autism Spectrum Disorder (ASD) who struggle to recognize facial expressions and emotions. This project explores the potential of real time FER based on the use of local regions of interest combined with a machine learning approach. Histogram of Oriented Gradients (HOG) was implemented for feature extraction, along with 3 different classifiers, 2 based on k-Nearest Neighbor and 1 using Support Vector Machine (SVM) classification. Model performance was compared using accuracy of randomly selected validation sets after training on random training sets of the Oulu-CASIA database. Image classes were distributed evenly, and accuracies of up to 98.44% were observed with small variation depending on data distributions. The region selection methodology provided a compromise between accuracy and number of extracted features, and validated the hypothesis a focus on smaller informative regions performs just as well as the entire image.**

Index Terms— **Autism Spectrum Disorder (ASD), Facial Emotion Recognition (FER), Feature Extraction (FE), Machine Learning, Oulu-CASIA**

I. 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, behaviors, and interests of individuals [1]. Estimates reveal 1 out of 59 individuals are affected by ASD [2]. Children suffering from ASD are accustomed to a certain routine and any deviation from the normalcy can cause psychological and emotional challenges to the child as well as an increased stress levels for the caregiver [3]. An individually adjusted virtual world combined with a reward system in the form of a gaming platform and technical affinity of most ASD children creates a suitable atmosphere for treatment.

When delivering a certain message, 55% of its efficacy is based on the facial component [4]. Machine Learning techniques are required to transform an image of a face into a corresponding emotion. Feature Extraction (FE) methods are needed to translate the pixel intensities to relevant relations, such as shapes or textures. The appearance based FE method of Histogram of Oriented Gradients (HOG) was used to make this transformation, since they are simple dense image descriptors that capture the orientation and gradient of a shape [5].

This study takes a different approach to FER by utilizing the object detection algorithm of Viola-Jones [6] to extract local regions of interest from images. The classification was performed using 2 classifier models for K-Nearest Neighbor (KNN) and 1 model for Support Vector Machines (SVM). The Oulu-CASIA [7] database is used to test the performance of the different models. The generated data is statistically analyzed to assess classification accuracy.

The aim of this study is to show localized region classification, with fewer features, performs just as well as classification based on the entire image.

II. System Description

A. Methodology

Fig. 1 illustrates the outline of the proposed model. The regions extracted were chosen as the face, mouth and left eyebrow. The face region was used as a basis for comparison of this study's approach. The left eyebrow was considered since it yielded better segmentation results, and tests performed on the eyebrow pair and right eyebrow showed similar classification results. After the correct segmentation of these regions the FE method of HOG was implemented, the different parameters were tuned, by running classifications using KNN model EUC1 on varying HOG Cell Sizes and Block Sizes, to get the optimal settings for FE. The HOG data of the mouth and eyebrow regions were combined to form one vector, while the face region was taken as generated.

After choosing the optimal FE parameters of cell size 10x10 pixels with a block size of 2x2 pixels, traditional machine learning algorithms were used because of their low computational costs and speed. Two different models of KNN, EUC1 and COS7, and one model of SVM, 1vs1, were used as classifiers. The second KNN model was based on the optimization results from fine tuning the classifier parameters using the "hyper-parameter-optimization" [8] setting in the MATLAB with a Bayesian optimization algorithm. Models were tested using data from the Oulu-CASIA database with a 70% training and 30% validation set split ratio.

B. Region Detection

To eliminate background noise and highlight the face of the subject, the cascade detection algorithm of Viola-Jones [6] was implemented. The process was combined with different functions for the different regions. For the Face

*Research supported by the Institute of Technical Medicine (ITeM).

H. Arabian, K. Möller are with the Institute of Technical Medicine (ITeM), 78054 VS-Schwenningen, Germany (phone: +49 (0)7720-307-4390; e-mail: H.Arabian@hs-furtwangen.de).

V. Wagner-Hartl is with the department of Industrial Technologies and Dean of Engineering Psychology at Hochschule Furtwangen University, 78532 Tuttlingen, Germany.

J. Geoffrey Chase, Center for Bioengineering, Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand.

region the "FrontalFaceLBP" [8] model with Merge Threshold of 5 and no Region of Interest (ROI) were set. For the Mouth region the "Mouth" [8] model with Merge Threshold 4, Min Size of 45x40 pixels and ROI of bottom half of the Face was set. For the Eyebrow region the "LEFTEyeCART" [8] model with Merge Threshold 4 and no ROI was used to find the left eye after which the origin of the detected rectangle, top left corner, was re-scaled by a factor of 0.9 in x and 0.95 in the y direction, along with extension of the width by 15 pixels and reduction of the height by a factor of 1.5, odd numbers were rounded down [9]. These functions are located in the vision Cascade Object Detector toolbox of MATLAB [8].

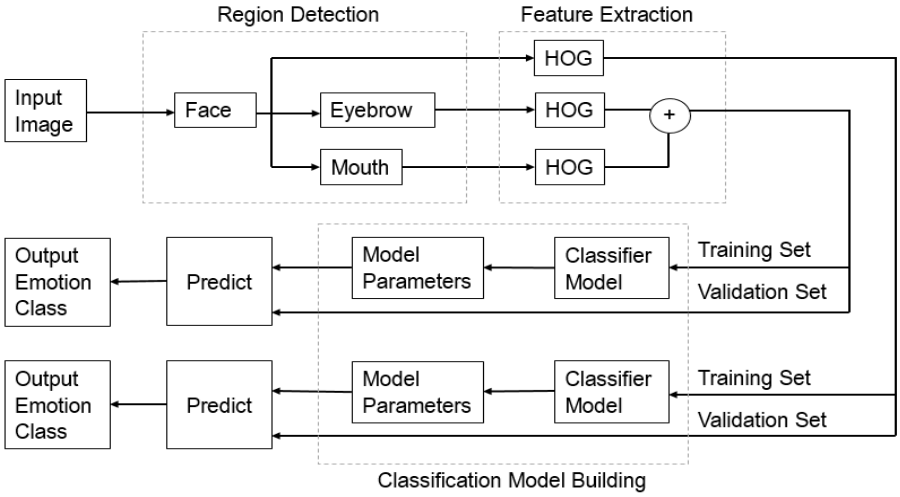

Figure 1. Outline of proposed system model

After the Face detection sequence, RGB images were transformed into grey scale then resized to pre-selected sizes of 227x227 pixels for the face and 50x80 pixels for both mouth and eyebrow regions. The hierarchy of first Face, then Eyebrow and then Mouth region was implemented. If the detection algorithms failed to detect any of the regions, the image was replaced by a zero matrix of same size and at the end of the detection loop, the AND operator was used with the Boolean operators of all 3 regions, such that any image of entire zeros for any of the regions was neglected and removed from the data set.

C. Model Parameters

The FE parameters of HOG was implemented with a cell size of 10x10 pixels, a block size of 2x2 pixels, with 9 number of pixel neighbors, and 9 histogram bins with no signed orientation. The HOG data of the mouth and eyebrow were combined to form 1 vector Mouth + Eyebrow, while the face region was taken as generated [9].

After generating the features, two different KNN models and 1 SVM model were used as classifiers. The first KNN model was set as EUC1 with K value of 1 nearest neighbor, Euclidean distance function, equal distance weights, exhaustive nearest neighbor search and no standardization of data. The second KNN model was chosen after performing a hyper parameter optimization function and selecting the highest K value for both regions. This model was COS7 with K=74, Cosine distance function, Squared-Inverse distance weights, exhaustive search and no standardization of data [8].

The SVM model was 1vs1 [8], which was based on the Error Correcting Output Code (ECOC) model of the 1vs1 approach. The solver of Sequential Minimal Optimization with $1e^{-3}$ as delta gradient tolerance, and iteration limit of $1e^{8}$ for global minima search limit was used [9].

D. Performance Criteria

Both regions of Face and Mouth + Eyebrow were tested, with 100 iterations for KNN models and 10 iterations for SVM. Each iteration chose a random selection of images from the dataset to be distributed into the training and validation sets according to the 70% training and 30% validation split. The reduced number of SVM model iterations was due to the fact of low time efficiency of each individual iteration and substantially heavy computational costs.

The average true positive accuracies predicted from the validation set were used as the performance criteria of the model and study approach. The accuracy data collected from all iterations of each model were tabulated and the statistical results of mean, median (Med), standard deviation (SD), coefficient of variance (CV), inter quartile range (IQR), Skewness, and 95% confidence intervals (CI) were analyzed. The data was also plotted as a Boxplot to showcase any outliers that are 1.5 times greater than that of the IQR. The confusion matrix chart was also analyzed to determine the misclassified classes.

D. Database Description

The Oulu-CASIA database consists of videos or image sequences from 80 different subjects expressing 6 different emotions of Anger, Disgust, Fear, Happiness, Sadness and Surprise. Each of the image sequences starts with a neutral expression and ends with a strong expression of the particular emotion [7]. The image sequences of the original RGB, of visible light with strong illumination lighting were selected for this study. The dataset selected consisted of 10,379 images in total, each representing facial portraits.

III. Experimental Results

A. Data Selection and Distribution

Table I summarizes the localized region detection sequence implemented. This method proved quite effective. In 10.36% of the images in the chosen Oulu-CASIA database due to missed detection of region of interest processing failed. Those images were excluded from the analysis. The new dataset for modelling was composed of 9,304 images distributed near equally between the emotion classes.

This equivalent distribution helped equalize opportunities for correct classification by not giving a bias towards a certain emotion class. This also helped in the even distribution of the images when splitting them into training and validation sets.

Fig. 2 represents the different regions detected within an image of a subject from the Oulu-CASIA database. The detection sequence successfully located the region of interest seen in the bounding box and highlighted the importance of region detection by removing background noises that affect emotion classification. The ROI detection sequences performed very well when dealing with the classes of Fear and Happiness with an average of 3.5% of the images being omitted. Therefore, the chosen parameters for the ROI detection algorithms worked effectively.

Table II summarizes the prediction results from testing the validation set on the corresponding trained models. The statistical data shows the EUC1 model performed the best with greater than 98% mean accuracy for both regions, it also showed the lowest variations in data distribution. The standard deviations were all below a 0.35% margin for both regions. The COS7 model also performed well, reaching accuracy levels greater than 90% mean accuracy and variations lower than 0.9% for both regions. The 1vs1 model had high accuracy, but varied in data distribution. Due to the small amount of generated data a conclusion was not drawn.

Since the EUC1 model showed the best performance, it was selected for use in a closed loop app developed for FER.

Fig. 3 shows that the Mouth + Eyebrow region had a difference range of less than 2 % between minimum and maximum accuracies, and only 1 outlier. The Face region did slightly better with a difference range of only 1% and no outlier presence. The Median of both the regions were situated in the middle of the IQR box. This result shows the implemented classifier model, performed well in dealing with small and large feature data. It also strengthened the hypothesis smaller regions perform just as well as larger ones.

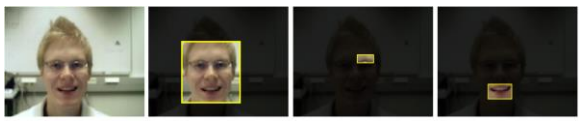

Figure 2. Regions detected for Face, Eyebrow and Mouth. Image of subject from Oulu-CASIA database [6]

Examining the Confusion Matrix charts of Table III and Table IV for the EUC1 model, the Face region performed slightly better than the Mouth + Eyebrow, with misclassification rates lower than a 0.55% margin compared to less than 1% for the Mouth + Eyebrow region. The best classification performance was for the Happiness class for both regions and the lowest was for Anger and Surprise classes for both Mouth + Eyebrow and Face regions respectively. Misclassification distribution was similar in both regions confirming small feature (Mouth + Eyebrow region) maintains the same accuracy as a larger feature data, such as the (Entire) Face region.

TABLE I. Distribution of Images of the Oulu-CASIA Database

Emotion Class	Database	Selected	% Neglected	Training	Validation	% Class Distribution
Anger (Ang)	1,790	1,541	13.91	1,079	462	16.56
Disgust (Dis)	1,633	1,432	12.31	1,002	430	15.39
Fear (Fea)	1,796	1,738	3.23	1,217	521	18.68
Happy (Hap)	1,791	1,725	3.69	1,208	517	18.54
Sadness (Sad)	1,668	1,437	13.85	1,006	431	15.45
Surprise (Sup)	1,701	1,431	15.87	1,002	429	15.38
Total	10,379	9,304	10.36	6,514	2,790	100.00

TABLE II. Statistical Results of Modelling

Accuracy %	Mouth + Eyebrow Region			Face Region		
	EUC1	*COS7*	*1vs1*	*EUC1*	*COS7*	*1vs1*
Mean ± SD	98.44 ± 0.26	92.14 ± 0.59	89.56 ± 0.54	98.94 ± 0.21	94.94 ± 0.48	98.84 ± 0.28
Lower 95% CI	98.39 ± 0.23	92.02 ± 0.52	89.17 ± 0.37	98.90 ± 0.19	94.85 ± 0.42	98.64 ± 0.19
Upper 95% CI	98.49 ± 0.30	92.26 ± 0.68	89.95 ± 0.99	98.99 ± 0.25	95.04 ± 0.56	99.04 ± 0.51
Median	98.42	92.15	89.66	98.96	94.95	98.76
IQR	0.32	0.79	0.93	0.29	0.54	0.47
Mean CV	0.27	0.64	0.61	0.21	0.51	0.28
Skewness	-0.20	-0.20	-0.41	-0.21	-0.40	0.38

TABLE III. Confusion Matrix Chart of EUC1 Model for Mouth + Eyebrow Region

Pre\ Act[a]	Ang	Dis	Fea	Hap	Sad	Sup
Ang	97.75%	0.80%	0.40%	0.13%	0.61%	0.32%
Dis	0.90%	97.94%	0.47%	0.16%	0.30%	0.23%
Fea	0.35%	0.11%	98.81%	0.20%	0.12%	0.42%
Hap	0.14%	0.07%	0.09%	99.23%	0.20%	0.27%
Sad	0.40%	0.30%	0.22%	0.23%	98.48%	0.37%
Sup	0.10%	0.17%	0.79%	0.53%	0.21%	98.21%

a. X-axis Predicted Classes, Y-axis True Classes

TABLE IV. Confusion Matrix Chart of EUC1 Model for Face Region

Pre\ Act[a]	Ang	Dis	Fea	Hap	Sad	Sup
Ang	98.58%	0.50%	0.53%	0.15%	0.23%	0.02%
Dis	0.39%	98.87%	0.38%	0.13%	0.18%	0.05%
Fea	0.27%	0.17%	99.35%	0.01%	0.02%	0.18%
Hap	0.03%	0.04%	0.02%	99.62%	0.10%	0.19%
Sad	0.12%	0.27%	0.30%	0.23%	98.61%	0.48%
Sup	0.21%	0.13%	0.51%	0.21%	0.51%	98.44%

a. X-axis Predicted Classes, Y-axis True Classes

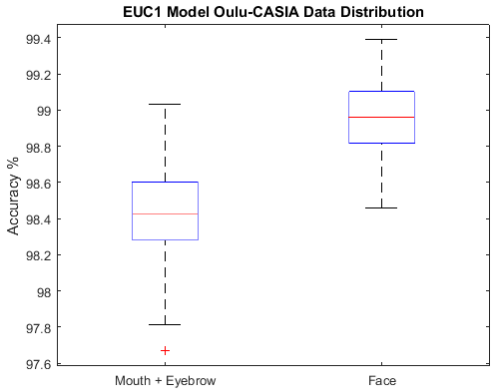

Figure 3. Boxplot representation of data generated from EUC1 model for both Mouth + Eyebrow and Face region

B. *Model Robustness Evaluation*

In order to evaluate the robustness of the trained EUC1 model, the Japanese Female Facial Expression (JAFFE) database was used for evaluation. The database consists of 213 images of Japanese female students in Grey scale color representing 7 emotion classes [10]. The Neutral class was excluded from validation because it is unavailable as a class in the Oulu-CASIA database.

The EUC1 model showed a mean accuracy of 38.92% ± 1.50 with a median of 38.89% and a range greater than 8%. Although the low accuracy means that the model is lacking robustness a closer look into the misclassifications was taken. The most misclassifications were of the negative emotional classes of Anger, Disgust, Fear and Sad, while that of the positive emotional classes of Happy and Surprise showed better robustness reaching values greater than 55% and 85%.

C. *State of the Art Comparison*

When comparing region segmentation against other state of the art results using the Oulu-CASIA database, this study recorded the highest accuracy with 98.44% when compared to the works of Zhang Huang et al. [11] with 86.25% and Jung et al. [12] with 81.46%.

When testing the models, in the closed loop app developed for FER, against images from subjects not found in the Oulu-CASIA or JAFFE database, the model showed better robustness at emotion classes of Happiness and Surprise. Hence, the overall approach shows significant promise, particularly for ASD applications.

IV. Conclusion

This study proposes a new method of localized region selection for FER modelling. The approach was able to achieve desirable accuracies reaching 98.44% mean accuracy on a big dataset of images. The statistical data helped verify that small regions of significant importance perform just as well as when dealing with the entire image as a whole. This approach also reduces the risk of focus of the classification algorithm on areas that are not relevant for classifying emotions. It is also a good compromise between number of generated features and maintaining accurate results.

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

[1] American Academy of Pediatrics, Committee on Children 2000-2001. "The Pediatrician's Role in the Diagnosis and Management of Autistic Spectrum Disorder in Children".

[2] Lauren Rylaarsdam and Alicia Guemez-Gamboa. 2019. "Genetic Causes and Modifiers of Autism Spectrum Disorder". Frontiers in cellular neuroscience 13, 385.

[3] Lugo-Marín J, Gisbert-Gustemps L, et al. "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". Res Autism Spectr Disord. 2021;83:101757.

[4] Albert Mehrabian. 2017. "Communication Without Words". In Communication Theory, C. David Mortensen, Ed. Routledge, New York NY.

[5] Dalal N., Triggs B. "Histograms of oriented gradients for human detection"; Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); San Diego, CA, USA. 20–25 June 2005; pp. 886–893.

[6] P. Viola and M. Jones. 2001. "Rapid object detection using a boosted cascade of simple features". CVPR 2001. IEEE Comput. Soc, I-511-I-518.

[7] Guoying Zhao, Xiaohua Huang,et al. (2011) Facial expression recognition from near-infrared videos. - Image and Vision Computing 29 (9), 607-619 DOI: https://doi.org/10.1016/j.imavis.2011.07.002

[8] MATLAB. (2020). version 9.9.0 (2020b). Natick, Massachusetts, USA: The MathWorks Inc.

[9] H. Arabian, V. Wagner-Hartl, et al. "Facial Emotion Recognition based on Localized Region Segmentation". 2021. DOI: https://doi.org/10.5281/zenodo.4922791.

[10] Michael J. Lyons, Miyuki Kamachi, et al. "Coding Facial Expressions with Gabor Wavelets (IVC Special Issue)". DOI: https://doi.org/10.5281/zenodo.4029679.

[11] Kaihao Zhang, Yongzhen Huang, et al. 2017. "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks". IEEE Signal Processing Society 26, 9, 4193–4203.

[12] Heechul Jung, Sihaeng Lee, et al. 2015 - 2015. "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition". In 2015 IEEE (ICCV). IEEE, 2983–299.