Comparison Study of Automated Facial Expression Recognition Models

Murilo de Souza Preto
Federal University of ABC
Santo André, SP, Brazil
Email: murilo.preto@aluno.ufabc.edu.br

Fernando Teubl Ferreira
Federal University of ABC
Santo André, SP, Brazil
Email: fernando.teubl@ufabc.edu.br

Celso Setsuo Kurashima
Federal University of ABC
Santo André, SP, Brazil
Email: celso.kurashima@ufabc.edu.br

***Abstract*—Facial expressions play a crucial role in human non-verbal communication, and in the psychology field there is a strong consensus on the existence of five key emotions: anger, fear, disgust, sadness, and happiness. This paper aims to evaluate multiple facial expression recognition detection models, assessing their performance across different machines and databases. By identifying the strengths and weaknesses of each option, the study seeks to comparatively determine the most suitable model for specific tasks or scenarios. For each computer, all databases were processed through the usage of the detection models, while measuring the required runtime for the facial expression detection. The detection models: Residual Masking Network and Deepface, were tested through the databases Extended Cohn-Kanade and AffectNet. The assessed data point towards an average higher accuracy for the model Residual Masking Network, but faster runtime for Deepface. Thereby, Deepface may be preferentially employed in scenarios where time constraints are a primary concern, there is limited processing capability available, or an emphasis on recognizing either happiness or neutral expressions, while Residual Masking Network might be favored in striving for a higher detection accuracy.**

I. Introduction

Facial expressions play a crucial role in human non-verbal communication, and while debates periodically happen in the psychology field regarding the validity of certain labeled emotions, such as "guilt", there is a strong consensus on the existence of five key emotions: anger, fear, disgust, sadness, and happiness [1].

Furthermore, facial expressions extend beyond the communication of emotions, and may also serve to convey the intention of an individual in social groups [2], to determine how people assess leadership capabilities [3], and even assist in detecting the drowsiness of a driver [4].

Given the rising prominence of computer vision, as well as machine learning, new detection methods for facial recognition, and facial expression recognition (FER), are constantly being developed and released, many of which as freeware, and often, as open-source projects.

With respect to this data, this research proposed to assess different methods of FER, mainly through the parameters of processing time and detection accuracy. Moreover, detection models were evaluated through different databases, including images captured in controlled and in-the-wild environments. Furthermore, tests were performed in different computers, including embedded systems, each with respective distinct specifications.

A. Objectives

In essence, this research aims to evaluate multiple FER detection models, assessing their performance across different machines and databases. By identifying the strengths and weaknesses of each option, the study seeks to comparatively determine the most suitable model for specific tasks or scenarios, mainly focusing on processing speed differences across different machines.

II. Related works

Beyond the practical applications of the researches listed in the Introduction, automated FER was used in a plethora of scenarios, such as for assessing the reaction of humans towards noise in urban centers [5], as well as predicting the self reported stress of a driver [6] (similarly to the previously cited study, which however assessed drowsiness). Moreover, some articles aim to aid in the implementation of these technologies. For instance, Li & Deng [7] offer a concise explanation of the main steps involved in deploying FER models, along with an extensive list of available technologies, in their article titled *Deep Facial Expression Recognition: A Survey*, which serves as a valuable reference for those seeking to initiate the implementation of basic FER models.

However, it is essential to highlight that the majority of research papers primarily concentrate on presenting the results obtained from the implementation of FER models, rather than the details of the implementation itself. Therefore, it is within this context that this research was motivated, and its primary objective is to contribute experimental data about the accuracy and, more importantly, the processing time of FER models when applied to diverse databases and machines. In doing so, this study aims to contribute to the understanding of practical application of FER models in various real-world scenarios.

III. Methodology

As for the Methodology, this section will be divided in four (4) subsections: Materials, containing information about the machines and databases required for this research; Image Pre-processing, regarding common image pre-processing methods; Facial Expression Recognition, in relation to the different

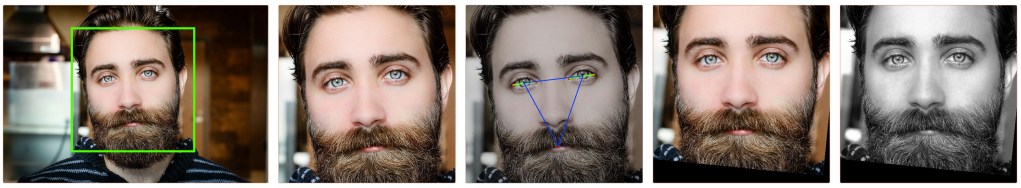

Fig. 1. Suggested pre-processing steps for Facial Expression Recognition. Original image available as open-source [8], further processing made by the author.

proposed models for FER; and Data evaluation, concerning which types of parameters would be assessed throughout this research.

A. Materials

The required materials for this research were: Machines, that is, computers with enough processing capabilities to handle image processing; as well as Databases, which allowed a predictable image input. Both subsets will be described in the subsequent particular sections.

1) Machines: Four distinct machines were utilized for processing images of the databases. They will be referenced in the following bold labels, across the paper, each respective to their comparative hardware and software specifications.

Also, no Graphical Processing Units (GPU) were utilized in favor of prioritising consistency and simulating real-world scenarios, but they can also be employed to provide additional detection acceleration.

- **High-performance desktop**: OS Windows 11, x64 Processor, Intel® Core™ i7-4790 CPU 3.60 GHz, 16 GB RAM.
- **Standard desktop**: OS Windows 10, AMD Ryzen 5 1400 Quad-Core CPU 3.20 GHz, 16 GB RAM.
- **Standard notebook**: OS Linux Mint 21.1 Cinnamon, Intel Core i5-7200U CPU 2.50GHz × 2, 8 GB RAM.
- **Raspberry Pi**: OS Raspberry Pi Full, 64 bits, system model 4B, Broadcom BCM2711 Quad core Cortex-A72 (ARM v8) 64-bit SoC 1.8GHz, 1 GB RAM.

2) Databases: The chosen databases were:

- **Extended Cohn-Kanade Dataset (CK+)** [9]: Database containing both grayscale and colorized images, taken in a controlled environment, with subjects ranging from eighteen (18) to fifty (50) years of age with varied genders and ethnicity. For this research, only labelled videos were utilized, the first image of a sequence was taken as a neutral expression, and the last three (3) as the respective labelled emotion;
- **AffectNet** [10]: Database containing colorized images, taken from varied internet sources in a non-controlled environment (in-the-wild), with respective manually labelled facial expressions. The database used for testing was the validation set, containing five hundred (500) images per facial expression.

B. Image Pre-processing

Image pre-processing is often used for two (2) purposes: for data augmentation, or to enhance the desired image processing. Since all the tested FER models were downloaded pre-trained, the data augmentation step was not necessary.

However, the processing enhancing steps were still desirable. For this, it was tested the added step of cropping the boundaries of a face from an image, aiming towards reducing the processing area to a minimum, as well as deploying a pose-alignment model, for horizontally aligning faces, with the objective of further normalizing data input.

These steps are depicted in Figure 1. From left to right, it is portrayed in the images the processes of: detecting the boundaries of a face, to cropping, horizontally aligning, and the optional histogram equalization.

For facial extraction, the chosen method was a Haar-cascade based model, from OpenCV [11]. In sequence, the chosen method for facial alignment was the face landmark predictor, from DLIB [12].

Furthermore, it is important to note that, even though in real-world scenarios – such as by receiving raw images from webcams – most pre-processing techniques would be useful, some steps were redundant for some databases. For instance, in AffectNet, validation images were already provided with the facial boundaries cropped, through a similar cascade-based OpenCV model to the one proposed – therefore rendering the additional pre-processing step unnecessary.

C. Facial Expression Recognition

As for the FER, the chosen models for testing will be listed, with a corresponding brief description.

- **Residual Masking Network** [13]: The FER is achieved through the usage of four convolutional Residual Masking Blocks, scoring facial features through the combination of feature maps in a fully connected layer, producing as an output one of seven facial expressions.
- **Deepface** [14]: Multi-purpose framework for facial attribute recognition. Allows for detection of age, gender, emotion and ethnicity, through the usage of state-of-the-art models [15]. In addition, the framework includes built-in image pre-processing (cropping and alignment), through the usage of the mentioned models. Images are also discriminated into seven (7) facial expression categories. For this study, the default VGG-Face model [16] was employed.

D. Data evaluation

The key parameters under comparison will be processing speed and accuracy of the FER models. For each database,

TABLE I
COMPARATIVE FACIAL EXPRESSION ACCURACY ACROSS DIFFERENT DATABASES AND MODELS

Database	Model	Facial Expression Accuracy (%)						
		Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
Ck+	RMN	85	85	56	96	92	86	85
	Deepface	13	12	20	85	48	47	72
AffectNet	RMN	62	45	40	82	55	58	61
	Deepface	26	06	32	76	30	16	52

the data will be tested through the usage of four (4) distinct computers, as well as two (2) recognition models, which amounts to eight (8) rounds of detection.

Since for a fixed database and model, the detection accuracy is not expected to vary, four (4) values will be compared (each for a different combination of database and detection model).

IV. RESULTS

The results are presented in two (2) different subsections: Accuracy, for data pertaining to matching facial expression recognition to labelled data; and Processing time, for the required runtime duration for processing the same data in different systems.

A. Accuracy

Initially, the primary accuracy metrics are presented in the Table I, where each row indicates the percentage of correct identifications for each facial expression, that is, the frequency of matching detected facial expressions to the respective ones labelled, corresponding to a database and detection model.

1) General Comparison: As for comparing both models, at first glance it is notable that the detection accuracy of *Happiness* performed consistently as the highest metric, while for facial expressions such as *Disgust* and *Fear*, the detection displayed a subpar performance. A direct comparison between the average accuracy is exhibited in the Table II.

TABLE II
COMPARISON OF DETECTION ACCURACY BETWEEN DIFFERENT MODELS ON EACH DATABASE.

Database	Model	Average Accuracy (%)
AffectNet	Residual Masking Network	57,6
	Deepface	34,0
Ck+	Residual Masking Network	83,6
	Deepface	42,4

In Table II, it is notable that for a single detection, the detection model Residual Masking Network performed with higher accuracy than the default model from the Deepface framework. For the AffectNet database, the performance difference was of 23.6%, while for the Ck+ dataset, the difference was of 41.2%.

However, though the average was notably different, the difference was mainly driven from the very low scores of accuracy for specific facial expressions, such as disgust. The comparison data can be seen in the Table III, where each row represents the average between the two (2) databases, for each facial expression.

TABLE III
AVERAGE DETECTION ACCURACY FOR EACH FACIAL EXPRESSION BETWEEN DATASETS AND DETECTION MODELS.

Expression	Model accuracy (%)	
	RMN	Deepface
Anger	73,5	19,5
Disgust	65,0	9,0
Fear	48,0	26,0
Happiness	89,0	80,5
Sadness	73,5	39,0
Surprise	72,0	31,5
Neutral	73,0	62,0

Where *RMN* stands for the average accuracy of the FER through the Residual Masking Network Model – of both datasets – and similarly is expressed in data in the *Deepface* column.

B. Processing time

The processing time across different computers is presented in the Table IV.

TABLE IV
PROCESSING TIME COMPARISON BETWEEN DIFFERENT COMPUTERS.

System	Database	Model	Processing time (s)
1	AffectNet	RMN	0.35 ± 0.04
		Deepface	0.09 ± 0.02
	Ck+	RMN	0.41 ± 0.02
		Deepface	0.13 ± 0.04
2	AffectNet	RMN	0.49 ± 0.04
		Deepface	0.08 ± 0.02
	Ck+	RMN	0.65 ± 0.04
		Deepface	0.14 ± 0.02
3	AffectNet	RMN	0.51 ± 0.03
		Deepface	0.10 ± 0.06
	Ck+	RMN	0.51 ± 0.04
		Deepface	0.18 ± 0.02
4	AffectNet	RMN	2 ± 2
		Deepface	0.2 ± 0.1
	Ck+	RMN	2.8 ± 0.5
		Deepface	0.4 ± 0.1

In contrast to the accuracy results, the Deepface model performed up to an order of magnitude faster than the Residual

Masking Network model. In the *Processing time* column, the median duration is expressed in seconds (s), with the respective following standard deviation.

Where, for each row in the *System* column, the number indicates the respective computers, ordered by ascending processing time:

- **1**: High performance desktop
- **2**: Standard desktop
- **3**: Standard Notebook
- **4**: Raspberry Pi

Further, analyzing the data on the Table IV, it is notable that the processing time deviated greatly for the Raspberry Pi, in comparison with the other systems. To the extent that the average of the median times between the first three (3) systems remained at 0.48 seconds for the Residual Masking Network model, and 0.12 seconds for the Deepface model. While, for the Raspberry Pi, the average was 2.4 and 0.3 seconds, respective to the both aforementioned models.

In average, the Residual Masking Network to Deepface processing time ratio (RMN:Deepface), was:

- 4:1, for the first three (3) computers;
- 8:1 for the Raspberry Pi;

Indicating that, as an example, for the Raspberry Pi, an image may take up to eight (8) times more seconds to be processed through the Residual Masking Network than the Deepface model.

V. Conclusion

The test results point towards the viability of both detection models. However, some applications may benefit in using one model rather than the other, for instance, if there is a strict time frame for the detection, where Deepface has been assessed to process images faster than Residual Masking Network. Nonetheless, the opposite is true for implementations which require a higher detection accuracy, which would point towards the usage of Residual Masking Network instead.

Across systems, Deepface performed several times faster than the RMN model. Which allows for, in real-world scenarios, using more than one detection from Deepface to compare to one recognition from RMN, favoring a weighted comparison between the two models, which may aid in bridging the gap between the detection accuracy of both models.

Overall, Deepface has shown good consistency in detecting the facial expressions of happiness and neutral, although it performed more poorly for the other facial expressions. Therefore, in applications where there is an interest in mainly the detection of happiness and neutral expressions, the usage of Deepface may be favored.

And, specifically for Raspberry Pi, the Deepface model performed more consistently, which could be beneficial. However, in the other systems where processing speed may not be a bottleneck, both models may be used interchangeably, based on the desired parameters.

Acknowledgment

Murilo de Souza Preto was under the Scientific Initiation program of UFABC in Edital 04/2021 PDPD and thereafter under UFABC Scholarship PIC program in Edital 04/2022. Additionally, this work was partially under FAPESP Grant 2022/10909-5.

References

[1] P. Ekman, "What scientists who study emotion agree about," *Perspectives on Psychological Science*, vol. 11, no. 1, pp. 31–34, 2016, pMID: 26817724. [Online]. Available: https://doi.org/10.1177/1745691615596992

[2] J. Stouten and D. De Cremer, "'Seeing is believing': The effects of facial expressions of emotion and verbal communication in social dilemmas," *Journal of Behavioral Decision Making*, vol. 23, no. 3, pp. 271–287, Jul. 2010. [Online]. Available: https://onlinelibrary-wiley.ez42.periodicos.capes.gov.br/doi/full/10.1002/bdm.659

[3] S. Trichas, B. Schyns, R. Lord, and R. Hall, "'Facing' leaders: Facial expression and leadership perception," *The Leadership Quarterly*, vol. 28, no. 2, pp. 317–333, Apr. 2017. [Online]. Available: https://www-sciencedirect.ez42.periodicos.capes.gov.br/science/article/pii/S1048984316301515

[4] M. A. Assari and M. Rahmati, "Driver drowsiness detection using face expression recognition," in *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Nov. 2011, pp. 337–341.

[5] Q. Meng, X. Hu, J. Kang, and Y. Wu, "On the effectiveness of facial expression recognition for evaluation of urban sound perception," *Science of The Total Environment*, vol. 710, p. 135484, 03 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0048969719354774

[6] C. Bustos, N. Elhaouij, A. Sole-Ribalta, J. Borge-Holthoefer, A. Lapedriza, and R. Picard, "Predicting driver self-reported stress by analyzing the road scene," *arXiv:2109.13225 [cs]*, 09 2021. [Online]. Available: https://arxiv.org/abs/2109.13225

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, jul 2022. [Online]. Available: https://doi.org/10.1109%2Ftaffc.2020.2981446

[8] "Free Image on Pixabay - Beard, Face, Man, Model, Mustache," 2016. [Online]. Available: https://pixabay.com/photos/beard-face-man-model-mustache-1845166/

[9] 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 *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 94–101.

[10] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, jan 2019. [Online]. Available: https://doi.org/10.1109%2Ftaffc.2017.2740923

[11] "OpenCV: Cascade Classifier," Jul. 2023, [Online; accessed 26. Jul. 2023]. [Online]. Available: https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html

[12] May 2022, [Online; accessed 26. Jul. 2023]. [Online]. Available: http://dlib.net/face_alignment.py.html

[13] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4513–4519.

[14] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICEET53442.2021.9659697

[15] "DeepFace: Closing the Gap to Human-Level Performance in Face Verification - Meta Research | Meta Research," Aug. 2023, [Online; accessed 13. Aug. 2023]. [Online]. Available: https://research.facebook.com/publications/deepface-closing-the-gap-to-human-level-performance-in-face-verification

[16] "Visual Geometry Group - University of Oxford," Aug. 2023, [Online; accessed 13. Aug. 2023]. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/software/vgg_face