Recognition of Emotions through Facial Geometry with Normalized Landmarks

Alessandra Alaniz Macedo
ale.alaniz@usp.br
University of São Paulo (USP)
FFCLRP, DCM, PPG-CA
Ribeirão Preto, São Paulo, Brazil

Leandro Persona
Fernando Meloni
leandro.persona@alumni.usp.br
melonifernando@yahoo.com.br
University of São Paulo, FFCLRP, DCM, PPG-CA
Ribeirão Preto, São Paulo, Brazil

ABSTRACT

Emotion recognition holds pivotal significance in human social interactions, as it entails the discernment of facial patterns intricately linked to diverse emotional states. The scientific, artistic, medical, and marketing domains have all demonstrated substantial interest in comprehending emotions, resulting in the emergence and refinement of techniques and computational methodologies to facilitate automated emotion recognition. In this study, we introduce a novel method named REGL (Recognizing Emotions through Facial Expression and Landmark normalization) aimed at recognizing facial expressions and human emotions depicted in images. REGL comprises a sequential set of steps designed to minimize sample variability, thereby facilitating a finer calibration of the informative aspects that delineate facial patterns. REGL carries out the normalization of facial fiducial points, called landmarks. Through the use of landmark positions, the reliability of the emotion recognition process is significantly improved. REGL also exploits classifiers explicitly tailored for the accurate identification of facial emotions. As related works, the outcomes of our experimentation yielded an average accuracy over 90% by employing Machine Learning algorithms. Differently, we have experimented REGL with varied architectures and datasets including racial factors. We surpass related works considering the following contributions: the REGL method represents an enhanced approach in terms of hit rate and response time, and REGL generates resilient outcomes by demonstrating reduced reliance on both the training set and classifier architecture. Moreover, REGL demonstrated excellent performance in terms of response time enabling low-cost and real-time processing, particularly suitable for devices with limited processing capabilities, such as cellphones. We intend to foster the advancement of robust assistive technologies, facilitate enhancements in computational synthesis techniques, and computational resources.

KEYWORDS

Multimedia Processing, Affective Computing, Machine Learning-Multimodal Interaction, Facial Patterns, Image Understanding.

1 INTRODUCTION

The recognition of emotions is an intrinsic part of human relationships. Even in early childhood, children learn to map and interpret the emotions of others, utilizing the result as an indicator of their surrounding context [11, 23, 27]. The mapping is possible because humans typically translate their emotions into detectable physical movements, particularly facial expressions, which are vital for social interactions. Facial expressions are ubiquitous behaviors and exhibit weak dependence on cultural factors as demonstrated by [13]. Thus, same researchers identified the following seven different universal emotions across cultures: fear, anger, sadness, happiness, surprise, disgust, along with the neutral emotion [13]. Due to their distinct and stable patterns, facial expressions can be recognized even in unfamiliar individuals [1]. As a result, recognition can be organized with significant potential for automation, enabling machines to interpret a select range of human emotions.

In recent years, facial expression-based methods for emotion recognition have witnessed rapid advancements due to the growing scientific, medical, and commercial interest in the field [22, 26, 43, 50]. The most common approaches for recognizing individuals and facial expressions focus on automated pattern detection in digital images. Notably, social media platforms, mobile devices, and digital cameras, now possess the capability to discern whether a person is smiling or not [5]. Assistive technologies are developed to aid individuals with behavioral syndromes (e.g., autism and mood disorders) [6, 26, 36]. By enabling individuals with disabilities to react differently upon perceiving expressed emotions, these technologies contribute to enhancing their social interactions. In summary, automatic facial expression recognition offers new avenues for interaction between humans and machines, facilitated by the detection of both voluntary and involuntary facial movements.

In general, the interpretation of information from digital images involves automation through Machine Learning (ML) [7]. However, image manipulation requires initial pre-processing steps [44] to enable the detection of specific shapes such as a human face. Next pre-trained models scan segments of the image and use probabilities to confirm patterns [48]. Some models are capable of detecting human faces in images with high accuracy rates, above 90% [45, 48]. Once the face is identified, the next step in evaluating facial expressions is to map the Regions of Interest (ROIs), which are specific facial structures such as eyes, nose, mouth, chin, etc. The relative positions of these structures are then marked as landmarks, which are assigned bi or tridimensional coordinates, adding an additional layer of information.

The use of landmarks provides a straightforward and objective way to recognize facial expressions by comparing patterns for both facial identification (facial recognition) and changes in facial

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2966-2753

patterns (expression and emotion recognition). Although implementations may vary, both approaches consider the coordinates of landmarks as problem variables (reference values) for calculating distances (e.g., Euclidean, cosine, etc.) [15, 29, 32]. In the case of emotions, the muscle movements responsible for each facial expression cause geometric changes in the relative positions of ROIs [49]. These changes alter the coordinates of landmarks and the values of distances between them. Given that a neutral face exhibits different relative distances from a fearful or smiling face, the set of variations can be utilized to identify expressions [32]. The generalizability of emotional responses among humans tends to simplify the problem, enabling the identification of which distances are most affected by each type of facial expression. At the end of the analysis of these variations, ML classifiers can be trained for the recognition of human emotions [48].

Despite the current methods yielding satisfactory results for specific contexts, such as photos acquired in controlled environments, there are still significant challenges related to the subject. The main problems involve the variability of patterns found in facial shapes, poses and in images obtained under real-world conditions (uncontrolled variables such as lighting, brightness, distance from the capturing device, etc.) [32, 41]. Differences in facial shapes tend to introduce noise into the problem, making training dependent on factors such as race, age, sex, etc. Aspects like focal length, luminosity, framing, people's poses (rotated face images and expression of sentiment), and hardware configurations significantly affect the concentration and location of pixels, adding noise to the data. These sources of variability negatively impact the performance of classifiers, making the ML process more challenging [7, 32]. Our hypothesis is that the reduction of variability in the data can contribute to improving computational performance. However, the domain of information in images is represented by pixels, and thus traditional techniques for data alignment and normalization need to be adapted for this context, requiring computationally creative and conceptually elaborate strategies.

According to Oge and Gonzales, one of the major challenges faced by algorithms in Digital Image Processing (DIP) in uncontrolled environments is the difficulty of achieving good results under varying conditions of luminosity and contrast [14, 18]. These conditions introduce noise and variability into the data, thereby posing significant challenges for algorithm performance. Additionally, the relative positioning of objects within the scene tends to generate variability, further reducing the efficiency of the algorithms.

In this paper, we present a novel method for emotion recognition in digital images, called REGL (Recognizing Emotions through Facial Expression and Landmark normalization), which operates under various conditions of variability, including scale, rotation, racial factors, and image acquisition. The method incorporates well-known image manipulation techniques such as histogram equalization and facial alignment to harmonize the information and enable appropriate normalization. A set of normalization steps is performed to reduce data variability, which finally produced an optimized context to identify facial expressions. Subsequently, classifiers suitable for the methodology of the study are employed to assess potential performance gains. The processing and normalization steps aimed at reducing data variability and simplifying the problem. As mentioned, we hypothesized that reducing variability would lead to classifiers with improved performance, regardless of the ML method employed. Thus, the development of methods that harmonize image information has the potential to yield better results.

The remainder of this paper is structured as follows: a brief review of image deblurring and attention mechanism is provided in Section 2. The proposed REGL method is discussed in Section 3. The results of the experiments are presented in Section 4. Section 5 brings related work. Finally, Section 6 provides final remarks and future work.

2 BACKGROUND

In the early 1970s, Paul Ekman conducted a groundbreaking scientific experiment that revolutionized the field of human emotion recognition [13]. At that time, it was believed that individuals used their facial muscles according to a set of social conventions and expressions shaped by societal interactions, much like languages, with each region of the world having its own variations.

Ekman captured numerous images of men and women displaying various facial expressions. He then traveled to Brazil, Argentina, and Japan to conduct his experiments. To his surprise, individuals from different countries obtained the same results in classifying the images. The experiment was further extended to the forests of Papua New Guinea in Oceania, reaching the most remote and isolated villages. Even among the inhabitants of these regions, the results did not differ, leading to the conclusion that human emotions expressed through facial expressions are universal and independent of ethnic and social factors.

Another significant contribution of Ekman's work was the creation of the Facial Action Coding System (FACS), a system for classifying human emotions. FACS provides a standardized framework for systematically categorizing the physical expression of emotions, allowing for the labeling of any anatomically possible facial expression. Figure 1 presents the six primary emotions, along with the neutral expression, subdivided into action units, which are the building blocks of FACS and their main differences.

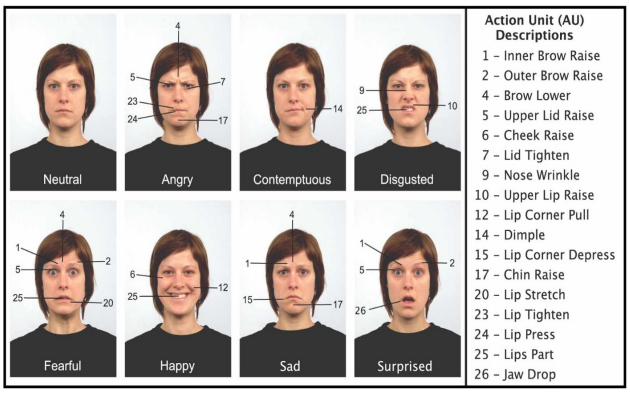

Figure 1: Illustration of FACS (Facial Action Coding System) [27]

This section offers the necessary context for our method by providing general information about the topic of our REGL method. To effectively present this context, we have divided the section into *Imaging Process, Face Detection, Landmarks Setting* subsection, and *Classification* subsection. Each subsection focuses on a specific aspect, providing a comprehensive overview of the underlying concepts and techniques related to our proposal.

2.1 Imaging Process, Face Detection, and Landmarks Setting

The real-time process of interpreting emotions from images includes managing images, e.g. pixel normalization, detecting faces, extracting features, and classifying the patterns [7, 18]. The initial pre-processing of images works out to minimize noise, like pixels acquired outside the standard [44], reducing variance from data. This step allows training more precise classifiers and achieves good performance even if limited datasets are employed for training.

The next step is analyzing human faces in images, which starts by detecting the region of the image that contains the aimed object [25]. In the early 2000s, Viola and Jones developed a method capable of detecting human faces with sufficient speed and accuracy for use in popular cameras [45]. Currently, more efficient and reliable solutions exist, such as the use of Gradient Histograms oriented (HOG), specifically for this purpose [16, 34]. HOG is a digital image processing technique widely used in computer vision and multimedia processing. It characterizes objects by their shape and texture, by evaluating the distribution of the density gradient or the edge directions [10]. It has been used as a reference in the facial detection and recognition of various types of objects [18]. Models trained to detect human faces in images can scan segments through the image, employing probabilities to determine bounds and confirm the whole pattern [48]. Some available models can detect human faces in images with high accuracy, even in low-quality images such as [2, 41, 45, 48].

Once face identification is confirmed, the next step is to identify the Regions of Interest (ROIs), like eyes, mouth, nose, chin, or any relevant structure related to facial expressions. The facial landmarks are subsets of the shape prediction problem, which involves locating key points and the overall shape of an object. For instance, the DLib library, which is available in C, C++, and Python, is commonly used to extract landmarks and provides a feature matrix of the x and y coordinates with 68 facial points, including eyes, nose, mouth, and face bounds. Landmarks may assume bi or three-dimensional coordinates according to the implementation, and their relative positions in the face carry information from relative distances of face structures, which can be useful for face recognition or emotion analysis. Regarding facial expressions, muscular movements lead to changes in the relative positions of face structures and, therefore, the relative distances between landmarks shall reflect the face features [44]. For this reason, relative distances between landmarks are useful for objectively analyzing facial expressions and emotions, or even for face recognition. In both cases, the landmark coordinates serve as reference values for calculating distances using various metrics such as Euclidean, Manhattan, or Minkowski distances [15, 30, 32], despite each implementation may greatly differ from the others.

The use of fixed common distances, such as the distance between the eyes, to normalize facial landmarks in images with different scales has been proposed as a way to make the landmarks comparable across all faces [24]. However, this normalization approach is not highly effective in reducing variability related to image acquisition conditions, particularly rotation and racial variations. Certain individuals may exhibit prominent facial characteristics that can negatively influence the performance of Machine Learning algorithms. For instance, the width of the mouth can be a critical factor in identifying the emotion of happiness, which is often manifested through a smile. In such scenarios, a classifier trained on the character Joker, the villain from the Batman superhero series, would face difficulties in converging and distinguishing a person's expression of happiness when using the Euclidean distance metric for landmark normalization.

Figure 2 illustrates the automatic localization of 68 facial landmarks, which aim to identify various structural components of the human face, including the face contour, eyebrows, eyes, nose, mouth, and others. Regarding the number of detected landmarks, the range between 60 and 80 coordinates predominates, appearing in nearly half of the studies in the literature.

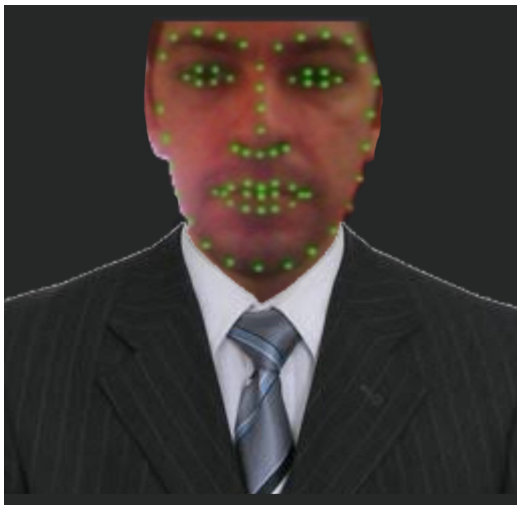

Figure 2: Locations of all 68 facial landmarks (adapted [35]).

With the extraction of facial landmarks, the Image Processing stage is completed. The processed data transitions from the pixel space of the image to the two-dimensional (x and y) coordinates representing the positions of the facial landmarks, reducing the dimensionality of the problem being studied. Mathematically, rotation is a linear transformation that involves spatial coordinates, in this case, in two dimensions, preserving the magnitude of vector lengths and orientation in physical space [47]. Therefore, using coordinates tends to introduce less variability than using inter-landmark distances, while also avoiding the additional processing time required to compute distance metrics.

Frontalization techniques enable the artificial synthesis of frontal views for various types of objects, including human faces, from rotated original sources. Their use tends to substantially improve the performance of classification and recognition systems that rely

on images by normalizing the variations that may occur during the acquisition process [20]. Additionally, the training and testing stages employed by Machine Learning algorithms can be performed with standardized samples in a consistent position. Image frontalization techniques can be classified into two distinct forms, as stated by [46]: appearance-based frontalization, where the original image is rendered to the frontal view, and coordinate-based frontalization, which uses a model trained with images of the object under study.

2.2 Classification

Machine learning classifiers are algorithms that learn patterns and relationships within data to make predictions or classify new instances. These classifiers play a crucial role in emotion recognition, such as detection of smiles, fraud detection, and recommendation systems. There are several types of machine learning classifiers, each with its own structures, rules, processes, and reasoning. The following five classifiers were exploited by our proposal.

The *Perceptron* is a connectionist algorithm that can effectively identify complex nonlinear relationships between input and output data [51]. It has only one input and one output layer and does not have any hidden layers. The K-nearest neighbor (KNN) algorithm is an instance-based learning classification technique that classifies a sample based on the labels of its k-nearest neighbors in a feature space [4]. KNN algorithm is usually implemented using the Euclidean distance as the distance metric between the samples.

The *Decision Tree* (D3) algorithm creates rules for learning and mimics human logical reasoning for classification, by splitting the problem's attributes based on the highest performance gain at each division[21]. Decision trees are made up of a collection of nodes that store information at their ends. The Random Forest (RFC) algorithm is an ensemble learning method that constructs multiple decision trees, each of which is trained on a random subset of the input features and data samples [40]. The final prediction is then made by aggregating the predictions of all individual decision trees. This method can reduce overfitting and improve the model's generalization ability[12].

The *Multilayer Perceptron (MLP)* is a neural network structure that consists of multiple layers, including an input layer, one or more hidden layers, and an output layer[21]. MLP can handle nonlinearly separable problems due to the nonlinearity introduced by the activation functions in each neuron. SVM stands for *Support Vector Machine*, which is a powerful and widely used supervised machine learning algorithm. It is primarily used for classification tasks but can also be adapted for regression and outlier detection.

3 REGL

The REGL (Recognizing Emotions through Facial Expression and Landmark normalization) method aims at recognizing facial expressions and human emotions depicted in digital images. The method extracts relative positional data from facial structures and calculates a more accurate measure of facial muscle movements, such as those produced by facial expressions. The steps of the REGL method, as shown in Figure 3, are as follows:

(1) **Histogram Equalization** for reduction of radiometric variability (brightness and contrast).
(2) **Gray Scale Conversion** for reduction of the three color channels to a single channel.
(3) **Facial Detection** for reduction of the search area by delineating the Region of Interest.
(4) **Landmark Extraction** for changing in data dimensionality, where pixels are replaced by the two-dimensional coordinates of landmarks. After the changes, the domain comprises only 136 variables (68 coordinates for the x-axis and another 68 for the y-axis).
(5) **Min-max Normalization of Coordinates** for reduction of the scale factor, which accounts for the proximity difference between the actor and the capturing device. Innovation in the emotion recognition process.
(6) **Frontalization** for reduction of geometric variability caused by various rotations in both the x-axis and y-axis. The coordinates are adjusted to simulate a frontal position.
(7) **Normalization by Actor's Face (Delta standard)** eliminates anatomical variations.
(8) **Coordinate Vector** for concatenation of the coordinates into a vector to facilitate the induction of Machine Learning algorithms. The final dimension of the vector is 1 row for each actor with 136 columns, i.e., 68 columns for the x-coordinates and 68 columns for the y-coordinates.

The first two steps involve pre-processing tasks to enable facial detection. After, the REGL method encompasses three main stages – (1) extraction and normalization of landmark coordinates, (2) frontalization, and normalization, and (3) extraction of relative position measures of the facial expression in the image with respect to the neutral face of the actor. These steps precede the construction of Machine Learning-based algorithms, i.e., they pertain to data pre-processing aimed at achieving better standardization of input information. Next, we present details considering each main stage.

3.1 Manipulation of Landmarks

After delineating the Region of Interest in each image, the landmark extraction occurs changing pixels by the two-dimensional coordinates of landmarks, as mentioned. The advantage is the reduction in data dimensionality. Furthermore, to minimize the effects of scale variability of the cordinates, the coordinates are normalized using the minimum and maximum (min-max) values for each axis, as described by Equation 1 for the horizontal coordinates on the x axis and Equation 2 for the vertical coordinates on the y axis. This normalization process has a constant computational cost of O(1), and it is defined as follows:

$$\widehat{x_i} = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

$$\widehat{y_i} = \frac{y_i - min(y)}{max(y) - min(y)} \tag{2}$$

Thus, all x and y coordinates are scaled to values between zero and one, minimizing the scale effect. It is important to note that this transformation does not alter the proportions between the coordinates. Therefore, the rescaled face remains similar to the original, enabling its use for facial recognition purposes. Hence, the methodology presented here requires prior knowledge of the actor depicted in the image.

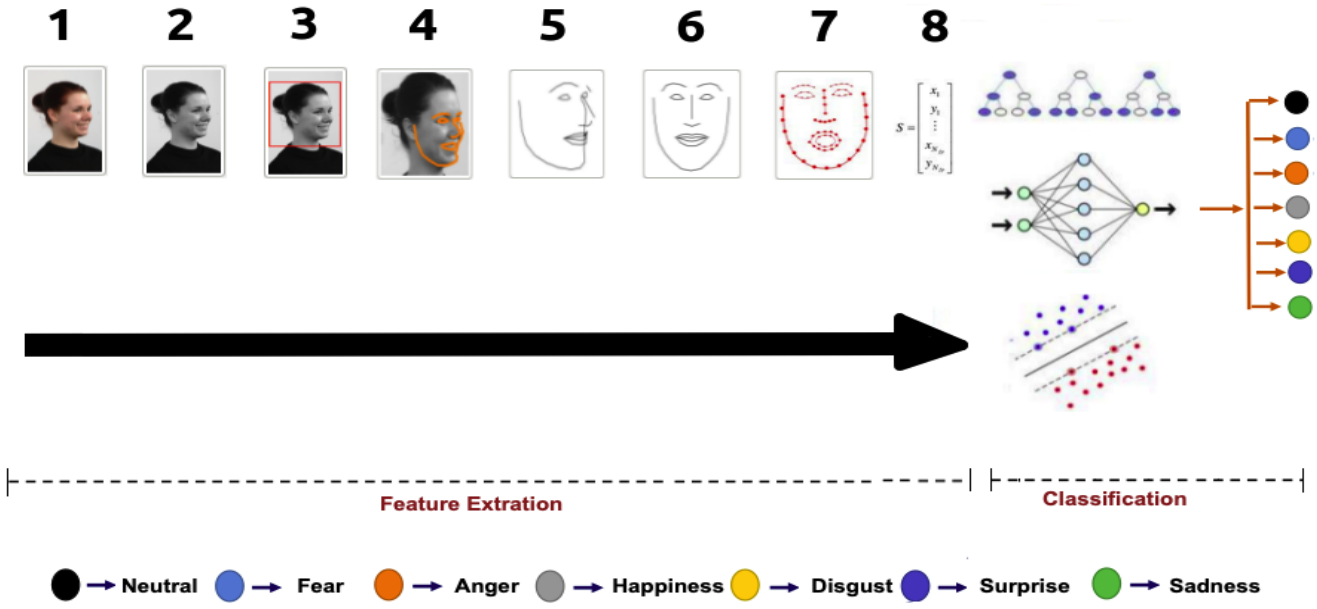

Figure 3: REGL: Recognizing Emotions through Facial Expression and Landmark normalization. Steps: 1 - Original image (raw data); 2 - Image processing to gray-scale image; 3 - Pre-trained model recognizes the face in the image; 4 - Landmarks are set; 5 - Another pre-trained model transforms the 2D landmarks into 3D landmarks; 6 - The 3D landmarks are rotated to the frontal position; 7 - A new round of face recognition and landmarks setting is performed; 8 - Coordinates of the evaluated expression are compared with coordinates of the neutral face of the same person, and the delta divergences are used to create a data vector and train classifiers.

3.2 Coordinates Frontalization and Normalization

The process of frontalization aims to reduce variations in face rotation, both horizontally and vertically, with respect to the image plane, by centering the pose in a frontal manner. Before the frontalization, it is necessary to recognize the variations in facial expressions.

The REGL method first needs to recognize variations in facial expressions because anatomical differences among different actors in images can be a significant source of noise, making it challenging to detect morpho-geometric patterns of the face when aiming for emotion recognition. In light of this, we propose evaluating the variation in patterns when a facial expression is performed. This approach requires a neutral facial pattern instead of evaluating static morpho-geometric patterns, as is typically done in facial expression detection [9, 33, 37]. Specifically, we propose that a frontal image A of the resting face (neutral facial expression), serving as a reference, undergoes facial detection, landmark extraction, coordinate normalization, and frontalization, resulting in the generation of 136 reference coordinates, which form the vector $\overrightarrow{A}$.

A second image B, the subject of evaluation, undergoes the same process, generating another set of 136 coordinates and forming the vector $\overrightarrow{B}$. This leads to the creation of the final information vector, $\overrightarrow{\Delta}_{AB}$, consisting of 136 variables, where each coordinate i is obtained as $\Delta_{AB}i = Ai - Bi$. Therefore, $\overrightarrow{\Delta}_{AB}$ contains the information about the relative variation of the frontalized coordinates of B with respect to the coordinates of the resting face A.

Since the expression of interest in B is always compared to the expression of the same actor in a neutral position, which has been previously labeled and known in A, problems of anatomical variability are minimized. As an advantage, it is possible not only to classify a static geometric pattern, such as the geometric pattern characterizing a smile, but also to **quantitatively measure the deformation produced by this movement** (See Figure 3 - steps 6, 7 and 8). We began the recognition of emotions by detecting smiles, as detailed in [35]. Next, we present our overall algorithm that implements the REGL method.

3.3 An Overall Algorithm

In summary, Algorithm 1 outlines the technical steps of facial emotion recognition in the REGL method. The input consists of all the images from the facial expression databases as Cohn-Kanade, RafD, NimStim, KDEF, and Jaffe (see further), and the output is a reusable Machine Learning Model (MLM) for emotion recognition.

The algorithm iterates over all input images G (line 2) and for each image (line 3), initiates image processing and facial detection (line 4). If successful (line 5), the processing to reduce data variability

runs sequentially landmark extraction (line 6), normalization of x-coordinates using min-max normalization (line 7), normalization of y-coordinates using min-max normalization (line 8), frontalization of coordinates (line 9), and normalization by the actor's face (delta pattern) (line 10). Finally, the coordinates are inserted into the information vector (line 11). After processing all the images, the final vector undergoes the classification process and the creation of the MLM (line 14).

Algorithm 1 - REGL

```
                    Input: Datasets (G)
                  Output..: ML Models (MLM)
1:  BEGIN
2:  WHILE G has images
3:      Load Image G(I)
4:      face ← face detection G(I)
5:      IF face ≠ ∅ SO
6:      coordinates ← Extraction of 68 landmarks
7:      coordinates ← Norm. coordinates x
8:      coordinates ← Norm. coordinates y
9:      coordinates ← Coordinates Frontalization
10:     coordinates ← Normalization (Δ)
11:     coordinatesArray ← coordinates
12:     END-IF
13: END-WHILE
14: MLM ← Classifier(coordinatesArray)
15: RETURN MLM
16: END
```

4 EXPERIMENTS AND RESULTS

This section presents the experiments conducted considering the datasets used, the obtained results, and the discussions regarding the processing of the REGL method.

4.1 Datasets

A facial expression database is a collection of digital images or video clips featuring different actors. Its content is essential for training, testing, and validating Machine Learning (ML) algorithms, as well as for the development of facial recognition and facial expression recognition systems, which encompass emotions. These databases of images are often guided by the theoretical basis of human emotions [1], which assumes the existence of six different types of facial expressions: happiness, fear, disgust, anger, surprise, and sadness, in addition to the neutral (or indifferent) expression. To experiment REGL, we selected different databases aiming to augment the variability of noises in the manipulated images. The free facial expression databases used in our study were:

- **Extended Cohn-Kanade Dataset (CK+):** attempts to standardize facial expressions according to the reference of FACS proposed by [13]. It contains frontal images with all seven universal emotions. CK+ was published in 2000's by [8].
- **The Japanese Female Facial Expression (JAFFE) Database:** comprises 210 images displaying the seven universal facial expressions, presented by ten Japanese women [31] in a frontal position. The images were obtained by the Department of Psychology at Kyushu University, Japan.
- **Radboud Faces Database (RafD):** consists of images of 67 actors (including Caucasian men and women, European Caucasian children, and Moroccan men) [28]. Following the FACS methodology, all actors were trained to express the following emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral.
- **The Karolinska Directed Emotional Faces (KDEF):** comprises 4,900 images with all universal emotions of seventy actors [17]. Each expression is captured from five different angles. The image set was developed by the Department of Clinical Neuroscience, Psychology Section, at the Karolinska Institute in Sweden.
- **The Nimstim set of Facial Expressions:** is well-known in medical literature [42]. Although it is rarely used for emotion recognition, it has received over 2,000 citations in scientific papers. Nimstim contains 672 frontal images of 43 professional actors, including 18 women and 25 men, aged between 21 and 30 years. All universal emotions are represented.

The reason for using different datasets was twofold: (i) to manipulate various ethnicities, ensuring that the classifier could generalize without being influenced by ethnic traits, and (ii) to consider sources of data variability, including variations in geometry, lighting, and rotation.

4.2 Training and Evaluation

The proposed REGL method was evaluated using the mentioned datasets in its sequence of steps culminates in the creation of a feature vector, comprising a set of 136 coordinates, after performing all normalization steps, which serves as the input for the SVM, MLP, Random Forests, Extra Trees, and Decision Tree Machine Learning algorithms [1]. For Decision Tree, the maximum node depth was set to 4. Depths below this configuration failed to generalize the categories effectively, while increasing the depth did not improve the results. Both Random Forests and Extra Trees used the same depth configuration as Decision Trees, with the number of trees set to 1000. Fewer trees did not consistently converge on the categories, while more trees exponentially increased execution time, making processing impractical. The MLP algorithm was configured with three main settings: 3000 iterations, a learning rate of 0.001, and the Adam optimizer to activate the neural network. This setup produced results proportional to processing time and model accuracy. The SVM algorithm used a radial basis function kernel to separate the categories and a C parameter of 5 to penalize incorrect classifications. These algorithms assessed the quality of the extracted features and the classification of emotions [2].

[1]The optimal parameters for the machine learning algorithms were exhaustively tested on the HPC (high-performance computing) server Aguia4 at the University of São Paulo. The cluster consists of 128 physical servers, each with 20 cores and 512 GB of RAM. The processor used is an Intel(R) Xeon CPU E7-2870 at 2.40 GHz, with a 256 TB filesystem for temporary files.

[2]All algorithms were implemented and processed on a laptop with an Intel(R) Core i5 eighth-generation processor, model 8265U at 3.9 GHz, with 8 GB of RAM, a 128 GB SSD, and an integrated Intel UHD graphics card. The software environment included Linux Ubuntu 18.04 and Python 3.6.9.

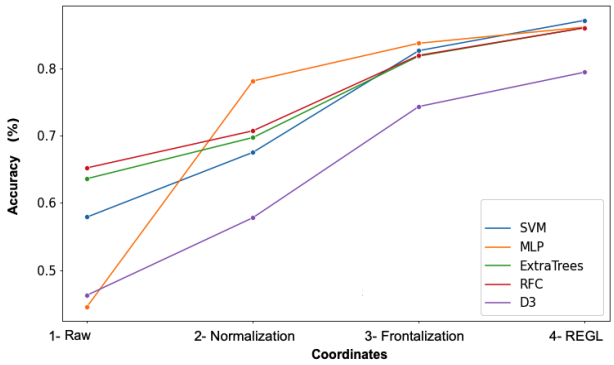

Figure 4: Accuracy Evolution - REGL Method

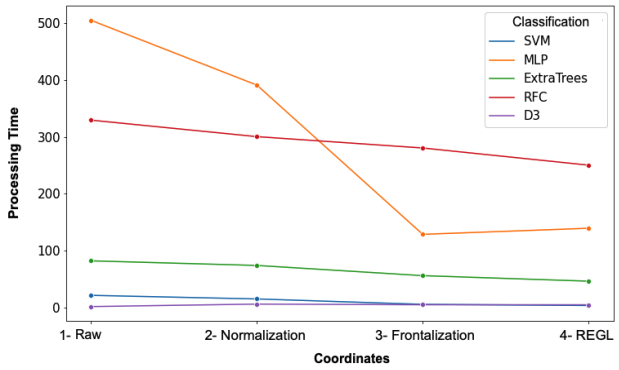

Figure 5: Processing Time - REGL Method

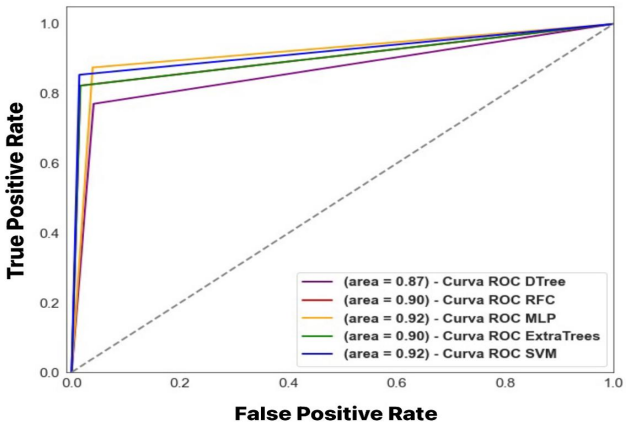

Figure 6: ROC Curve - Emotion Recognition with the REGL Method

To present the results, graphics depicting the percentages of accuracy and error were employed. ROC (Receiver Operating Characteristic) curves were created based on the true positive (TP) and false positive (FP) rates to represent the discriminative capability of the classifiers, i.e., their ability to correctly classify specific emotions. In the ROC graphics, points closer to (0,1), indicating higher TP and lower FP, represent a more consistent and generalizable classification.

Figure 4 showcases the results of emotion recognition through the sequential execution of the REGL method steps. It is worth highlighting that the succession of coordinate normalization techniques played a crucial role in the improvement and accuracy of the results, irrespective of the algorithm employed for classification.

The best performance of the REGL method was achieved using the SVM algorithm, yielding an accuracy of 87.1% and a processing time of only 4.33 seconds. It is important to note that the experiments utilized 10-fold cross-validation during training. It suggested that emotions are also universally recognizable by machines, and their recognition can be effectively performed artificially.

Another important aspect to be analyzed regarding the REGL method is the processing time of the algorithms in relation to the chaining of normalization steps. According to Figure 5, the data is consistent and demonstrates that the steps of the REGL method decrease the processing time of all the experimented algorithms. SVM and D3 achieved the best constant performance. Therefore, the reduction of sample variability in the coordinates, provided by min-max normalization, frontalization, and normalization by actor's face, contributes to achieve a good performance of ML algorithms responsible for human emotion recognition and optimizes the processing time required in each step.

Figure 6 presents a comparative analysis of the performance of various algorithms used in the emotion recognition process, after implementing all the variability reduction steps, through a ROC graphic. The dashed diagonal line represents the performance of a random classifier and serves as a reference for evaluating the others. Points above the diagonal in the ROC space represent better classification than points below the diagonal. A perfect classifier is one that produces a point close to $(0, 1)$ and an area close to 1. This means that the classifier has a zero false positive rate and a 100% true positive rate. Analyzing Figure 6, the ROC curves of the SVM and MLP algorithms yielded more accurate results than the others. In conclusion, the average performance of SVM was the best for all measures evaluated.

5 RELATED WORK

Methods of emotion recognition encompass a variety of approaches aimed at identifying and categorizing human emotions based on observable cues. Differently of REGL, these methods do not usually explore these set of techniques of manipulation of landmarks, frontalization and coordinates normalization sequentially. Here, we present related work of our proposal. In the Background section, we have already described methods used for face detection and landmark extraction.

In Testa et al., the authors conducted a literature review of studies that discover the synthesis of facial expressions through landmark extraction [41]. The predominant facial structures detected in these studies include the eyes, eyebrows, mouth, face contour, and nose. The authors observed that machine learning approaches are the most widely used and few studies used metrics to evaluate the results.

Álvarez, Luengo, and Lawrence presented a method for emotion recognition using the Euclidean distance between landmarks in the eye and mouth regions [1]. The technique of relative error distribution was employed as a validation method for accuracy. The final average accuracy and standard deviation were respectively 94.53% ± 2.47% using just the LabeledFacesInTheWild (LFW) dataset. Cui developed a method for smile recognition in images containing people, utilizing the Euclidean distance as an attribute to train an algorithm called Extreme Learning Machine [9]. An accuracy of 93.40% was achieved on the LFW dataset too.

Salman, Madani, and Kissi proposed a facial expression recognition method for training a decision tree called Classification and Regression Tree, which utilizes the measurements of the distances between the width and height of the mouth as the feature vector [39].

Hassner et al. explored a simpler approach using an unmodified 3D reference surface, which approximates the shape of all input faces [20]. This facilitated the development of a direct, efficient, and easily implementable method for frontalization. Importantly, it generated aesthetically pleasing frontal views and proved surprisingly effective for face recognition and gender estimation. In the LabeledFacesInTheWild (LFW) dataset, 97.5% of the 1432 images were successfully frontalized.

Jia et al. published a review article presenting spontaneous and posed facial expression databases and various computer vision-based detection methods, including those specific to smile detection [38]. They highlighted the importance of generalization ability in emotion detection and the detection of specific emotions based on unique facial features. Considering multimodal input, Bohy et al. developed a deep learning-based multimodal smile and laugh classification system, considering the use of audio and vision-based models as well as a fusion approach [3]

Guyon and Elisseeff demonstrated, through tests with linear and nonlinear classifiers, that feature selection reduces the variability of the studied problem [19]. This is because certain classifiers struggle with duplicate data, and as a result, accuracy tends to increase.

The choice of method for emotion recognition depends on several factors, including the available data, application context, and desired accuracy. Using different datasets, REGL demonstrated equivalent accuracy to other similar methods measured that considered specific datasets. When compared to related works, REGL addresses a different set of techniques and shows excellent response time, which is a fundamental requirement for real-time emotion recognition in mobile applications. This is due to normalized landmarks are generalized two-dimensional coordinates, enabling low-cost and real-time processing, particularly suitable for devices with limited processing capabilities, such as cellphones.

6 FINAL REMARKS

Emotions provide our first means of nonverbal communication developed throughout our lives. Through emotions, humans are able to interact with others and the environment in which they are immersed. This interaction is possible because humans almost always translate their emotions into detectable physical movements, such as facial expressions, which are essential for social interactions.

Despite being a trivial mechanism easily recognized by all human beings, emotion recognition is a challenging task for machines and computers. Thus, the aim of this work was to develop an artificial emotion recognition method called REGL, to extract and analyze the morphometric characteristics of the facial region, similar to the method used by human beings. REGL combined various techniques of digital image processing and statistical methods to evaluate and reduce variability in the images used. Among them, noteworthy are the normalization of coordinates using min-max, capable of optimizing the effects of scale factor, frontalization, responsible for reducing the effects caused by face rotation, and delta normalization, which uses the actor's neutral face to identify other emotions, thereby minimizing the effects of anatomical and racial variations.

The results of our experiments indicated the efficiency of the REGL method in the classification and artificial recognition of human emotions, resulting in an accuracy rate above 90% for all processed facial expression from diverse databases. Regarding only smile detection, a final accuracy rate close to 95% was achieved, surpassing previous works in the literature. REGL surpassed time processing of methods that evaluated this variable. Here, we did not exploit quantitative aspects of REGL because our focus was classification. However, REGL can measure the intensity and speed of movements. We have used it in SofiaFala [3].

Convolutional Neural Networks has been explored for emotion recognition despite their drawbacks: high computational demands, the need for large labeled datasets, sensitivity to data variability, overfitting due to unbalanced data, and difficult to interpret. These limitations suggest that alternative methods as REGL may be useful.

The main contribution extracted from this work is the creation of the min-max normalization technique for facial landmark coordinates, reducing the scale factor effect in images. The main limitation involving the REGL method is associated with the fact that REGL is tailored for contexts whose the identity of the actor is assured. Therefore, experiments considering uncontrolled environments will demand coupling a routine of face recognition. Moreover, this work will be continued in terms of: (i) analyzes of the performance of the REGL with a greater number of facial landmark coordinates; (ii) analyzes of the REGL with three-dimensional coordinates of facial landmarks; (iii) evaluation of the performance of the REGL method using other databases, preferably with black and Asian actors because the used databases did not incorporate images with them; (iv) assessment of the performance of the REGL method in video-based applications, (v) investigation of facial recognition with coordinates normalized by min-max and frontalization; (vi) use of the quantitative responses of the REGL method, especially considering fear and surprise.

[3] https://dcm.ffclrp.usp.br/sofiafala/

ACKNOWLEDGMENTS

This research was developed with the help of HPC resources provided by the Information Technology Superintendence of the University of São Paulo.

The presented method was discussed in the SofiaFala project supported by CNPq (Brazil), through the process 442533/2016-0, which supports the development of an application for mobile devices to assist children with Down Syndrome, to support training speech therapy for speech development.

REFERENCES

[1] Mauricio Alvarez, David Luengo, and Neil Lawrence. 2013. Linear Latent Force Models Using Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 11 (Nov 2013), 2693–2705. https://doi.org/10.1109/TPAMI.2013.86

[2] Mitra B., Sharma K., Acharya S., Mishra P., and Guglani A. 2022. Real-time Smile Detection using Integrated ML Model. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, Madurai, India, 1374–1381. https://doi.org/10.1109/ICICCS53718.2022.9788399

[3] Hugo Bohy, Kevin El Haddad, and Thierry Dutoit. 2022. A New Perspective on Smiling and Laughter Detection: Intensity Levels Matter. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. https://doi.org/10.1109/ACII55700.2022.9953896

[4] Guilherme Campos, Arthur Zimek, Joerg Sander, Ricardo Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael Houle. 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery* 30 (07 2016). https://doi.org/10.1007/s10618-015-0444-8

[5] V. Chaugule, D. Abhishek, A. Vijayakumar, P. B. Ramteke, and S. G. Koolagudi. 2016. Product review based on optimized facial expression detection. In *2016 Ninth International Conference on Contemporary Computing (IC3)*. IEEE, Noida, India, 1–6. https://doi.org/10.1109/IC3.2016.7880213

[6] Yufang Cheng and Shuhui Ling. 2008. 3D Animated Facial Expression and Autism in Taiwan. In *IEEE International Conference on Advanced Learning Technologies (ICALT 2008)*. IEEE Computer Society, Los Alamitos, CA, USA, 17–19. https://doi.org/10.1109/ICALT.2008.220

[7] Francois Chollet. 2017. *Deep Learning with Python* (1st ed.). Manning Publications Co., Greenwich, CT, USA.

[8] Jeffrey Cohn and Takeo Kanade. 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*, 94 – 101. https://doi.org/10.1109/CVPRW.2010.5543262

[9] Dongshun Cui, Guang-Bin Huang, and Tianchi Liu. 2018. ELM based smile detection using Distance Vector. *Pattern Recognition* 79 (2018), 356–369. https://doi.org/10.1016/j.patcog.2018.02.019

[10] D Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, San Diego, CA, USA, 886–893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[11] Charles Darwin. 2013. *The Expression of the Emotions in Man and Animals*. Cambridge University Press, England. https://doi.org/10.1017/CBO9781139833813

[12] Alex Davies and Zoubin Ghahramani. 2014. The Random Forest Kernel and other kernels for big data from random partitions. arXiv:1402.4293 [stat.ML]

[13] Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124–129. https://doi.org/10.1037/h0030377

[14] Hugo. Filho, Oge Marques; Vieira Neto. 1999. *Processamento Digital de Imagens*. Brasport, Brasil. 30–31 pages.

[15] Gabriel Garrido and Prateek Joshi. 2018. *OpenCV 3.X with Python By Example: Make the most of OpenCV and Python to build applications for object recognition and augmented reality* (2nd ed.). Packt Publishing, US.

[16] A. T. Ghorbani, G; Targhi and M. Dehshibi. 2015. HOG and LBP: Towards a robust face recognition system. In *2015 Tenth International Conference on Digital Information Management (ICDIM)*. IEEE, Jeju, South Korea, 138–141. https://doi.org/10.1109/ICDIM.2015.7381860

[17] Ellen Goeleven, Rudi De Raedt, Lemke Leyman, and Bruno Verschuere. 2008. The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion* 22, 6 (2008), 1094–1118. https://doi.org/10.1080/02699930701626582 arXiv:https://doi.org/10.1080/02699930701626582

[18] Rafael C Gonzales and Richard E. Woods. 2008. *Digital Image Processing* (3rd ed.). Pearson, New Jersey, US.

[19] Isabelle Guyon and André Elisseeff. 2003. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* 3 (March 2003), 1157–1182.

[20] T. Hassner, E. Harel, S.and Paz, and R. Enbar. 2015. Effective face frontalization in unconstrained images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, US, 4295–4304. https://doi.org/10.1109/CVPR.2015.7299058

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction* (2 ed.). Springer, USA. http://www-stat.stanford.edu/~tibs/ElemStatLearn/

[22] Jiabei He, Xiaoyu Wen, and Juxiang Zhou. 2023. Advances and Application of Facial Expression and Learning Emotion Recognition in Classroom. In *Proceedings of the 2023 6th International Conference on Image and Graphics Processing* (Chongqing, China) *(ICIGP '23)*. Association for Computing Machinery, New York, NY, USA, 23–30. https://doi.org/10.1145/3582649.3582670

[23] Ursula Hess. 2001. The Communication of Emotion. In *Emotions, Qualia and Consciousness*. Singapore, 397–409. https://doi.org/10.1142/9789812810687_0031

[24] Nurulhuda Ismail and Mas Idayu Md. Sabri. 2009. Review of Existing Algorithms for Face Detection and Recognition. In *Proceedings of the 8th WSEAS International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics* (Puerto De La Cruz, Tenerife, Canary Islands, Spain) *(CIMMACS'09)*. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 30–39.

[25] A. Kumar, K.M. Baalamurugan, and B. Balamurugan. 2022. Real-Time Facial Components Detection Using Haar Classifiers. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, Salem, India, 01–08. https://doi.org/10.1109/ICAAIC53929.2022.9793034

[26] Uttama Lahiri, Esube Bekele, Elizabeth Dohrmann, Zachary Warren, and Nilanjan Sarkar. 2011. Design of a Virtual Reality Based Adaptive Response Technology for Children with Autism Spectrum Disorder. In *Affective Computing and Intelligent Interaction*, Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 165–174.

[27] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. 2010. Presentation and validation of the Radboud Faces Database. *Cognition and Emotion* 24, 8 (2010), 1377–1388. https://doi.org/10.1080/02699930903485076 arXiv:https://doi.org/10.1080/02699930903485076

[28] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. 2010. Presentation and validation of the Radboud Faces Database. *Cognition and Emotion* 24, 8 (2010), 1377–1388. https://doi.org/10.1080/02699930903485076 arXiv:https://doi.org/10.1080/02699930903485076

[29] K. Li, F. Xu, J. Wang, Q. Dai, and Y. Liu. 2012. A data-driven approach for facial expression synthesis in video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, RI, US, 57–64. https://doi.org/10.1109/CVPR.2012.6247658

[30] Shan Li and Weihong Deng. 2018. Deep Facial Expression Recognition: A Survey. *Computing Research Repository (CoRR)* abs/1804.08348 (2018). https://doi.org/10.1109/TAFFC.2020.2981446

[31] Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. 2017. Japanese Female Facial Expression (JAFFE) Database. (7 2017). https://doi.org/10.6084/m9.figshare.5245003.v2

[32] Dhwani Mehta, Mohammad Faridul Haque Siddiqui, and Ahmad Y. Javaid. 2018. Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality. *Sensors* 18, 2 (2018). https://doi.org/10.3390/s18020416

[33] Karnati Mohan, Ayan Seal, Ondrej Krejcar, and Anis Yazidi. 2021. Facial Expression Recognition Using Local Gravitational Force Descriptor-Based Deep Convolution Neural Networks. *IEEE Transactions on Instrumentation and Measurement* 70 (2021), 1–12. https://doi.org/10.1109/TIM.2020.3031835

[34] A Monzo, D; Albiol and M. J. Mossi. 2010. A Comparative Study of Facial Landmark Localization Methods for Face Recognition Using HOG descriptors. In *2010 20th International Conference on Pattern Recognition*. IEEE, Istanbul, Turkey, 1330–1333. https://doi.org/10.1109/ICPR.2010.1145

[35] Leandro Persona, Fernando Meloni, and Alessandra Macedo. 2023. An accurate real-time method to detect the smile facial expression. In *Anais do XXIX Simpósio Brasileiro de Sistemas Multimídia e Web* (Ribeirão Preto/SP). SBC, Porto Alegre, RS, Brasil, 46–55. https://sol.sbc.org.br/index.php/webmedia/article/view/25865

[36] Rosalind W. Picard. 2016. Automating the Recognition of Stress and Emotion: From Lab to Real-World Impact. *IEEE MultiMedia* 23, 3 (July 2016), 3–7. https://doi.org/10.1109/MMUL.2016.38

[37] I.Michael Revina and W.R. Sam Emmanuel. 2018. A Survey on Human Face Expression Recognition Techniques. *Journal of King Saud University - Computer and Information Sciences* (2018). https://doi.org/10.1016/j.jksuci.2018.09.002

[38] Jia S, Wang S, Hu C., Webster PJ, and Li X. 2021. Detection of Genuine and Posed Facial Expressions of Emotion: Databases and Methods. *Front. Psychol. - Sec. Perception Science* 11 (15 January 2021), 12p. https://doi.org/10.3389/fpsyg.2020.580287

[39] F. Z. SALMAN, A. MADANI, and M. KISSI. 2016. Facial Expression Recognition Using Decision Trees. In *2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV)*. IEEE, Beni Mellal, Morocco, 125–130. https://doi.org/10.1109/CGiV.2016.33

[40] Paul F. Smith, Siva Ganesh, and Ping Liu. 2013. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal*

of Neuroscience Methods 220, 1 (2013), 85–91. https://doi.org/10.1016/j.jneumeth.2013.08.024

[41] Rafael Luiz Testa, Cléber Gimenez Corrêa, Ariane Machado-Lima, and Fátima L. S. Nunes. 2019. Synthesis of Facial Expressions in Photographs: Characteristics, Approaches, and Challenges. *ACM Comput. Surv.* 51, 6, Article 124 (jan 2019), 35 pages. https://doi.org/10.1145/3292652

[42] Nim Tottenham, James Tanaka, Andrew Leon, Thomas Mccarry, Marcella Nurse, Todd Hare, David Marcus, Alissa Westerlund, Bj Casey, and Charles Nelson. 2009. The NimStim set of Facial Expressions: Judgments from Untrained Research Participants. *Psychiatry research* 168 (07 2009), 242–9. https://doi.org/10.1016/j.psychres.2008.05.006

[43] Sana Ullah and Wenhong Tian. 2021. A Systematic Literature Review of Recognition of Compound Facial Expression of Emotions. In *Proceedings of the 2020 4th International Conference on Video and Image Processing* (Xi'an, China) *(ICVIP '20)*. Association for Computing Machinery, New York, NY, USA, 116–121. https://doi.org/10.1145/3447450.3447469

[44] Jean Vaillancourt. 2010. Statistical Methods for Data Mining and Knowledge Discovery. In *Proceedings of the 8th International Conference on Formal Concept Analysis* (Agadir, Morocco) *(ICFCA'10)*. Springer-Verlag, Berlin, Heidelberg, 51–60.

[45] Paul Viola and Michael J. Jones. 2001. Robust Real-Time Face Detection. *International Journal of Computer Vision* 57, 2 (2001), 137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb

[46] V. Vonikakis and S. Winkler. 2020. Identity-Invariant Facial Landmark Frontalization For Facial Expression Analysis. In *International Conference on Image Processing (ICIP)*. 2020 IEEE ICIP, Abu Dhabi, United Arab Emirates, 2281–2285. https://doi.org/10.1109/ICIP40778.2020.9190989

[47] P. Winterle. 2014. *Vetores e Geometria Analítica*. MAKRON. https://books.google.com.br/books?id=AKhivgAACAAJ

[48] Yue Wu and Qiang Ji. 2018. Facial Landmark Detection: A Literature Survey. *International Journal of Computer Vision* 2 (2018), 115–142. https://doi.org/10.1007/s11263-018-1097-z

[49] L. Xie, W.; Shen and J. Jiang. 2017. A Novel Transient Wrinkle Detection Algorithm and Its Application for Expression Synthesis. *IEEE Transactions on Multimedia* 19, 2 (Feb 2017), 279–292. https://doi.org/10.1109/TMM.2016.2614429

[50] W. XIE, L. SHEB, M. YANG, and Q. HOU. 2015. Lighting difference based wrinkle mapping for expression synthesis. In *2015 8th International Congress on Image and Signal Processing (CISP)*. IEEE, Shenyang, China, 636–641. https://doi.org/10.1109/CISP.2015.7407956

[51] Xiaoming Zhao and Shiqing Zhang. 2016. A Review on Facial Expression Recognition: Feature Extraction and Classification. *IETE Technical Review* 33, 5 (2016), 505–517. https://doi.org/10.1080/02564602.2015.1117403 arXiv:https://doi.org/10.1080/02564602.2015.1117403