frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 31 January 2024
DOI 10.3389/fcomp.2024.1359471

OPEN ACCESS

EDITED BY
Cesar Collazos,
University of Cauca, Colombia

REVIEWED BY
Tiago Thompsen Primo,
Federal University of Pelotas, Brazil
Jeferson Arango López,
University of Caldas, Colombia
Gustavo Eduardo Constain Moreno,
National Open and Distance University, Colombia

*CORRESPONDENCE
Gabriel M. Ramírez V.
✉ gramirez@udemedellin.edu.co
Fernando Moreira
✉ fmoreira@upt.pt

RECEIVED 21 December 2023
ACCEPTED 19 January 2024
PUBLISHED 31 January 2024

CITATION
Ballesteros JA, Ramírez V. GM, Moreira F, Solano A and Pelaez CA (2024) Facial emotion recognition through artificial intelligence. *Front. Comput. Sci.* 6:1359471. doi: 10.3389/fcomp.2024.1359471

Facial emotion recognition through artificial intelligence

Jesús A. Ballesteros[1], Gabriel M. Ramírez V.[2]*, Fernando Moreira[3]*, Andrés Solano[4] and Carlos A. Pelaez[4]

[1]Maestria en Inteligencia Artificial, Universidad Internacional de La Rioja, Logroño, Spain, [2]Facultad de Ingeniería, Universidad de Medellín, Medellín, Colombia, [3]REMIT, IJP, Universidade Portucalense, Porto and IEETA, Universidade de Aveiro, Aveiro, Portugal, [4]Departamento de Operaciones y Sistemas, Universidad Autónoma de Occidente, Cali, Colombia

This paper introduces a study employing artificial intelligence (AI) to utilize computer vision algorithms for detecting human emotions in video content during user interactions with diverse visual stimuli. The research aims to unveil the creation of software capable of emotion detection by leveraging AI algorithms and image processing pipelines to identify users' facial expressions. The process involves assessing users through images and facilitating the implementation of computer vision algorithms aligned with psychological theories defining emotions and their recognizable features. The study demonstrates the feasibility of emotion recognition through convolutional neural networks (CNN) and software development and training based on facial expressions. The results highlight successful emotion identification; however, precision improvement necessitates further training for contexts with more diverse images and additional algorithms to distinguish closely related emotional patterns. The discussion and conclusions emphasize the potential of A.I. and computer vision algorithms in emotion detection, providing insights into software development, ongoing training, and the evolving landscape of emotion recognition technology. Further training is necessary for contexts with more diverse images, alongside additional algorithms that can effectively distinguish between facial expressions depicting closely related emotional patterns, enhancing certainty and accuracy.

KEYWORDS

facial emotion, recognition, A.I., convolutional neural network, images

1 Introduction

Affective computing is an interdisciplinary field that involves studying and developing systems capable of understanding and interpreting human emotions (Banafa, 2016). One of the primary motivations for research in this field is the simulation of empathy: endowing machines with the ability to detect and interpret users' emotional states and thus generate adaptive behavior based on the recognized information.

Facial expressions are crucial in communication and convey complex mental states during interaction. In non-verbal communication, the face transmits emotions (Darwin and Prodger, 1996). Using machine learning techniques such as face recognition, information obtained from facial expressions can be processed to infer their emotional state.

Affective computing, which recognizes user emotional states, proposes to enrich the form of user-machine interaction. A system with this capability could generate more appropriate responses considering users' emotional states (Banafa, 2016).

The application of affective computing offers a wide range of possibilities. In marketing, analyzing emotions is instrumental in determining the impact of a given advertisement or product on the public. An increasing number of companies are betting on projects related

to affective computing, such as the detection and prevention of stress in workers or the development of video games capable of adapting to players.

This paper presents the development of software capable of detecting a user's emotions through computer vision techniques using A.I. algorithms, considering the theories of emotions and how to evaluate emotions with different algorithms and thus determine people's emotions.

The development of software capable of detecting the emotions of a user through computer vision techniques using A.I. algorithms, specifically neuronal convolutional networks; face recognition is performed using the framework Multitask Cascade Convolutional Networks (MTCNN), considering the theories of emotions and how to evaluate emotions with different algorithms and thus determine the emotions of people.

The paper is structured as follows: Section 1 Introduction, Section 2 Materials and methods, Section 3 Results, and Section 5 Discussions.

2 Materials and methods

This section delves into the foundational elements, encompassing the psychological dimensions of emotions and the various classification theories. This study extends to the technical facets of facial recognition, exploring current techniques employed for object recognition and image classification. Furthermore, it touches upon the intricacies of developing emotion recognition software using A.I. algorithms.

2.1 Background

Emotions are pivotal in mammals, providing essential information for survival and environmental adaptation. Perception of emotions is defined as the ability to take appropriate actions or direct thoughts and identify emotions in oneself or other individuals (Salovey and Mayer, 1990).

It is crucial to differentiate between emotion and feeling. Emotions arise unconsciously and rapidly, requiring no explicit mental processing. In contrast, feelings are consciously elaborated from the emotions experienced by the individual (Darwin and Prodger, 1996).

Emotions manifest primarily as physical responses characterized by specific physiological activation patterns. However, it is also noteworthy that different emotions can share similar physiological responses; fear and anger both increase heart rate. The same emotion can elicit various responses, such as fleeing, fighting, or experiencing paralysis in danger or intense fear (Darwin and Prodger, 1996).

2.1.1 Psychology of emotion

Emotions play a vital role in mammals, providing essential information for survival and environmental adaptation. Emotion perception can be defined as the ability to take actions or direct thoughts appropriately and identify emotions in oneself or other individuals (Darwin and Prodger, 1996).

Emotions arise briefly unconsciously without requiring explicit mental processing. Primarily, emotions are physical responses, which are represented by a characteristic's physiological activation pattern. Sometimes, two or more emotions may share specific physiological responses (Banafa, 2016).

In 1994, the psychologist and anthropologist Paul Ekman proposed six emotions not determined by sociocultural factors (Ekman et al., 1969; Ekman, 1994). This set of emotions consists of joy, anger, fear, disgust, and surprise. These emotions were called fundamental or universal because they are closely related to survival behaviors and evolutionary patterns in the human species (Salovey and Mayer, 1990). Subsequently, this initial set of universal emotions would be expanded with contempt (García, 2013).

The set of universal emotions has given rise to different models that try to explain the great variety of emotions that exist by combining two or more basic emotions. Attempts to classify emotions have generated circumplex models based on fuzzy categories and factorial models. These models propose using the opposite extremes of the emotional categories to respond to humans' emotional states (Russell, 1980).

Plutchik's circumplex model consists of eight basic emotions: joy, trust, surprise, aversion, sadness, anticipation, and anger. Plutchik argues that the emotional states described in his model are similar. For Plutchik (2001) this similarity facilitated combining one or more of these emotional states to obtain a more complex emotion.

Represent the relationship between emotion categories. The vertical dimensions of the cone represent the intensity of the emotion. In contrast, the cone sections represent the degree of similarity in the intensity of the emotions. In the displayed model, the blanks represent the secondary emotions resulting from combining two primary emotions.

Russell (1997) argues that affective states result from two independent neurophysiological systems: one is in charge of establishing the valence of the emotion, and the other is related to the state of activation of the individual. Therefore, emotions can be interpreted as the linear combination of these two dimensions (Russell, 1997). This theory describes the organization of emotional states based on two orthogonal axes. The horizontal axis represents the valence dimension, while the vertical axis represents the activation dimension.

2.1.2 Facial recognition

Face recognition originated in the 1960's when a research team led by W. Bledsoe conducted experiments to determine whether a computer could recognize human faces. Bledsoe's team sought to establish relationships between the minutiae of the human face so that the computer could find a set of matches that would allow recognition of those faces (Plutchik, 2001).

Bledsoe's experiments could have been more successful, but they were vital in laying the groundwork for using biometric information in face recognition (Bledsoe, 1966). For many years, the techniques used in face recognition did not develop significantly until 2001, when Paul Viola and Jones published a method for object detection that offered previously unheard-of hit rates.

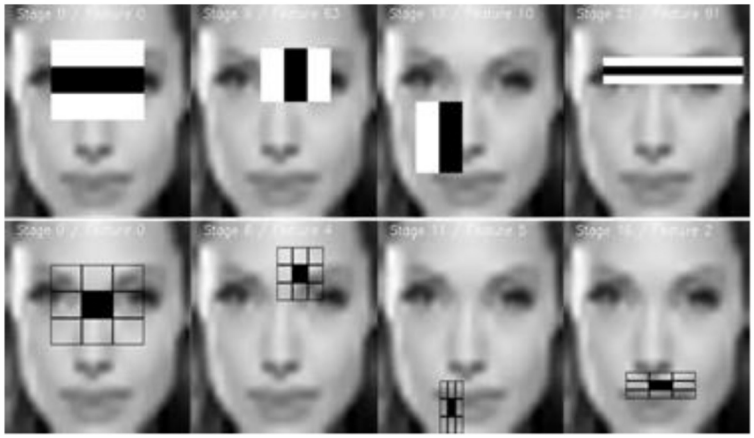


FIGURE 1
Haar characteristics.

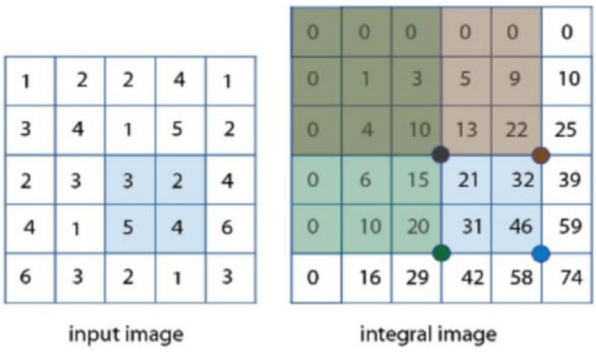


FIGURE 2
Integral image calculation (Schapire, 2013).

2.1.3 Viola-Jones method

The Viola-Jones method is one of the most widely used techniques for face recognition tasks today. This algorithm comprises two phases: a first phase of training an AdaBoost classifier (Wang, 2014) and a second stage of detection using the classifier with unknown images (Simonyan and Zisserman, 2014). This method uses the image's Haar features instead of the pixel level; see Figure 1. The Haar features of an image represent the difference in intensity in adjacent areas, which allows the detection of intensity changes in different orientations. They are calculated as the difference in the sum of the pixels of two or more contiguous rectangular areas based on the light intensity of the pixels.

The Viola-Jones method proposes using an intermediate, integral image to reduce the degree of complexity when dealing with images. As seen in Figure 2, the intermediate image represents the original image, where each point corresponds to the sum of the pixels located to its left and above it (Bledsoe, 1966). This type of transformation on the image reduces the algorithm's complexity from O(N^2) to O(1).

Once the features are obtained by this fast "Integral Image" technique, they are categorized. The Viola-Jones method uses a variation of the AdaBoost classification algorithm (Schapire, 2013), which handles selecting a reduced set of features. The AdaBoost algorithm (Schapire, 2013), developed by Freud and Schapire, is a type of classifier that learns by a chain combination of smaller classifiers. The original idea of AdaBoost was to improve the performance of other classification algorithms.

AdaBoost separately trains each classifier that composes it. These classifiers input the information that its predecessor could not correctly classify. The algorithm selects each iteration's feature set with the lowest error rate. When a classifier can recognize a region in an image, that region is passed to the next classifier, which is how the feature set is selected. In AdaBoost, the classifiers focus on a single feature; therefore, a new feature and a new classifier will be chosen at each iteration.

2.1.4 CNN—convolutional neural networks

Nowadays, the high availability of labeled data and the continuous improvement of GPUs have played a crucial role in developing new face recognition algorithms based on Deep Convolutional Neural Networks (DCNs). The widespread use of these neural networks was motivated by the 2012 ImageNet championship (Krizhevsky et al., 2017).

CNNs are formed by several layers of neurons, with the convolutional layer being the most important one. As shown in Figure 3, the input of this layer is a vector of pixel values. Its operation is based on applying a series of filters that move through the image to obtain the layer outputs (Simonyan and Zisserman, 2014).

In addition to the filter size used in the convolutional layers, two fundamental parameters modify its behavior: stride and padding. The stride controls how the filter moves through the image. When the stride size of the stride is increased, the convolutional layer is fixed in more distant areas of the image, which implies a dimensionality reduction. Dimensionality reduction is the prevalent use of the Zero-Padding technique, which fills the edges

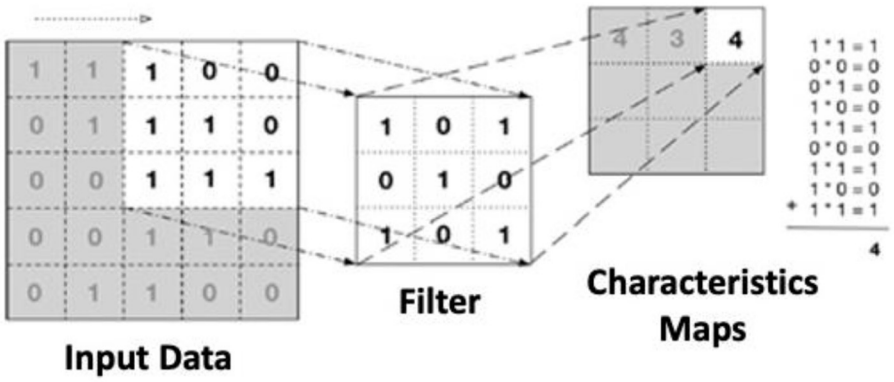


FIGURE 3
Convolution algorithm operation by applying a 3x3 filter (Sotil, 2022).

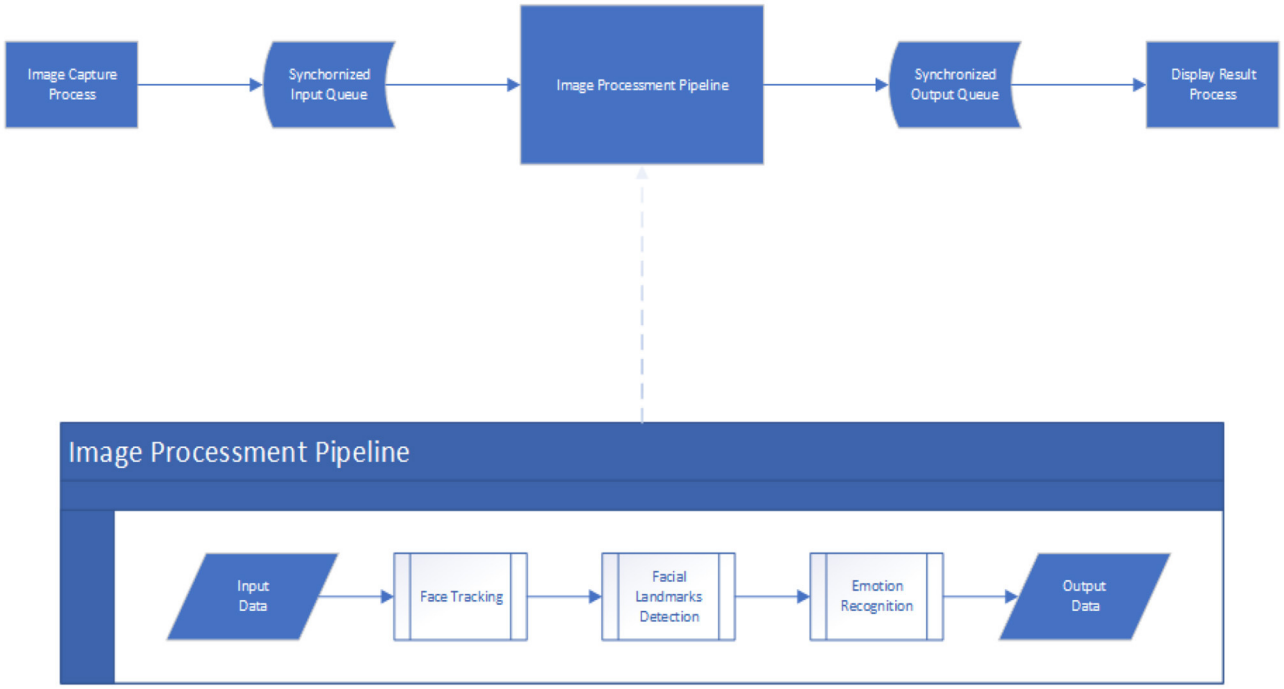


FIGURE 4
Software architecture.

of the output obtained by applying filters with zeros (Mathworks, 2023).

Other layers are intermixed with the convolutional layers, the intermediate layers. The purpose of these layers is to eliminate the non-linearities while maintaining the dimensions to improve the robustness and avoid overtraining of the neural network. These layers employ RELU activation units, which are computationally more efficient than the traditional Sigmoid or Tangh (Thomas et al., 2005).

As one goes deeper into the neural network and traverses more convolutional layers, one obtains activation maps that represent more complex features of more significant regions in the image: the first layers recognize more basic image structures in a small region. In comparison, deeper layers obtain higher-level representations from the elements recognized in the first layers (Simonyan and Zisserman, 2014).

The architecture of the AlexNet convolutional neural network (Krizhevsky et al., 2017) won the 2012 ImageNet championship. AlexNet was a turning point in the field of computer vision.

AlexNet was a turning point in the field of computer vision. Following its success, new architectures based on more complex CNNs appear year after year, capable of beating the record set in the previous year. In addition, this architecture was the first to use ReLU activation units, and today, this

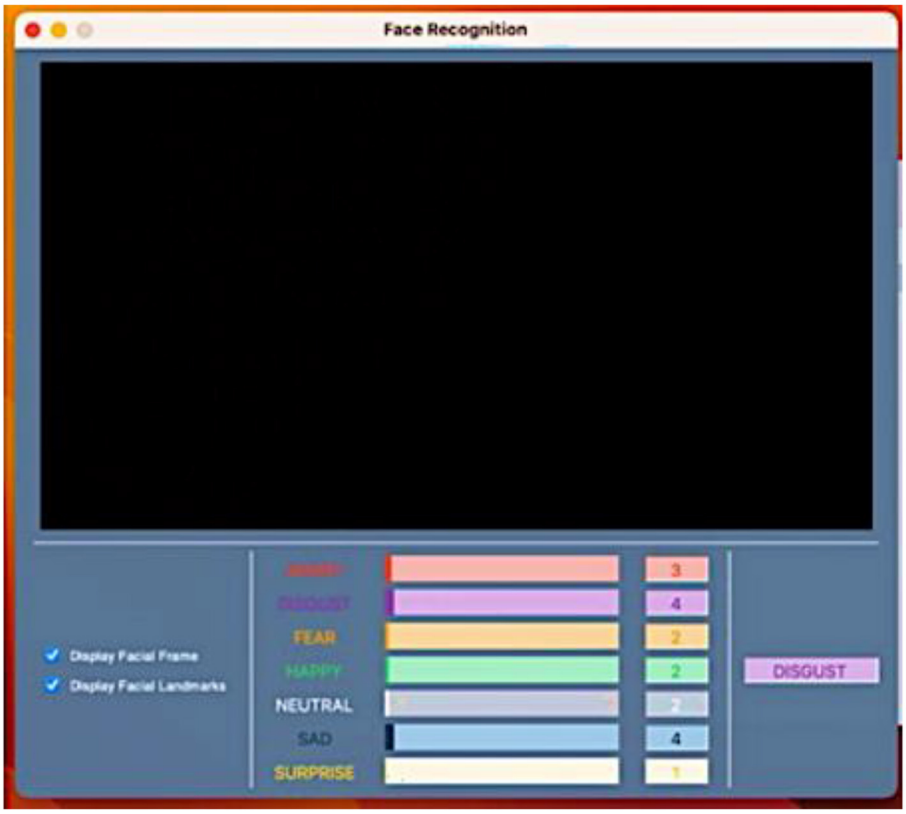


FIGURE 5
Graphical interface.

activation unit is the most widely used in the world of deep learning.

2.1.5 Intermediate layer

In this type of neural network architecture, other layers are intermingled with the convolutional and intermediate layers. These layers aim to produce dimension-preserving non-linearities to improve the neural network's robustness and avoid overtraining (Simonyan and Zisserman, 2014).

These layers are based on the ReLU activation units. These units are computationally more efficient than the traditional sigmoid or tangh. ReLU activation layers transform negative input values to 0 using $f(x) = \max(0, x)$.

As you go deeper into the neural network and go through more convolutional layers, you get activation maps that represent more complex features of more significant regions in the image: the first layers recognize more basic image structures in a small region, whereas deeper layers will get higher level representations from the elements recognized in the first convolutional layers for much larger image regions (Simonyan and Zisserman, 2014).

2.1.6 Exit layer

A fully connected layer is the last layer of a convolutional network for classification problems. This layer takes as input the output of the last convolutional layer and returns as output a vector of dimension N, where N is the number of classes for image classification. Typically, this is done using a Softmax layer (Simonyan and Zisserman, 2014).

2.1.7 Stride and padding

In addition to filter size, two fundamental parameters modify the behavior of a convolutional layer: stride and padding.

The stride controls how the filter moves across the image. Increasing the stride size causes the features obtained from the convolutional layer to be fixed in more distant areas of the image, which implies a dimensionality reduction (Simonyan and Zisserman, 2014).

The formula can calculate the resulting size of the output: [(N-F)/S]+1, where N is the size of the input NxN, F is the size of the filter being applied, and S is the size of the

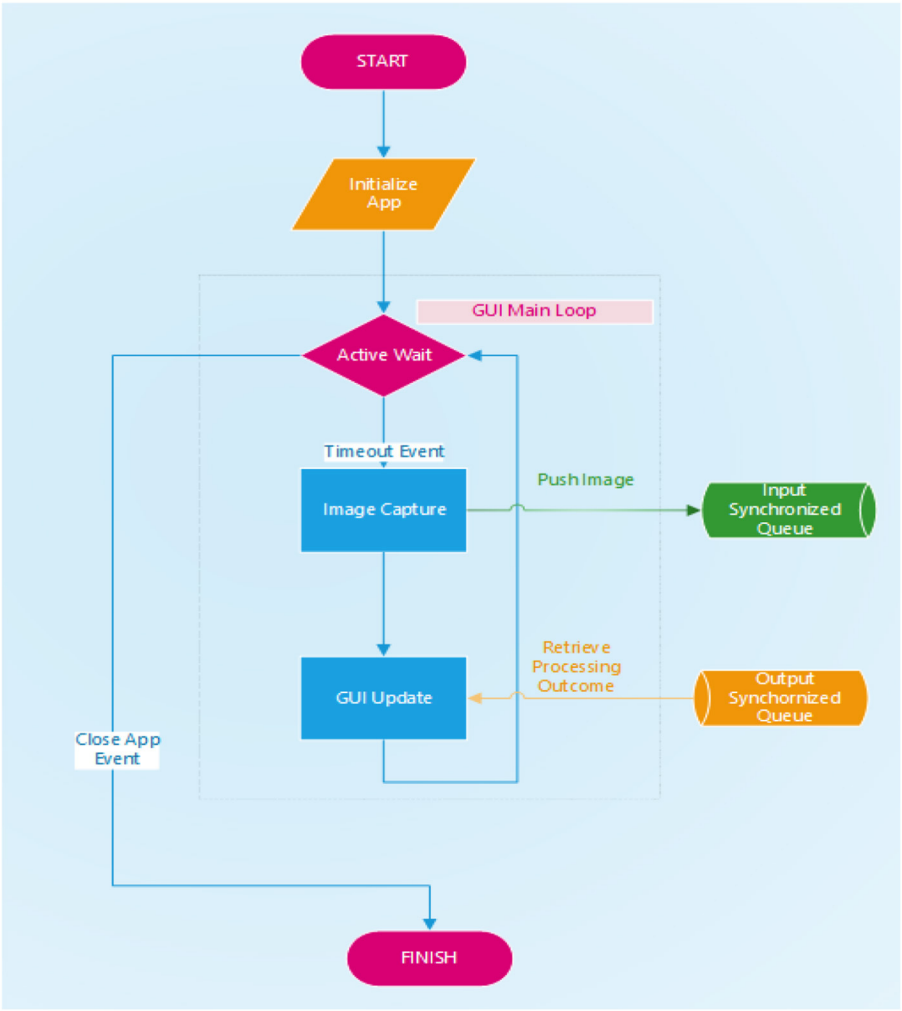


FIGURE 6
Application flowchart.

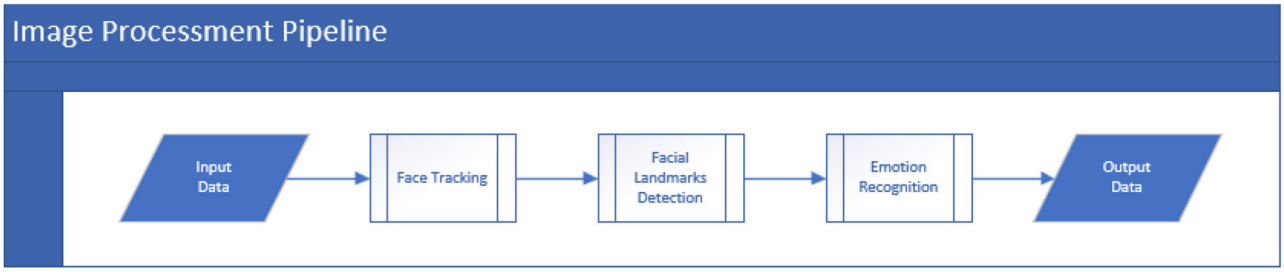


FIGURE 7
Image processing pipeline.

stride with which the filter is shifted (Simonyan and Zisserman, 2014).

As observed, applying filters always implies a reduction in the dimensionality of the resulting images. If several convolutional layers are applied in succession, the image's dimensionality may be excessively reduced.

The zero-padding technique is widespread for preventing dimensionality reduction, thus preserving the exact dimensions of both the input and output images. This technique involves padding the edges of the output obtained with zeros, thereby "recovering" the pixels lost when applying filters and maintaining the dimensionality of the input (Simonyan and Zisserman, 2014).

2.1.8 Max pooling layer

Max pooling layers are common in most CNN architectures. Typically positioned after ReLU layers, their primary function is to diminish the size of the obtained representations, thereby reducing the number of parameters necessary in the network. This reduction in computational complexity makes the representations more efficient and significantly contributes to mitigating overtraining in the neural network—a technique commonly called downsampling (Simonyan and Zisserman, 2014).

The operational principle of max pooling layers relies on filters, typically of size two and a stride of 2. In this process, the output of each filter in the pooling layer corresponds to the maximum value within the region it covers.

The reduction of the representation occurs in the spatial plane of the image but does not affect the number of features obtained. The precise location of a feature becomes less critical than its relative position concerning other features extracted from the image.

FIGURE 8
Facial recognition along with the 64 key points.

2.2 Development

2.2.1 Problem statement

Facial expressions play a crucial role in communication, serving as a primary means of conveying intricate mental states during human interaction. Central to non-verbal communication, the human face is the primary vehicle for expressing emotions (Darwin and Prodger, 1996). Six basic emotions, considered universal and innate, are associated with distinct facial expressions: disgust, fear, joy, anger, surprise, and sadness (Ekman, 1994). Studies by Ekman and Friesen (Ekman, 1999) indicate that while subjects can adequately recognize these emotions, perfect identification still needs to be discovered.

Expressing play an emotion involves intricate muscular configurations and creating recognizable patterns. Key facial areas, such as the forehead-eyebrows, eyes-eyelids, and the lower part of the face (around the mouth), primarily contribute to manifesting these emotions (Plutchik and Conte, 1997).

Emotions hold a fundamental role in human decision-making. Integrating new technologies, particularly those rooted in artificial intelligence, is becoming increasingly prevalent. These technologies can recommend movies and series based on user preferences, tailor advertising according to search and purchase history, and learn from daily routines and user interactions with applications. Affective computing, which recognizes users' emotional states, enhances user-machine interaction by generating responses that align with users' emotional states (Banafa, 2016).

2.2.1.1 Research question

How do we develop emotion recognition software using computer vision techniques through the recognition of facial expressions?

2.2.2 Objectives

- General objective:

To develop an emotion recognition software using computer vision techniques that allows the detection of users' emotions through the recognition of facial expressions.

- Specific objectives:

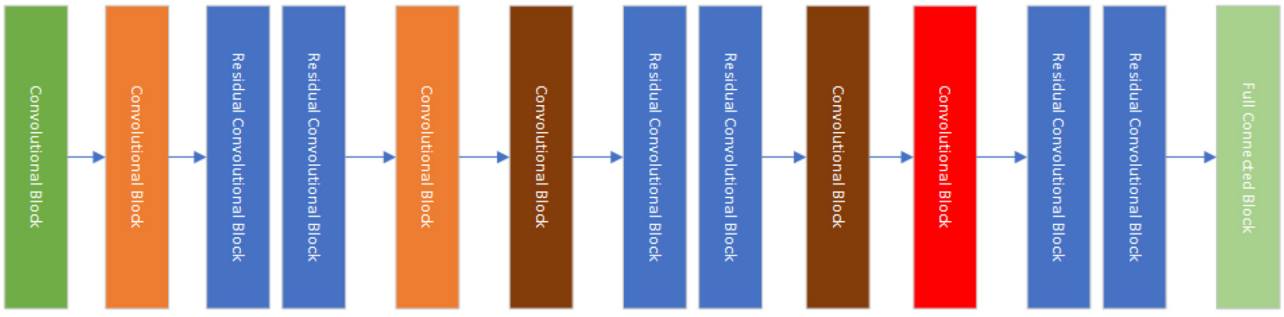


FIGURE 9
ResNet architecture for emotion classification.

1. Design the software architecture that defines the end-to-end of an application that allows the detection of users' emotions through a capture device.
2. Implement a graphical interface that supports the application and allows real-time visualization of the images provided by the capture device. In addition, this graphical interface must be able to display the results obtained by processing the input images.
3. Implement a software component to perform facial recognition of users by employing an input image from the capture device. This component must be sufficiently computationally efficient to ensure the correct operation of the application in real-time.
4. Implement a software component to classify emotions from a user's facial expressions. This component must be sufficiently computationally efficient to ensure the correct functioning of the application in real-time.
5. Validate the obtained application. Evaluating the system's performance will provide the necessary information to determine the feasibility of emotion detection by recognizing users' facial expressions.

2.2.3 Methodology and the research approach

This work's objectives necessitate using a descriptive applied research methodology, employing a mixed approach that combines qualitative and quantitative data. Video data collected from user interactions will undergo classification and conversion using a neural network and a classification model to identify emotions based on established emotional concepts, resulting in qualitative and quantitative data. This mixed research approach allows for a comprehensive analysis (Hernández Sampieri et al., 2003).

The SCRUM methodology will guide the software development process, an AGILE approach that enables flexible adaptation to project needs for optimal results. SCRUM operates on short, fixed-duration iterations known as Sprints, culminating in an intermediate product with part of the required functionality at the end of each iteration (Albaladejo et al., 2021).

The functional requirements have been categorized into modules, organized into three main blocks based on their functionalities within the application: GUI, facial recognition, and emotion recognition. The GUI module handles the visualization of input data and results obtained from emotion recognition. The facial recognition module, the initial part of the pipeline, processes input data by detecting the user's face. Finally, the Emotion Recognition module, positioned as the pipeline's last segment, processes input data to identify the expressed emotion on the user's face.

2.2.4 Software architecture

Developing a complex software system requires the design of architectures that allow developers to extend and modify the system's capabilities. For this, it is essential to design the architectures in software blocks, which allow dividing the software solution into a set of independent modules that communicate with each other; this facilitates the development and scalability of the system. The main advantage of modular architectures is the reduction in development time because they allow the extension or modification of the functionality of previously developed modules.

The development of the proposed emotion recognition application, which utilizes computer vision techniques, adheres to the architecture outlined in Figure 4. This architecture comprises the image capture process, processing pipeline, and result visualization process. The solution's design follows a producer-consumer model using synchronized queues.

The image capture process is responsible for initializing and collecting data from the capture device, adapting it, and making it available by the image processing pipeline. This pipeline consists of three steps: face tracking, facial landmarks detection, and emotion recognition. The results obtained due to the input image processing are dumped into the output synchronized queue. The result visualization process collects and paints the result in the graphical interface.

The class diagram traces the system's structure according to its composing classes. As shown in the diagram, the components that integrate the processing pipeline implement the abstract class AbstractImgProcessor to homogenize the public methods required during pipeline execution. It is worth mentioning that, unlike languages such as Java, there is no interface mechanism to define the specifications and behaviors other classes will implement in Python.

2.2.5 Graphical interface

The graphical interface is the main module of the application. In addition to the representation of the elements to be displayed by the users, this module is responsible for the start of components and the orchestration of the application.

During the start phase, all the components involved in the application are instantiated and configured, the capture device is initialized, and the asynchronous tasks executed during the application life cycle are planned.

Figure 9 shows the logic for developing the solution using the execution flowchart. The orchestration of the application is based on the event-driven design pattern by generating a timeout event every 20 ms. This event generates a constant flow of images from the capture device. The input images are deposited in an asynchronous line and consumed by the processing pipeline.

The data generated by the pipeline are deposited by the pipeline in another asynchronous line to be consumed by the main application thread. From the data deposited in this queue, the interface's graphical components will be updated, for example, the visualization of the input images obtained by the capture device or the update of the emotion recognition results.

- Viewer: This graphic component displays the images obtained from the capture device. This component also represents the information generated by the facial recognition of the input image.
- Boxes: The graphical interface has been configured with two checkboxes or active ones by default. These checkboxes allow activation or deactivation of the display of the results obtained from facial recognition of the input images.

- Display Facial Frame: activates or deactivates the frame that delimits the area of the input image detected as a face.
- Display Facial Landmarks: Enables or disables the display of the 64 critical landmarks detected on a face. This display helps determine whether face detection is being performed correctly.

- Status bars: Each bar corresponds to a specific emotion. These bars represent the probability rate assigned to an emotion based on the results of processing the input image.
- Classification status: The emotion recognized with the highest probability rate from an input image.

The graphical interface module comprises two differentiated phases: component start and application orchestration as shown in Figure 5. During the start phase, all the components involved in the application logic are instantiated and configured. In addition, the trained models are loaded into memory, the capture device is initialized, and the asynchronous tasks executed during the application life cycle are planned.

As shown in the flowchart in Figure 6, the orchestration occurs place in the main loop of the GUI. The main loop captures and handles events triggered by the components that integrate the graphical interface. It is characterized by having an event-driven wait mechanism, which implies that the main program thread is suspended until a component of the interface is reached.

Orchestration of the application occurs place in the main loop of the GUI. This main loop is used to orchestrate the application components and the event handling of the GUI.

2.2.6 Image processing pipeline

The image processing pipeline defines the steps necessary for emotion recognition from the input data. This pipeline is executed by an asynchronous task that consumes the data deposited in the input line. The results are deposited in the output line for the updating graphic components. Figure 7 presents the steps this processing pipeline executes to recognize user emotions. Emotion recognition is performed by executing the following three steps defined in this processing pipeline: Face Tracking, Facial Landmarks Detection, and Emotion Recognition.

Step 1. Face Tracking: The Face Tracking step obtains the coordinates or bounding box that delimits the faces detected in the input image, trimming the detected area and making the result available for the next component of the Image Processing Pipeline.

The face recognition task uses the Multitask Cascade Convolutional Networks framework (MTCNN; Zhang et al., 2016). In addition to providing high accuracy in face recognition, this tool offers excellent performance for real-time applications even through a CPU. The MTCNN architecture, see Figure 9, comprises three cascaded convolutional networks: P-Net, Q-Net, and O-Net.

Step 2. Facial Landmarks Detection: During the Facial Landmarks Detection step, the detection of the 64 critical points on a face is based on the results obtained from the facial recognition of the input image. The objective of this phase is to determine whether the face is being captured correctly or not. This information will be of great use when positioning the capture device.

TABLE 1 Training hyperparameters.

Training parameters	Value
Epoch number	50
Batch size	32
Learning rate	1e-1
Optimization	Stochastic Gradient Descent
Loss function	Cross Entry Loss

The detection of the 64 critical points of the face is performed using CNNs. The network architecture chosen was Xception Net. This architecture is characterized by being built as a linear stack of depth-separable convolution layers with residual connections, making it easy to define and modify as shown in Figure 8.

Step 3. Emotion Recognition: Emotional recognition will be performed from the facial expression extracted during Step 1: Face Tracking. The classification of facial expressions occurs place using CNNs. The architecture designed to implement this classifier, as shown in Figure 9, corresponds to an adaptation of the ResNet architecture (Centeno, 2021; Zhao et al., 2022).

The model used to classify facial expressions was trained using the public dataset FER-2013. This dataset comprises face images with a size of 48x48 pixels in grayscale. FER-2013 contains 28,000 labeled images for the training dataset and 3,500 for the validation dataset. Each image in the dataset corresponds to one of the following seven categories: 0 = angry, 1 = disgust, 2 = fear, 3 = happy, 4 = sad, 5 = surprise, 6 = neutral (Kaggle, 2019). Table 1 shows the hyperparameters used during model training. Metrics were obtained using the set of images to validate the trained model (Sambare, 2023).

3 Results

The developed emotion classification and detection software was initially tested with images to obtain results for sad, fearful, angry, surprised, disgusted, and happy emotions. Then, the classification and emotion detection tests were performed with images and multimedia content of users in real-time to have a control group to compare the results obtained and define the accuracy of the detection of emotions.

Once the software tool for emotion recognition has been implemented, this document chapter will be devoted to the experiments to evaluate the application's performance. The tests were carried out in two different scenarios: images and user tests.

3.1 Tests with images

This test will comprise 19 representative images for each emotional category, except the "Neutral" category, which will not be evaluated in this test. The author selects the images according to his criteria, classifying them into one category or another.

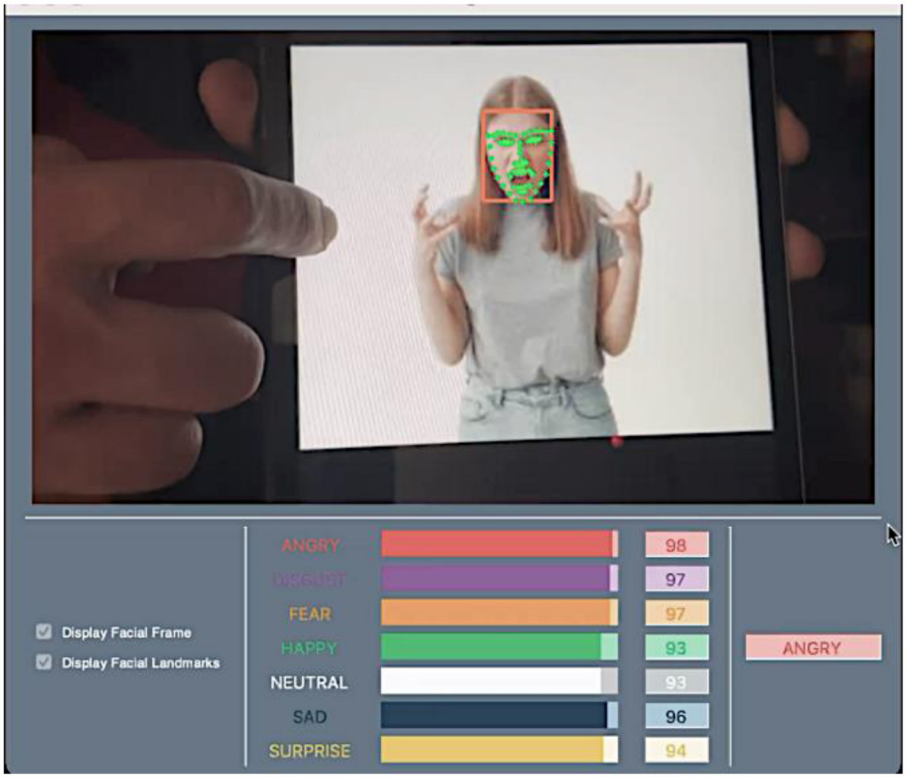


FIGURE 10
Test with an image.

The application will process these images, and the same image will be kept in front of the capture device until the application returns a single result. If the application cannot return a single result, the image's emotion recognition will be considered a failure.

Figure 10 corresponds to a fragment of the tests performed on the images selected by the author for the "Angry" category. It can be seen that the facial recognition of the person in the image is performed correctly, showing the characteristic points of the face. From the information displayed on the interface, it is possible to verify that the following can be verified.

This test compares the author's emotion classification ability with the system's ability. It is essential to mention that this test has a high degree of subjectivity due to the selection of images. The selected images are completely decontextualized, so there will be images that could fit into several categories.

Table 2 presents the results obtained from the image tests as a confusion matrix. The "Neutral/No Clear" row has extended this confusion matrix. The purpose of this row is to collect the results of those images for which the system has not shown a clear result, oscillating between two or more categories.

The confusion matrix allows for visualization of the performance of the classification algorithm with static images during the tests. Each column of this matrix represents the number of predictions for each class. At the same time, the rows of this matrix correspond to the number of instances in the actual class. With the information collected in the confusion matrix, conclusions can be drawn about the types of hits or misses generated by the model during image testing.

The accuracy metric refers to the number of correct optimistic predictions; as seen in Table 2, all classes obtained an accuracy value above 82%. This value obtained can lead to confusion because it can be interpreted as an indicator that the model has a high predictive capacity.

The accuracy metric, the proportion of images that have been correctly classified, is obtained. As can be seen in Table 2, three categories have had an "accuracy" above 60%: "angry," "surprised," and "happy." At the same time, all the other categories have obtained a value below 50%. The value obtained from this metric can be helpful when estimating the model's predictive capacity.

The results showed that some of these images previously classified by the author could perfectly fit into several emotional categories; this reinforces the idea expressed in the multidimensionality of emotions: two emotions can share or have elements in common in their physiological activation patterns.

Comparing the results obtained with the metrics of the model evaluation represented by Table 3, it can be observed that the "accuracy" metric of the test results follows the same progression, except for the emotions "angry," "surprise," and "happy," which improve slightly concerning the model validation.

The explanation for why the rest of the emotional categories has such a low hit rate is shown in Figure 3: Russell's Circumplex Model (Russell, 1980). As can be seen, the emotional categories detected by the model are all located in quadrants I and II.

TABLE 2 Test results with images.

Category	Sad	Fear	Angry	Surprise	Disgust	Happy
Sad	9	1	1	-	6	-
Fear	-	8	-	1	-	-
Angry	1	1	12	-	2	-
Surprise	-	2	-	14	-	-
Disgust	1	-	-	-	8	-
Happy	-	1	-	2	2	18
Neutral/No Clear	8	1	6	2	1	-
Precision	0.47	0.44	0.63	0.73	0.42	1
Accuracy	0.83	0.93	0.90	0.93	0.89	0.95

TABLE 3 Classification model evaluation metrics.

Category	Precision	Recall	F1-score	Support
Angry	0.49	0.43	0.46	958
Disgust	0.49	0.55	0.52	111
Fear	0.42	0.25	0.32	1,024
Happy	0.87	0.70	0.77	1,774
Neutral	0.39	0.78	0.52	1,233
Sad	0.54	0.18	0.27	1,247
Surprise	0.53	0.84	0.65	831

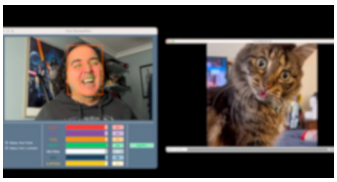
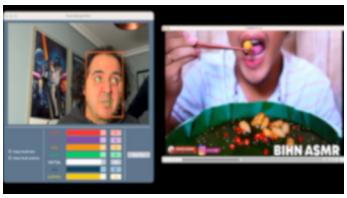
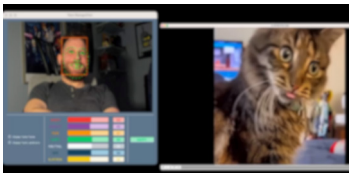
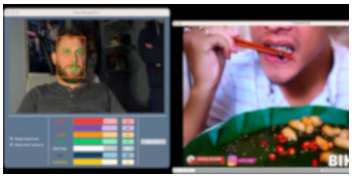
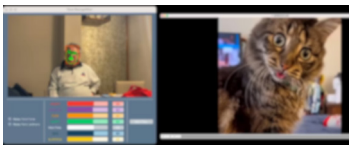
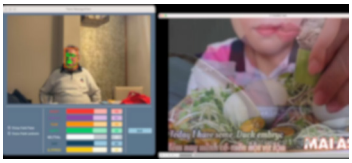
TABLE 4 Test with video.

I.D.	Content	Goal
Vídeo#1	Comic scenes starring domestic animals	Detection of positive emotions such as "happy" or "surprise."
Vídeo#2	Unpleasant scenes of people consuming unusual foods	Detection of negative emotions such as "disgust" or "fear."

The emotions "happy" and "surprise" are in quadrant I. These two emotions are differentiated by their levels of pleasure and activation. "Happy" has the maximum value of pleasure and the minimum activation, while "surprise" has the opposite. Therefore, they could be complementary emotions in the same quadrant.

In quadrant II, the rest of the emotions detected by the application are located. As can be seen, the emotions "disgust," "anger," and "fear" are located consecutively in the middle zone of quadrant II. The unpleasantness activation difference between these emotions is minimal. According to Russell's circumplex model, "angry" corresponds to the middle value between "disgust" and "fear," which would explain why it has the highest hit rate in this quadrant.

TABLE 5 Test results with users.

User	Id	Result	Evidence
User#1	1	OK	
	2	OK	
User#2	1	OK	
	2	OK	
User#3	1	N/A	
	2	N/A	

3.2 Tests with video

This set of tests was carried out through the participation of three volunteers individually. Each test is composed of two audiovisual contents that will be played. During the playback of each video, the developed tool monitors the facial expressions of the volunteers.

The goal of these tests is to determine the ability of the developed tool to detect the emotions generated by a specific

message. Each video used in these tests was designed to provoke emotion in the viewer Table 4.

The results of the tests with volunteer #1 in Table 5 could have been more satisfactory. During the tests with video #1, it was possible to recognize positive emotions caused by the content played. In the case of video #2, the emotion "neutral" was detected most of the time during the playback. Some negative emotions were detected in very short periods but were not significant enough to consider the test as OK.

The results of the tests with volunteer #2 in Table 5 could have been more satisfactory. During the tests with video #1, it was possible to recognize positive emotions caused by the content played. In the case of video #2, it was impossible to recognize any negative emotions expressed by volunteer #2. During the playback of video #2, some positive expressions could be recognized; curiously, some scenes made volunteer #2 smile. Still, the tests performed with video #2 can be considered as K.O. because no negative emotions were detected.

The results of the tests with volunteer #3, presented in Table 5, needed to be more satisfactory. During the tests performed with video #1, the individual showed no expressivity that the system could recognize. In the case of video #2, no expressivity of any kind could be detected either. The negative emotions recorded correspond to the characteristic facial expression of volunteer #3. The results obtained by testing with volunteer #3 show no assessment of the application's performance.

The results obtained from the tests with video #1 were satisfactory for most volunteers. Positive emotions could be detected throughout the video playback, with "Happy" being the predominant emotion. Volunteers #1 and #2, in the same age range, showed positive emotions during the playback of video #1. In contrast, volunteer #3, who was older, did not show any expressiveness when faced with the same content.

In contrast, the results obtained from the tests performed with video #2 were negative. The tool has not been able to recognize any of the volunteer's emotions directly related to the target of the content. The following points could explain these results.

a) The trained model has much lower hit rates for negative emotions, making recognizing these emotions more challenging. In addition, volunteers seem more "reluctant" to exhibit negative emotions. As seen in the evidence recordings, there is a notable difference between the degree of expressivity during the visualization of the contents used in the tests. This fact and the low hit rate of negative emotions could explain the results obtained.
b) Test design. The content of video #2 did not have the necessary impact on the volunteers, who maintained a relatively neutral expression throughout the test. This behavior may be explained by the fact that users are aware that their facial expressions are being monitored during the tests, implying a certain degree of conditioning when expressing negative emotions.
c) Another possible argument to explain the results obtained during the playback of video#2 may be based on the law of apparent reality. Nico Fridja stated that an individual's perception of something real derives from an emotional response. Conversely, if the individual does not directly observe the event, it may lead to thinking that it is not real and, therefore, does not engage emotionally (Frijda, 2017). This statement by Nico Fridja raises the question: would volunteers react if they had experienced the test content *in situ*? Possibly not.

4 Discussions

Drawing from the results obtained in tests with volunteers, as outlined in the paper, and considering the inherent complexities in accurately classifying emotions, constructing a system based on computer vision techniques capable of precisely categorizing the entire spectrum of user emotions is challenging (Monteith et al., 2022). This paper emphasizes the non-trivial nature of correctly identifying emotions, asserting that emotions must be comprehensively understood in all dimensions. The physiological activation patterns that characterize emotions may share elements, intensifying the intricacies of emotion identification.

Leveraging artificial intelligence (AI) for emotion recognition profoundly benefits individuals and humanity (Tanabe et al., 2023). The ability of AI systems to accurately perceive and interpret human emotions opens avenues for enhanced human-computer interaction, personalized user experiences, security, education, military, health, entertainment, and improved mental health support (Chollet, 2017; Lu, 2022). As highlighted by recent studies, emotion-aware systems can contribute to developing empathetic A.I. applications, ranging from virtual assistants that respond to users' emotional states to mental health tools capable of providing timely and contextually sensitive interventions. Furthermore, integrating emotion recognition in A.I. aligns with the growing emphasis on human-centric technology, fostering a more nuanced and adaptive relationship between individuals and the digital realm (Lee and Park, 2022).

Observations during the development and evaluation of the tool led to the realization that an image should convey more information for accurate emotion determination. While certain facial expressions, such as joy or surprise, are easily recognizable without prior context, this is only true in specific cases. Referencing Russell's circumplex model reveals the complexity inherent in this study. Quadrant I encompass emotions like "Happy" and "Surprise," whereas Quadrant II houses over 57% of the emotional categories recognized by the developed application: "Disgust," "Anger," "Fear," and proximately, "Sad." Notably, these four categories, representing 25% of Russell's model, exhibit significant differences, complicating accurate identification.

However, this complexity does not deny the feasibility of developing an emotion recognition system based on computer vision techniques (Ghotbi, 2023). The strategy should incorporate additional characteristics to augment the information beyond that derived from the analysis of individual facial expressions, raise the possibility of retraining the model with another dataset with more images, increase the dimensions and resolution of the images, and expand the tests with more users and with more videos, which in

theory should improve the metrics of accuracy and certainty in the classification of emotions (Zhao et al., 2022).

Considering the results, the overarching goal of creating a tool capable of recognizing user emotions using computer vision techniques has been realized. The developed tool successfully fulfills all the objectives outlined in this study. The application showcases real-time results from the recognition of user emotions. The components responsible for face detection, facial landmark recognition, and facial expression classification were designed using computer vision techniques based on convolutional neural networks. Three distinct convolutional network architectures, chosen for their computational efficiency, were employed in this study.

Future work should focus on obtaining greater accuracy in classifying a user's facial expressions. Undoubtedly, providing context to the detected facial expressions is a prerequisite for the correct classification of emotions.

Affective computing is a field in which much research still needs to be done, and it has enormous potential for society. As has happened to virtual assistants, systems based on affective computing will soon be fully integrated into society, offering users a completely personalized experience.

The work presented in this paper is a small sample of the effort in affective computing to achieve the correct emotional identification of users interacting with a system. It highlights the complexity of the problem involved.

The tool developed throughout this study has several areas for improvement that result in a need for more accuracy in recognizing the proposed set of emotions. This fact is separate from the importance of the results and conclusions obtained but demonstrates the difficulty involved in something as seemingly simple for humans as emotion recognition.

Future work should focus on obtaining greater precision when classifying a user's facial expressions. Providing context to the detected facial expressions is necessary for correctly classifying emotions; the possibility of increasing the size of the images with superior resolution and dimension and testing with more users in video to verify if better results can be obtained in the recognition of emotions is proposed. Using LSTM networks, a new emotion classifier can be trained from a time series of facial expressions. A time series would provide some context to the classification because it would not be based on a single facial expression but on a sequence of images defining an emotion's expression.

The results of the tests performed with the expressions are confusing or difficult to classify, raise the possibility of retraining the model with another dataset with more images, increasing the dimensions and resolution of the images, and expanding the tests with more users and with more videos, which in theory should improve the metrics of accuracy and certainty in the classification of emotions. However, other authors raise the concept of microexpressions in short periods, which would be another study to develop.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

JB: Conceptualization, Investigation, Methodology, Writing—original draft, Writing—review & editing. GR: Conceptualization, Investigation, Methodology, Writing—original draft, Writing—review & editing. FM: Conceptualization, Funding acquisition, Investigation, Methodology, Writing—original draft, Writing—review & editing. AS: Conceptualization, Investigation, Methodology, Writing—original draft, Writing—review & editing. CP: Conceptualization, Investigation, Methodology, Writing—original draft, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. FCT supported this work—Fundação para a Ciência e a Tecnologia, I.P. (Project UIDB/05105/2020).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Albaladejo, X., Díaz, J. R., Quesada, A. X., and Iglesias, J. (2021). *Proyectos agiles.org*. Available online at: https://proyectosagiles.org/pm-partners (accessed July 12, 2023).

Banafa, A. (2016). *Qué es la computación afectiva?* OpenMind BBVA. Available online at: https://www.bbvaopenmind.com/tecnologia/mundo-digital/que-es-la-computacion-afectiva/ (accessed September 14, 2023).

Bledsoe, W. W. (1966). *Man-Machine Facial Recognition: Report on a Large-Scale Experiment. Technical Report PRI 22*. Palo Alto, CA: Panoramic Research.

Centeno, I. D. P. (2021). *MTCNN Face Detection Implementation for TensorFlow, as a Pip Package*. Available online at: https://github.com/ipazc/mtcnn (accessed September 14, 2023).

Chollet, F. (2017). "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1251–1258.

Darwin, C., and Prodger, P. (1996). *The Expression of the Emotions in Man and Animals*. Oxford: Oxford University Press.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychol. Bull.* 115, 268–287. doi: 10.1037/0033-2909.115.2.268

Ekman, P. (1999). Basic emotions. *Handb. Cogn. Emot.* 3, 45–60. doi: 10.1002/0470013494.ch3

Ekman, P., Sorenson, E., and Friesen, W. (1969). Pan-cultural elements in facial displays of emotion. *Science* 164, 86–88. doi: 10.1126/science.164.3875.86

Frijda, N. H. (2017). *The Laws of Emotion*. London: Psychology Press.

García, A. R. (2013). La educación emocional, el autoconcepto, la autoestima y su importancia en la infancia. *Estudios y propuestas socioeducativas*. 44, 241–257.

Ghotbi, N. (2023). The ethics of emotional artificial intelligence: a mixed method analysis. *Asian Bioethics Rev.* 15, 417–430. doi: 10.1007/s41649-022-00237-y

Hernández Sampieri, R., Fernández, C., and Baptista, L. C. (2003). *Metodología de la Investigación*. Chile: McGraw Hill.

Kaggle (2019). *FER—2013*. Available online at: https://www.kaggle.com/ (accessed October 5, 2023).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Lee, Y. S., and Park, W. H. (2022). Diagnosis of depressive disorder model on facial expression based on fast R-CNN. *Diagnostics* 12:317. doi: 10.3390/diagnostics12020317

Lu, X. (2022). Deep learning based emotion recognition and visualization of figural representation. *Front. Psychol.* 12:818833. doi: 10.3389/fpsyg.2021.818833

Mathworks (2023). *Integral Image*. Available online at: https://www.mathworks.com/help/images/integral-image.html (accessed October 16, 2023).

Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C., and Bauer, M. (2022). Commercial use of emotion artificial intelligence (AI): implications for psychiatry. *Curr. Psychiatr. Rep.* 24, 203–211. doi: 10.1007/s11920-022-01330-7

Plutchik, R. (2001). The nature of emotions. *Am. Scientist* 89, 334–350. doi: 10.1511/2001.28.334

Plutchik, R. E., and Conte, H. R. (1997). *Circumplex Models of Personality and Emotions*. Washington, DC: American Psychological Association.

Russell, J. A. (1980). A circumplex model of effect. *J. Personal. Soc. Psychol.* 39:1161. doi: 10.1037/h0077714

Russell, J. A. (1997). "Reading emotions from and into faces: resurrecting a dimensional-contextual perspective," in *The Psychology of Facial Expression*, eds J. A. Russell and J. M. Fernández-Dols (Cambridge University Press; Editions de la Maison des Sciences de l'Homme), 295–320.

Salovey, P., and Mayer, J. (1990). Emotional Intelligence. *Imag. Cogn. Personal.* 9, 185–211. doi: 10.2190/DUGG-P24E-52WK-6CDG

Sambare, M. (2023). *Kraggle. FER-013. Learn Facial Expresions From a Image*. Available online at: https://www.kaggle.com/datasets/msambare/fer2013 (accessed October 16, 2023).

Schapire, R. E. (2013). "Explaining adaboost," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (Berlin; Heidelberg: Springer), 37–52. doi: 10.1007/978-3-642-41136-6_5

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. doi: 10.48550/arXiv.1409.1556

Sotil, D. A. (2022). *RPubs*. Available online at: https://rpubs.com/ (accessed October 14, 2023).

Tanabe, H., Shiraishi, T., Sato, H., Nihei, M., Inoue, T., and Kuwabara, C. (2023). A concept for emotion recognition systems for children with profound intellectual and multiple disabilities based on artificial intelligence using physiological and motion signals. *Disabil. Rehabil. Assist. Technol.* 1–8. doi: 10.1080/17483107.2023.2170478

Thomas, J. R., Nelson, J. K., and Silverman, J. (2005). *Research Methods in Physical Activity, 5th Edn*. Champaign, IL: Human Kinetics.

Wang, Y. Q. (2014). An analysis of the Viola-Jones face detection algorithm. *Image Process. Line* 4, 128–148. doi: 10.5201/ipol.2014.104

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sign. Process. Lett.* 23, 1499–1503. doi: 10.1109/LSP.2016.2603342

Zhao, J., Wu, M., Zhou, L., Wang, X., and Jia, J. (2022). Cognitive psychology-based artificial intelligence review. *Front. Neurosci.* 16:1024316. doi: 10.3389/fnins.2022.1024316