USO DE INTELIGÊNCIA ARTIFICIAL PARA O RECONHECIMENTO DE EMOÇÕES EM IMAGENS

Aline Gadelha Oliveira
Graduanda em Engenharia de Controle e Automação
campus São Paulo, IFSP

Cauê Coelho Rangel
Graduando em Engenharia em Eletrônica
campus São Paulo, IFSP

Daniel de Andrade Moura
Graduando em Engenharia em Eletrônica
campus São Paulo, IFSP

Matheus de Luna Guerharthi
Graduando em Engenharia de Controle e Automação
campus São Paulo, IFSP

Ricardo Pires
Doutor em Sistemas Automáticos e Microeletrônicos
Docente no campus São Paulo do IFSP

RESUMO

O reconhecimento de elementos a partir da expressão facial é de fundamental importância na comunicação humana, pois ela faz parte da linguagem não verbal. Uma máquina que seja capaz de reconhecer emoções pode ser usada em pesquisas, na avaliação de pacientes em hospitais etc. Tendo isto em vista, o presente trabalho apresenta os resultados obtidos com o uso de um sistema de inteligência artificial desenvolvido utilizando a linguagem Python, em duas etapas: detecção de faces e reconhecimento de emoções a partir de imagens. Ao final do desenvolvimento e dos testes, o sistema alcançou acurácia de 92% na identificação de emoções contrastantes, por meio do uso de rede neural convolucional.

Palavras-chave: Inteligência Artificial; expressões faciais; reconhecimento de emoções; detecção de faces; processamento de imagens.

ABSTRACT

The recognition of elements from facial expression is of fundamental importance in human communication, as it is part of non-verbal language. A machine capable of recognizing emotions can be used in research, assessing patients in hospitals, etc. With this in mind, the present work presents the results obtained using an artificial intelligence system developed using the Python language in two stages: face detection and emotion recognition from images. At the end of the development and testing, the system achieved an accuracy of 92% in the identification of contrasting emotions, by the use of a convolution neural network.

Keywords: Artificial Intelligence; facial expressions; emotion recognition; face detection; image processing.

Introdução

Uma expressão facial é um movimento ou posicionamento de músculos faciais (Fridlund, 1997). O movimento facial humano resulta de cerca de 20 músculos que fazem a face transmitir emoções, dor e sentimentos. Estas expressões são fundamentais na comunicação humana (Silva et al., 2000).

O rosto é uma importante região de comunicação dos estados emocionais do corpo humano, sendo que alguns estudos alegam que muitas expressões faciais são inatas, e independem de aprendizado ou cultura (Silva e Silva, 1994).

O rosto humano apresenta várias informações como sinais faciais (Ekman apud Silva e Silva, 1994). Esses sinais podem ser divididos em quatro categorias:

a) Sinais estáticos – aqueles que pouco se alteram durante a vida do indivíduo, como estrutura óssea e cor da pele.

b) Sinais lentos – mudanças que ocorrem com o passar do tempo, mas que não se alteram instantaneamente, tais como rugas, cabelos e barba.

c) Sinais rápidos – aqueles que ocorrem em questão de segundos, podendo ser mais ou menos sutis, por exemplo, dilatação da pupila e contrações musculares.

d) Sinais artificiais – intervenções nos vetores de sinais lentos e estáticos, como óculos de grau, maquiagem e cirurgias plásticas.

Silva e Silva, apud Ekman e Friesen (1994), sustentam que se pode inferir a emoção de outrem somente a partir dos sinais rápidos, como os movimentos faciais e o tônus muscular.

Nos anos 1970, Paul Ekman (Slimani et al, 2018) demonstrou que certas expressões faciais são consideradas inatas e universais, sendo observadas em qualquer cultura; são elas a alegria, tristeza, medo, raiva, desprazer e surpresa. Essas expressões são chamadas de “emoções primárias” ou de “emoções” básicas.

As emoções fazem parte da linguagem não verbal e a sua interpretação é determinante para a compreensão da mensagem, afetando diretamente o resultado da comunicação (Nogueira e Faria, 2014). A partir do reconhecimento de emoções, pode-se evitar ou incentivar uma série de ações, daí a sua importância.

Apesar da relevância disso, até pouco tempo atrás, as máquinas não eram capazes de reconhecer emoções. Porém, com o advento da Inteligência Artificial, isto mudou e novas áreas de pesquisas surgiram. Uma delas é a área de Computação Afetiva ou Computação Emocional, na qual se estuda a interação entre a tecnologia e as emoções humanas, a fim de

desenvolver sistemas e mecanismos capazes de reconhecer, interpretar e responder às emoções humanas (Happy e Routray, 2015).

O reconhecimento das expressões faciais é importante no entendimento do comportamento humano, em sistemas de segurança, na interação entre humanos e computadores, no monitoramento de motoristas durante o trabalho e no ensino a distância (Slimani et al, 2018). Ao crescer numa determinada cultura, o ser humano aprende a decodificar símbolos e signos (Vigostki, 2008), inclusive aqueles presentes na face dos integrantes da comunidade em que vive. Porém, é importante ressaltar que uma mesma expressão pode ter interpretações diferentes dependendo de seu intérprete. A complexidade desta atividade é tamanha que as máquinas não conseguem interpretar estas expressões sem que sejam treinadas por meio das técnicas de Inteligência Artificial, sendo esta aprendizagem fonte de pesquisa em diversos campos.

Maqableh, Alzyoud e Zraqou (2023) apresentaram uma abordagem para a criação de um ambiente para aprendizado digital, no qual expressões faciais e o ritmo do batimento cardíaco dos estudantes foram usados para medir a efetividade do aprendizado e o nível de engajamento deles. Os resultados superaram aqueles obtidos por sistemas que não usavam essa abordagem.

Pesquisas como as de Lencioni e Zanella (2020) e de Nunes e Pacheco (2020) buscam reconhecer a dor a partir das expressões faciais de animais. Dado que os seres não humanos não se comunicam pela fala, criar mecanismos capazes de identificar a dor neles pode auxiliar no tratamento médico e aumentar as chances de cura e recuperação.

Também há pesquisas voltadas para a identificação de dor pelas expressões faciais em seres humanos. Uma delas é a pesquisa de Coutrin e Thomaz (2023), que investiga o uso de redes neurais convolucionais para identificar sinais de dor em recém-nascidos. A escolha deste público se dá pela dificuldade de comunicação entre o recém-nascido e os adultos. Embora possa parecer recente, a criação e investigação de software para detecção de dor em neonatos em pesquisas nacionais já tem mais de 10 anos, sendo possível encontrar pesquisas como a de Heiderich (2013).

Devido à importância e crescente demanda de aplicações com tal finalidade, o presente projeto visa a investigação de ferramentas para reconhecimento de emoções utilizando técnicas de Inteligência Artificial.

Estudo Bibliográfico

O reconhecimento da expressão facial por meio de Inteligência Artificial (IA) tem sido objeto de estudo de diversas pesquisas nos últimos anos. É um campo muito desafiador, que visa a estreitar a interação homem-máquina. O reconhecimento de emoções em seus semelhantes é algo inerente ao ser humano, e treinar uma máquina para reconhecê-las abre uma vasta gama de possibilidades nas relações humanas com os sistemas computadorizados.

Na literatura, muitos autores já abordaram o tema dos reconhecimento de emoções em expressões faciais, seja por métodos clássicos ou por uso de Redes Neurais. Os métodos clássicos consistem no processamento de imagens e reconhecimento de imagens, um processo de projeto de recurso quase artesanal, onde os dados das características são inseridos manualmente. Por outro lado, nos métodos baseados em Redes Neurais Convolucionais, em inglês, *Convolutional Neural Network* (CNNs), as características são aprendidas por meio da extração de características mais genéricas. As CNNs, por suas propriedades de extraírem características relevantes a partir das características genéricas, são usadas como um amplo "resolvedor de problemas", sendo muito aplicadas na área de processamento de imagem e reconhecimento de padrões (Canal et al, 2021).

Para encontrar um rosto numa imagem, o posicionamento dos componentes da face é primordial para uma melhor detecção. As marcações das posições, como os olhos alinhados na horizontal, contribuem para uma maior acurácia na detecção de um rosto numa imagem (Happy e Routray, 2015).

Happy e Routray (2015) também defendem que o mapeamento das regiões da face pode determinar quais áreas estão ativas, por meio de contrações musculares, como resposta a cada uma das chamadas emoções primárias. A imagem da face é dividida em uma grade determinada a partir da posição dos olhos e, a partir dessa posição principal (já que a posição dos olhos não se altera com a expressão facial), são mapeados em ordem o nariz, sobrancelhas e lábios, e posteriormente estabelecidas as marcações faciais. O posicionamento das marcações dos cantos de lábios e sobrancelhas é uma das principais características que ajudam a distinguir uma expressão da outra.

As regiões ativas são processadas para extrair as características únicas de cada expressão de emoção, e são comparadas uma contra a outra, em todas as combinações possíveis. Foi proposta no mesmo trabalho uma técnica de detecção das marcações faciais de forma automatizada e livre da fase de aprendizado, porém, esse método demanda um maior tempo de processamento. Todavia, mostrou um bom desempenho mesmo em imagens com baixa resolução. Experimentos com os bancos de imagens JAFFE (Japanese Female Facial Expressions) e Cohn-Kanade (CK) trouxeram resultados de 89,64% e 85,06% de acurácia, respectivamente.

Fundamentação Teórica

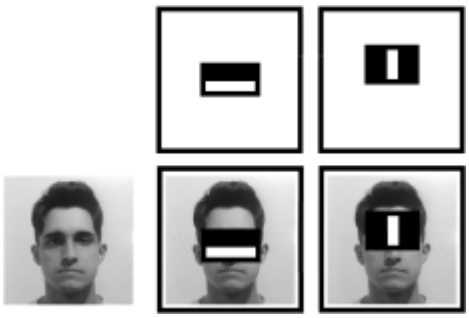
A fim de fornecer modelos analíticos para reconhecimento de expressões humanas a partir de imagens faciais, engenheiros, matemáticos e cientistas da computação estão explorando maneiras distintas de reproduzir abordagens capazes de implementar efetivamente algoritmos eficientes. Existem fortes correlações entre processamento de imagens, visão computacional, reconhecimento de padrões, inteligência artificial e campos de ciência que exploram este tópico. Abordagens interessantes podem ser verificadas na literatura. Em geral, especificamente devido às recentes conquistas em poder de processamento computacional e novas arquiteturas para computação de alto desempenho, aplicações para as áreas citadas, antes restritas por limitações computacionais, tiveram sua solução alcançada. (Canal et al, 2021).

Considerando o contexto de visão computacional, o reconhecimento de emoções comporta técnicas de processamento de imagem, reconhecimento de padrões e inteligência artificial, com aplicações que realizam análise emocional em pesquisas psicológicas e aprimoram as interações entre humanos e máquinas. Este processo envolve a identificação e interpretação dos padrões faciais para caracterizar o que define as expressões e quais são os fatores comuns. É tido que, para desenvolver aplicações cujo objetivo é o processamento de dados volumosos em geral, a linguagem Python, com suas vastas opções bibliotecas como OpenCV, NumPy e Pandas, torna-se de conveniente uso.

Antes de se buscar reconhecer uma face em uma imagem é necessário detectá-la, ou seja, perceber sua existência e delimitar sua localização. Dentre as possíveis técnicas e bibliotecas disponíveis para isso, o presente projeto destaca o uso de Haar Cascade, que se trata de um método de detecção de objetos baseado na identificação e padronização de características visuais (Huamán, 2023), implementado no OpenCV, que é uma biblioteca de processamento de imagens amplamente utilizada em aplicações diversas (Bradski, 2000). A técnica é baseada em padrões retangulares, denominados características Haar, utilizados para o cálculo de diferenças de intensidade em regiões específicas de uma imagem, podendo ser retângulos adjacentes ou sobrepostos, onde o valor de uma característica é a diferença entre a soma das intensidades dos pontos da imagem (*pixels*) em áreas brancas e a soma das intensidades dos *pixels* em áreas pretas, conforme a Figura 1. Nela, o padrão retangular com uma faixa preta sobre uma faixa branca favorece a detecção da área dos olhos, já que esta contém uma região larga e escura sob uma região clara. O padrão ao lado, na Figura 1,

favorece a detecção do nariz. Muitos desses padrões são combinados para, em conjunto, decidirem se a região em análise contém ou não uma face.

Figura 1 – Funcionamento do HaarCascade.



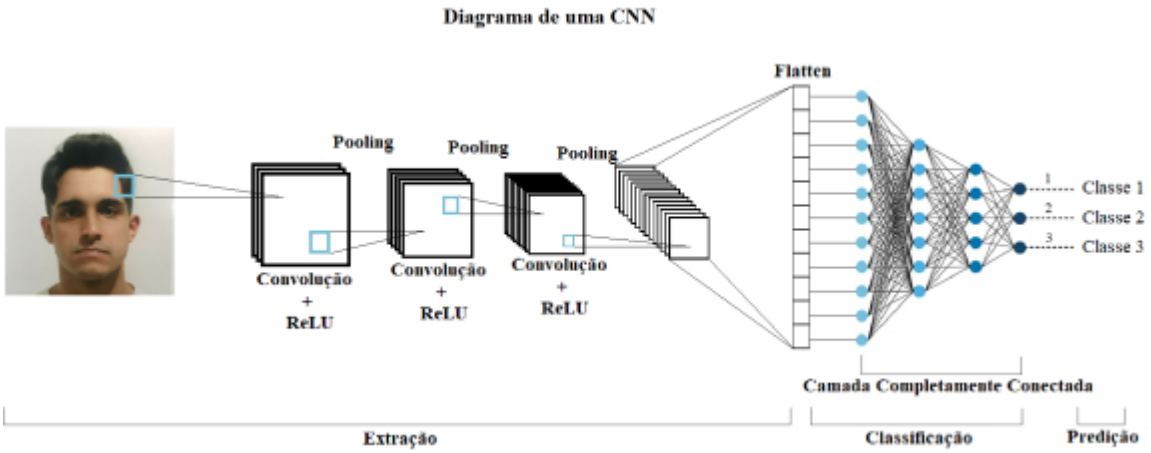
Fonte: autores.

O treino de um HaarCascade é realizado com grandes conjuntos de imagens, divididos entre positivas (contendo o objeto que se queira identificar) e negativas (não contendo o objeto). As características Haar são aplicadas em diversas escalas e posições nas imagens e são identificadas as relevâncias para a distinção entre exemplos positivos e negativos de imagens. Dessa forma, são descartadas as regiões que não contém o objeto a ser identificado. Para reconhecimento facial, um Haar Cascade pré-treinado é aplicado para identificar regiões em uma imagem que possa conter faces. Além do Haar Cascade, é possível utilizar HOG (*Histogram of Oriented Gradients*, ou histograma de gradientes orientados), que se trata de uma técnica de extração de características no contexto de visão computacional e processamento de imagens para detecção de objetos cuja ideia é capturar informações sobre a aparência e a forma de um objeto em uma imagem, levando em consideração gradientes de intensidade (Dalal e Triggs, 2005).

Em geral, após as etapas de detecção facial, é realizado o processo de reconhecimento de expressões, por meio do uso de Redes Neurais Convolucionais, que costumam ser aplicadas como solucionadores de problemas genéricos, onde algum sinal de entrada é decomposto em um conjunto de características invariantes, fornecendo mecanismos robustos para extrair características relevantes (como textura, cantos e pontos-chave). Após uma etapa de treinamento, o classificador estará pronto para interpretar a imagem de uma forma muito eficaz. Consequentemente, o uso de CNN está ascendendo ultimamente em muitas áreas, principalmente aquelas que envolvem processamento de imagens e reconhecimento de padrões (Canal et al, 2021).

A arquitetura de uma CNN é descrita conforme a Figura 2:

Figura 2 – Arquitetura de uma CNN.



Fonte: autores.

O funcionamento de uma CNN ocorre com o objetivo de extrair fatores relevantes sobre o objeto de análise, de forma que um filtro é aplicado sobre as regiões da imagem a fim de reconhecer padrões, de modo que se possa gerar mapas de características. A composição de uma CNN também conta com camadas de *pooling* (agrupamento), que reduzem as dimensões dos mapas de características (Shahriar, 2023). Após os agrupamentos, é inserida uma camada de achatamento (ou *flatten*), que conclui a etapa de preparação dos dados de entrada, alimentado a camada totalmente conectada, que é responsável por fazer a classificação dos dados. Objetivando um melhor desempenho para padrões complexos, entre as camadas é adicionada uma função de ativação não linear (ReLU), pois isso permite a identificação de relações não lineares entre os dados de entrada e de saída. Durante o treinamento, os pesos dos filtros nas camadas convolutivas e camadas totalmente conectadas são ajustados por meio de retropropagação. Por fim, é obtida uma função de ativação para a distribuição de probabilidades, de modo a atribuir uma classe ao objeto analisado. Para fazer a alimentação de uma CNN, é conveniente realizar partições no banco de dados. Vale ressaltar que o objetivo de uma CNN é aprender padrões. Dessa forma, só é possível cumprir essa tarefa de forma satisfatória caso haja validação de que o modelo de fato é capaz de realizar classificações corretamente. Por essa razão, o banco é dividido entre conjuntos, sendo esses, de treinamento e de testes ou validação. Em geral, o conjunto de treinamento recebe a maior parte dos dados, com o objetivo de “ensinar o sistema”, de modo que ocorram as calibrações dos pesos e filtros. O conjunto de testes é composto por uma fatia menor de dados, com o objetivo de avaliar o desempenho do modelo, a fim de comprovar se o sistema “aprendeu”, valendo ressaltar que os dados de testes não são usados no treino. A divisão dos

dados é crucial para que o modelo não se ajuste em excesso para os dados de treinamento e fique incapaz de generalizar para dados novos. O processo de treino em uma CNN é repetido para cada conjunto de dados alimentado à entrada, de modo que o sistema seja percorrido completamente. Cada apresentação completa ao sistema do conjunto completo de dados de treino é chamada "época". Visto que o treino de uma CNN tem por objetivo o ajuste dos pesos dos filtros, é necessário que os conjuntos sejam analisados pelos algoritmos em várias épocas, de modo a otimizar e aprimorar o desempenho do modelo. Vale ressaltar que pode ser conveniente estabelecer critérios de parada para as épocas, para evitar processamento em excesso.

Após a etapa de classificação, por vezes é conveniente que se tenha uma matriz de confusão, que é uma tabela que mostra o desempenho de um modelo de classificação em termos de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos e é de muita importância para evidenciar bem o desempenho de predições e fornecer uma melhor visualização dos dados.

Existem etapas a serem consideradas quando se pretende realizar a construção de um sistema de identificação de expressões faciais com aplicações de IA e visão computacional, das quais destacam-se (Canal et al, 2021):

a) O processo é iniciado com a captura de imagens diversas contendo faces. Pode ser realizado por parte dos construtores do sistema ou por meio do uso de banco de dados já existentes.

b) As imagens sofrem pré-processamento, com o objetivo de melhorar a qualidade e facilitar a detecção de características relevantes. Geralmente, é feita a conversão para escala de cinza, e redimensionamento.

c) É utilizado o Haar Cascade para detectar faces nas imagens.

d) Utilizando técnicas de aprendizado de máquina, é feito o treinamento de um modelo, como uma rede neural convolucional (CNN), com um conjunto de dados rotulados contendo expressões faciais associadas a emoções específicas.

e) O modelo treinado então é utilizado para realizar a classificação das expressões faciais analisadas, determinando a emoção associada à face detectada.

f) Os resultados podem então ser apresentados de forma visual ou utilizados para tomada de decisões, dependendo do contexto da aplicação.

Metodologia e Resultados

Considerando as etapas citadas (Canal et al, 2021) para a construção de um sistema de identificação de expressões faciais, foi dado início ao processo de desenvolvimento de código em linguagem Python, no ambiente do Google Colab, cuja escolha se deu devido à facilidade de compartilhamento de código e controle de versionamento. Conectou-se a aplicação ao banco de dados que, para o presente projeto, foi o banco de imagens Fer2013 (Disponível em https://www.kaggle.com/datasets/msambare/fer2013), que contém mais de 35 mil imagens classificadas em 7 categorias que representam tipos de expressões faciais, sendo essas: raiva (0), nojo (1), medo (2), felicidade (3), tristeza (4), surpresa (5) e neutra (6). A partição de treino consiste em aproximadamente 29 mil imagens, enquanto a partição de teste consiste de aproximadamente 3 mil imagens, assim como a partição de validação. Trata-se de dados fornecidos publicamente, portanto, de livre uso acadêmico, possibilitando estudos diversificados e livre *download*.

A preparação dos dados para a posterior alimentação de uma rede neural frequentemente demanda técnicas diversas de ajuste matemático para adequar e otimizar o processamento. Para tal demanda, o pacote NumPy fornece estruturas para manipulação de *arrays* multidimensionais, sendo essencial para computação científica (Harris et al, 2020). No contexto de processamento de imagens, assume papel conveniente para representar e manipular matrizes com *pixels* de imagens, facilitando a extração de características essenciais traduzidas de forma eficiente para a linguagem da máquina. Para complementar, Pandas é uma biblioteca para análise e manipulação de dados tabulares, conveniente à organização e análise de dados. Após o pré-processamento das imagens e preparação dos dados, é necessário que seja possível a construção de um modelo a ser treinado. Para tal, o Tensorflow é uma alternativa de amplo uso, por se tratar de uma biblioteca popular para a construção e treinamento de modelos de aprendizado de máquina (Abadi et al, 2015). No contexto de reconhecimento facial, é utilizado para criar modelos de redes neurais, como redes neurais convolucionais (CNN).

Para a etapa de classificação das predições, são necessárias iterações, e para tal, a ferramenta itertools é utilizada para criar e manipular iteradores, que são estruturas que permitem iteração sequencial sobre uma coleção de elementos. Oferece funções eficientes para criar e combinar iteráveis, permitindo realizar operações avançadas de iteração de maneira eficaz. (Rossum e Drake Jr, 1995).

Para ilustração de resultados, é conveniente o uso da ferramenta Matplotlib, que é uma biblioteca para gerar gráficos e visualizar dados em alta qualidade. Há uma grande variedade de funções para a criação de gráficos estáticos, interativos e de animações, além de

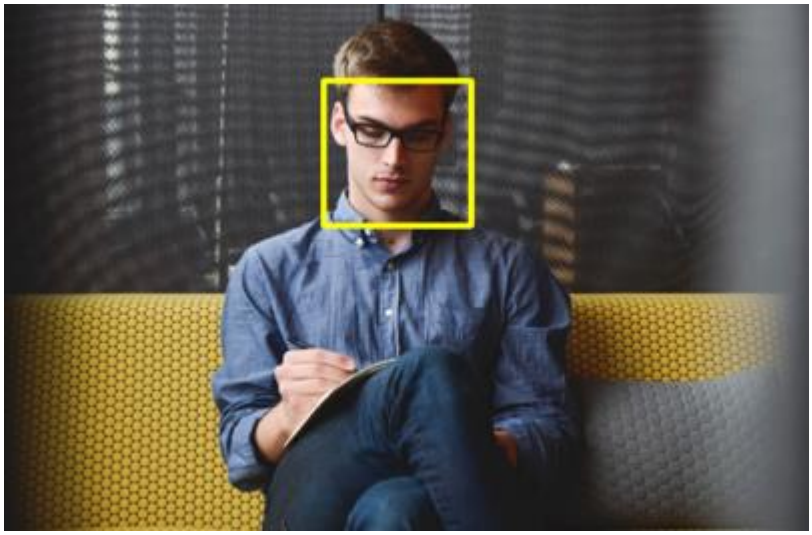
ser de fácil utilização, por ter uma sintaxe semelhante ao MATLAB (Hunter, 2007). A função "confusion_matrix" oferecida pela biblioteca "sklearn.metrics" é utilizada para avaliar o desempenho de algoritmos de classificação, comparando as predições do modelo com os valores reais. (Radoux e Bogaert, 2020).

a) Detecção de faces

A fim de comparar a eficácia de diferentes técnicas, foram desenvolvidos dois programas utilizando técnicas distintas para detecção de faces nas imagens. O primeiro utilizou o Haar Cascade como ferramenta e o segundo fez uso de HOG. Após a conexão com o banco de imagens, foi feita a extração dos arquivos com o objetivo de realizar detecções faciais. Para tal, foi feito o pré-processamento, que consistiu em carregar as imagens e modificar os padrões de cores, inicialmente em RGB, para escalas de cinza e foi feito o redimensionamento das imagens. Dessa forma, o tamanho das imagens foi reduzido, tornando o processamento mais facilitado. Para seguir, usou-se o Haar Cascade, utilizando o comando *cv2.CascadeClassifier* da biblioteca OpenCV, em conjunto com um arquivo em formato XML contendo um modelo previamente treinado para a detecção de faces, que para o caso de apenas uma face na imagem, apresentou resultados satisfatórios, conforme a Figura 3.

Foi observado, por meio de testes realizados com o uso de outras imagens com maior número de faces, que havia falsos positivos nos resultados, sendo essas detecções em regiões da imagem onde não existem faces. Por esta razão, foi necessário realizar o ajuste manualmente de parâmetros do Haar Cascade (scaleFactor e minNeighbors), que após diversos testes, apresentou resultados satisfatórios, mesmo quando consideradas imagens com maior complexidade (contendo um número grande de regiões com faces a serem detectadas), conforme a Figura 4.

Figura 3 – Detecção facial com Haar Cascade em imagem com uma face.



Fonte: pxhere (2023, alterado).

Figura 4 – Detecção facial em imagem com diversas faces.

Fonte: pxhere (2023, alterado).

Por resultado, foram obtidos comportamentos finais semelhantes para ambos os programas desenvolvidos. Entretanto, vale ressaltar que, para o programa que usou HOG, foi observada maior simplicidade e leveza, eliminando a necessidade do pré-processamento da imagem, além de não ser necessário o uso do arquivo em formato XML, como no código com Haar Cascade. Também é válida a observação de que, para o código com HOG, em nenhum momento ocorreram situações de falso positivo, sem a necessidade de ajuste de parâmetros.

b) Reconhecimento de expressões

Para dar início ao código, importaram-se as bibliotecas OpenCV, Numpy, Pandas, Matplotlib, ZipFile, TensorFlow, Itertools e Confusion_Matrix, que dotam de papel fundamental ao desempenho da aplicação. Com a aplicação conectada ao banco de dados, foram extraídos arquivos e feito a leitura de um arquivo no formato CSV (valores separados por vírgula), cuja estrutura é descrita na Figura 5:

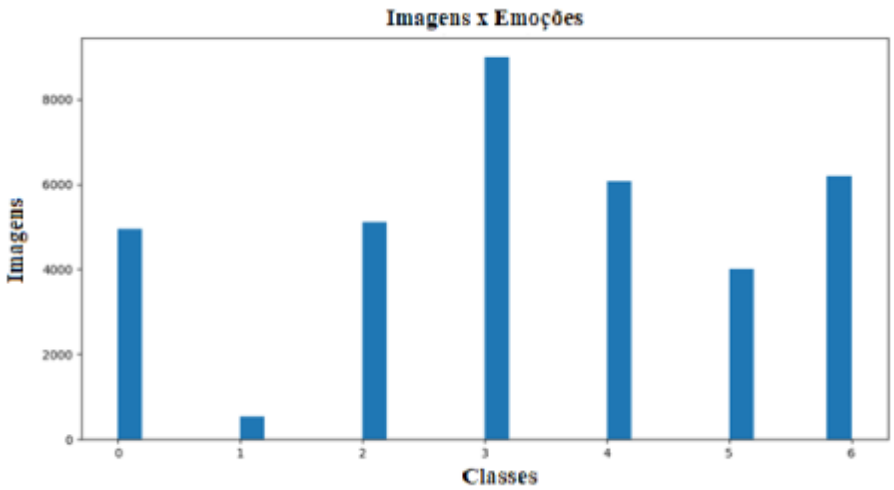
Figura 5 - Estrutura do arquivo CSV referente ao bando FER2013.

	emotion	pixels	Usage
35882	6	50 36 17 22 23 29 33 39 34 37 37 37 39 43 48 5...	PrivateTest
35883	3	178 174 172 173 181 188 191 194 196 199 200 20...	PrivateTest
35884	0	17 17 16 23 28 22 19 17 25 26 20 24 31 19 27 9...	PrivateTest
35885	3	30 28 28 29 31 30 42 68 79 81 77 67 67 71 63 6...	PrivateTest
35886	2	19 13 14 12 13 16 21 33 50 57 71 84 97 108 122...	PrivateTest

Fonte: autores.

A Figura 5 ilustra os últimos 5 últimos registros do banco de dados, organizados conforme: A primeira coluna refere-se ao número da imagem, evidenciando a quantidade de dados no banco, a segunda coluna descreve a classe (expressão facial) a qual o dado pertence e a terceira, refere-se aos *pixels* das imagens. Com base na leitura do arquivo CSV, foi feito um gráfico ilustrando a quantidade de imagens por classe, conforme Figura 6:

Figura 6 – Relação de dados por classe.

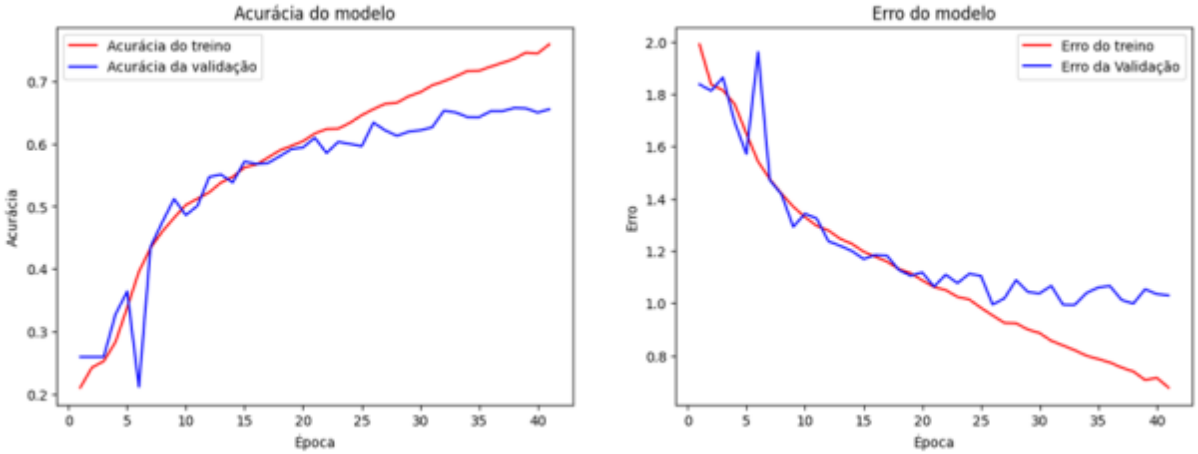


Fonte: autores.

A fim de facilitar o processamento, foi considerada a leitura dos conjuntos dos *pixels* das imagens, analisando assim, apenas um fator descritivo da imagem, não necessitando realizar a análise sobre as demais dimensões (largura, altura e escala de cores). Também foi manipulado o formato do banco para adequação ao TensorFlow e feita a normalização dos vetores para padronizar a faixa de valores dos parâmetros da rede neural (Isso ocorre porque há a diminuição na escala de valores que, em *pixels*, estavam em um intervalo de 0 a 255, e após a normalização, o intervalo passou a ser de 0 a 1).

Com os dados ajustados às condições da aplicação, foram importadas as ferramentas do TensorFlow para a posterior configuração da rede neural. Nessa configuração, foram definidas as camadas de convolução e *pooling*, com ativação ReLU e *dropouts* para desligamento aleatório de neurônios durante o processamento. Com rede configurada, foi feita a compilação e treinado o modelo. Para a configuração definida, foi verificada uma acurácia de 65% para a parcela de validação e plotados os gráficos correspondentes ao comportamento da aplicação, conforme a Figura 7.

Figura 7 – Comportamento do modelo para as 7 classes de expressões.



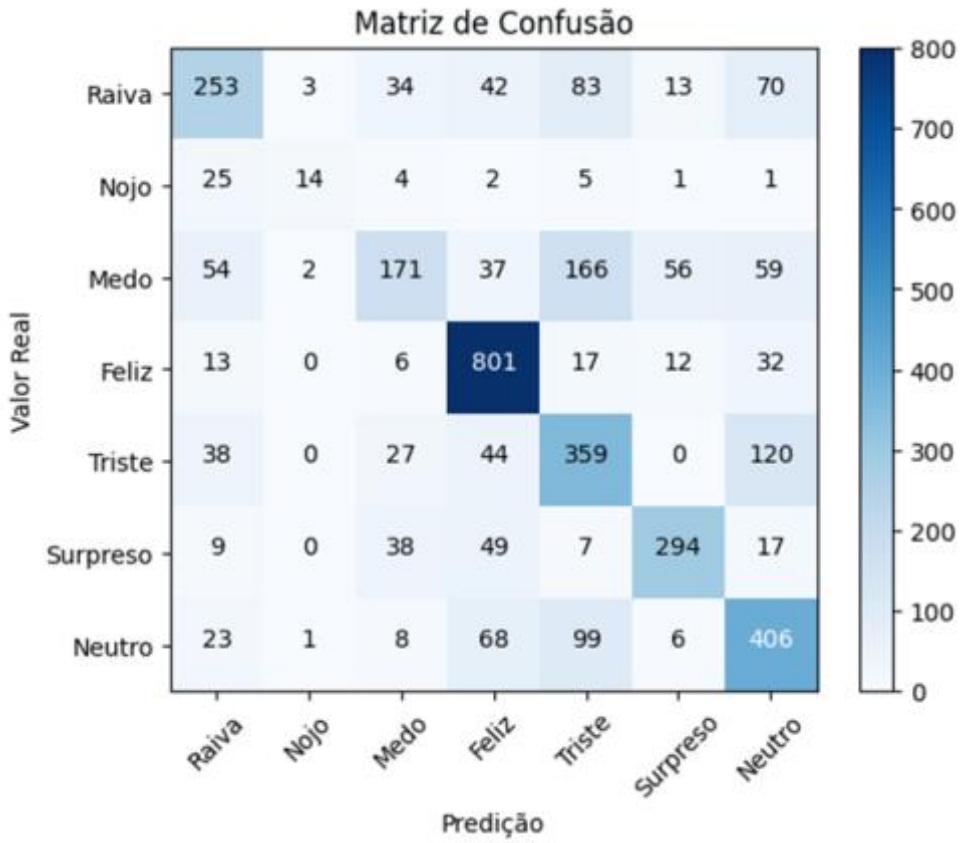
Fonte: autores.

Com base em análise dos gráficos plotados, é possível a observação de que a acurácia de validação foi mais próxima à acurácia de treino quando o treinamento se deu próximo à época de número 20 e em seguida, os resultados divergiram. Pode-se constatar comportamento semelhante quando considerada a plotagem referente aos valores dos erros de treino e validação. Vale ressaltar que, conforme aumenta o número de épocas, é possível que se obtenham resultados menos satisfatórios. Considerando que a análise foi feita em relação a sete classes, é tido que a acurácia de 65% é muito maior do que a de um classificador

aleatório, cuja acurácia seria próxima a 14%, pois a probabilidade de acerto seria de uma em sete para cada imagem.

A fim de melhorar a visualização dos resultados, foi construída uma matriz de confusão, conforme a Figura 8.

Figura 8 – Matriz de confusão.

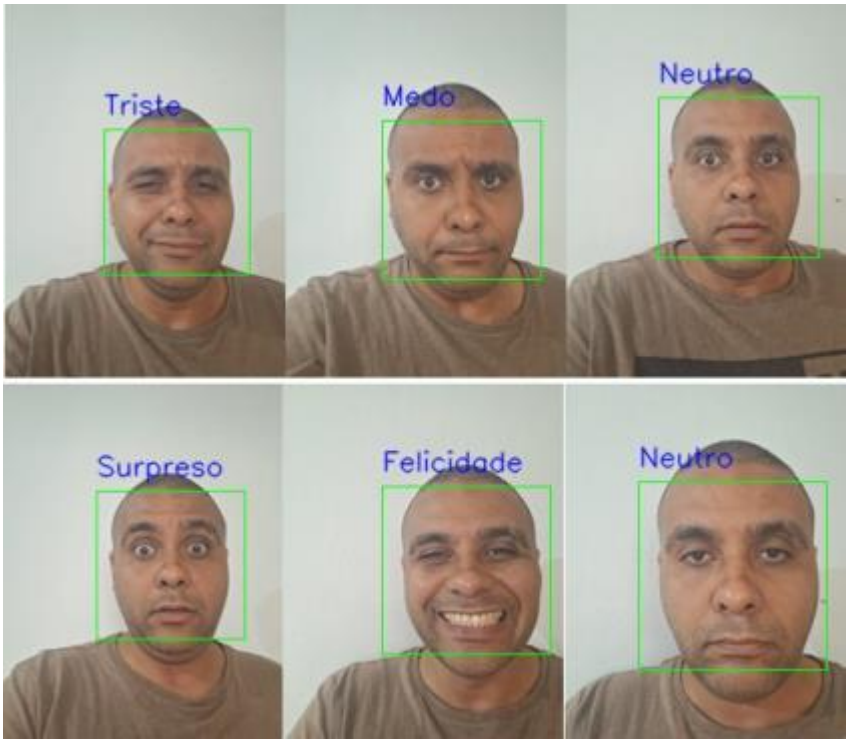


Fonte: autores.

Para a construção desta matriz, tem-se os valores reais para cada dado em relação à classificação definida por meio da predição do modelo. Dessa forma, organiza-se em escala de intensidade de cor (sendo a cor mais escura referente ao maior número de acertos e a cor mais clara referente ao menor número de acertos), valendo a observação de que a volumetria dos dados influi sobre o comportamento dos resultados. É possível constatar que para classes de expressões que são passíveis de confusão entre si (tristeza e medo, por exemplo), apresentam comportamento com maiores quantidades de erros.

Visando a ilustrar o comportamento da aplicação, foi utilizado HaarCascade para a detecção de faces e descrita a classe à qual o sistema atribui a expressão detectada, utilizando imagens do conjunto de teste, conforme a Figura 9.

Figura 9 – Ilustração do comportamento do sistema.

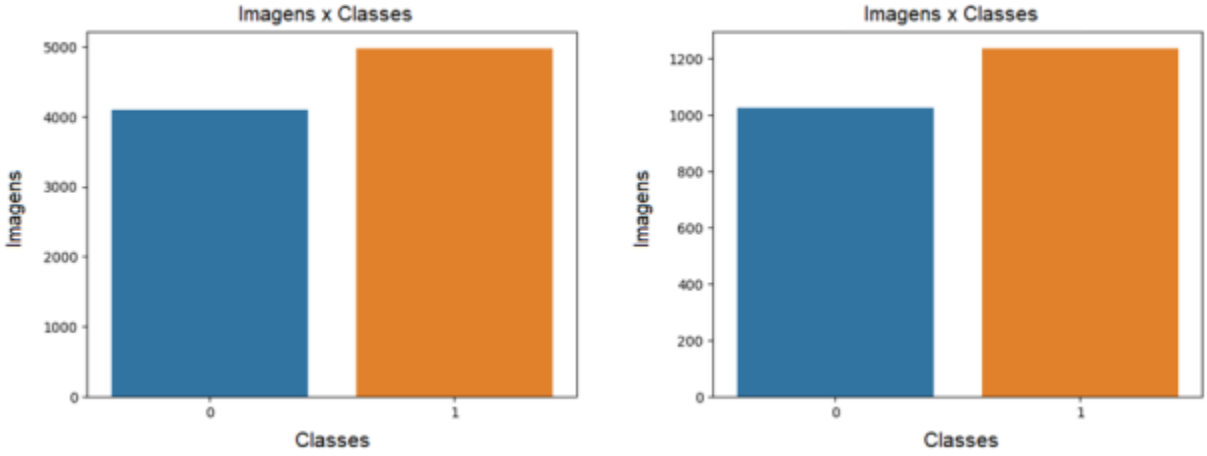


Fonte: autores.

Considerando que a configuração da rede neural apresentou resultados promissores, vista a acurácia de 65% obtida, foi dado prosseguimento, adequando a aplicação a estudos futuros, que consistem na detecção de expressões faciais que indicam dor. Para tal, foi necessária abstração para a construção de situações análogas às quais se presume possíveis. Dessa forma, considerou-se realizar novamente o treino da rede neural, considerando apenas duas classes. Os testes foram realizados em duas duplas de classes, sendo estas compostas, em um dos testes, por expressões que são passíveis de confusão entre si, e no outro, por expressões que são opostas entre si. Tal consideração se dá com o objetivo de presumir, com base nos resultados a se obter, qual seria o comportamento da aplicação para casos em que as expressões ocorrem com mais e menos intensidade.

Para o teste com classes que se confundem, foram consideradas as expressões de medo (0) e neutro (1). Os passos para a execução da aplicação são similares aos já descritos. Entretanto, o banco de dados sofreu modificações para comportar apenas as classes escolhidas para o teste, de forma que o conjunto de treinamento consistiu de aproximadamente 9 mil imagens, enquanto o conjunto de validação consistiu de aproximadamente 2 mil imagens, conforme a Figura 10.

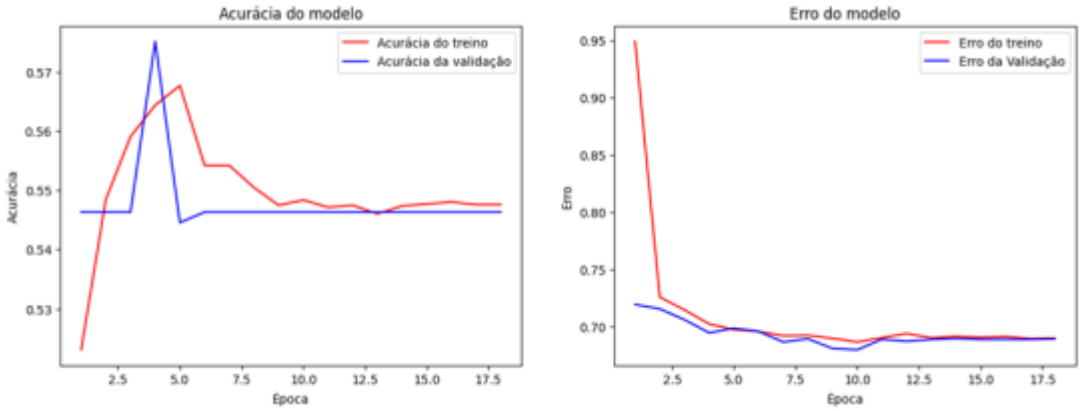
Figura 10 – Relação de imagens por classe, medo e neutro, nos conjuntos de treino e validação.



Fonte: autores.

Mantendo as configurações da rede neural, foi realizado o treinamento, que apresentou acurácia de validação de 57%, conforme a Figura 11.

Figura 11 – Comportamento da aplicação para classes que se confundem.



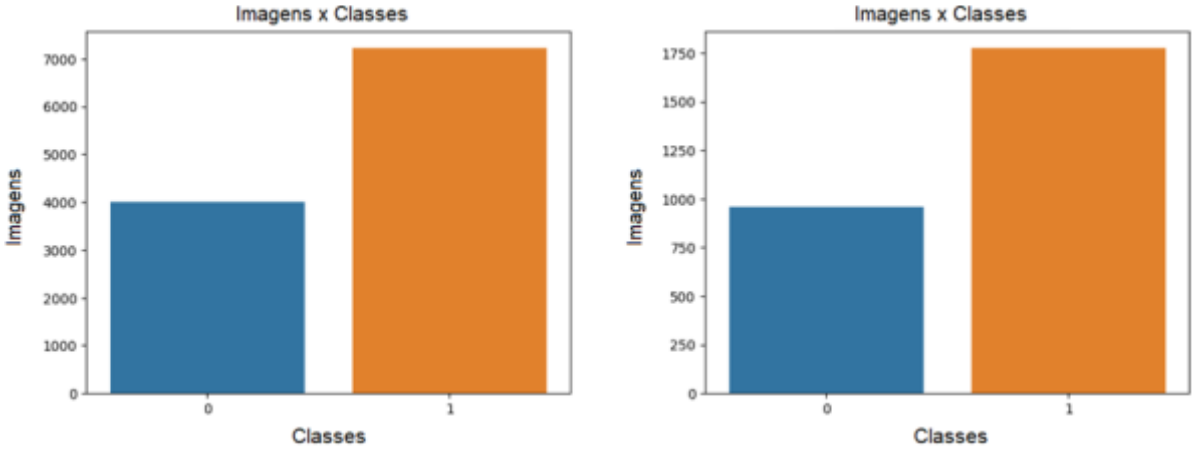
Fonte: autores.

Tal comportamento foi esperado, visto que o teste se deu de modo a considerar classes que se confundem. Dessa forma, foi possível presumir que, para uma situação em que a expressão se dá de forma menos intensa, a aplicação não apresenta resultados de alta acurácia.

Para o teste com classes muito diferentes entre si foram consideradas as expressões de raiva (0) e felicidade (1). Os passos para a execução da aplicação são similares aos já descritos. Entretanto, o banco de dados sofreu modificações para comportar apenas as classes escolhidas para o teste de forma que, o conjunto de treinamento consistiu de

aproximadamente 11 mil imagens, enquanto o conjunto de validação consistiu de aproximadamente 2 mil imagens, conforme a Figura 12.

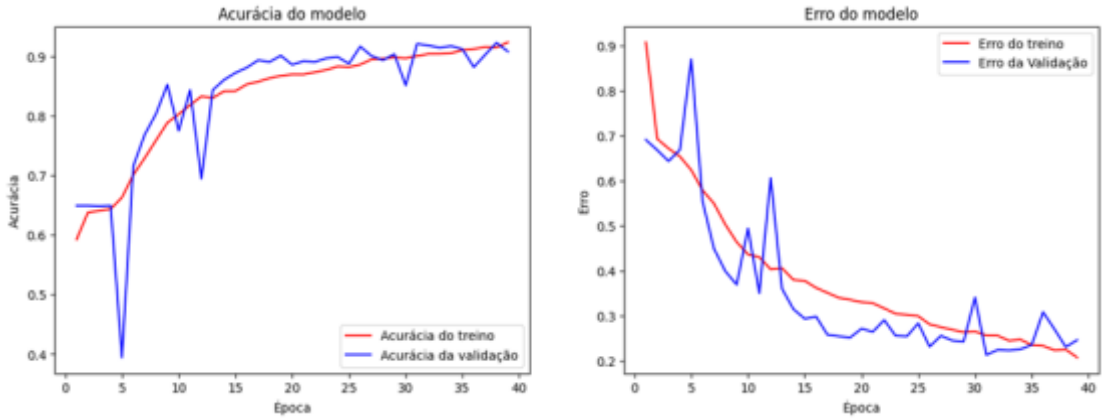
Figura 12 – Relação de imagens por classe, raiva e felicidade, nos conjuntos de treino e validação



Fonte: autores.

Mantendo as configurações da rede neural foi realizado o treinamento que apresentou acurácia de validação de 92% em seu pico, conforme plotagem:

Figura 13 - Comportamento da aplicação para classes de características opostas.



Fonte: autores.

Com base nos resultados obtidos, foi possível presumir que, para casos em que as expressões se dão de forma mais intensa, a aplicação apresenta acurácia mais alta.

Conclusão

O sistema desenvolvido apresentou desempenho promissor, com capacidade de reconhecimento de emoções com acurácia aproximada de 92%, se considerados pares de

classes bastante contrastantes. Entretanto, quando o objetivo foi discernir entre dois tipos de emoções de classes mais próximas, houve baixa acurácia (em torno de 60% ou menor). Com base nos dados obtidos, é possível a conclusão de que, para casos em que se observam maiores intensidades durante a realização de uma expressão facial, existem maiores chances de que os sistemas consigam predições mais acuradas, ao passo que, para os casos em que a intensidade é menor, os comportamentos dos sistemas ficam sujeitos a confusões. Vistas tais constatações, vale que, para contornar as situações de demonstrações de emoções com menores intensidades, poderia ser válida a construção de um banco de dados mais específico para o treinamento de identificação de menores distinções entre as classes, realizando uma análise mais aprofundada sobre o que se observa sobre sinais mais detalhados durante as mudanças entre uma expressão e outra.

Referências

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools, 2000.* Disponível em: https://www.researchgate.net/publication/233950935_The_Opencv_Library. Acesso em: 24 de novembro de 2023.

Coutrin, G. de A. S., & Thomaz, C. E. (2023). Redes Neurais Convolucionais para Avaliação de Dor Neonatal em Imagens de Face: Uma Análise Quantitativa e Qualitativa. In: *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde.* Disponível em https://sol.sbc.org.br/index.php/sbcas_estendido/article/view/25337. Acesso em 24 de novembro.

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

Ekman, P. (1972). *Emotion in the human face.* Malor Books.

Ekman, P. (2003). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.* New York, NY: Times Books.

Fridlund, A. J., et al. (1997). *Facial Expressions. The psychology of facial expression (1997), v. 103.*

Happy, S. L.; Routray, Aurobinda. Automatic facial expression recognition using features of salient facial patches. **IEEE transactions on Affective Computing**, v. 6, n. 1, p. 1-12, 2014.

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362.

Heiderich, T. M. (2013). Desenvolvimento de software para identificar a expressão facial de dor do recém-nascido. *Tese de doutorado apresentada à Unifesp em 2013.* Disponível em https://repositorio.unifesp.br/handle/11600/22891. Acesso em 24 de novembro.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering, 9*(3), 90-95.

Huamán, A. (2023). Cascade Classifier. Disponível em: https://docs.opencv.org/4.x/db/d28/tutorial_cascade_classifier.html Acesso em 24 de novembro de 2023.

Lencioni, G. C., & Zanella, A. J. (2020). Detecção automática de dor em equinos por reconhecimento facial. *Resumos, 2020.* Disponível em https://repositorio.usp.br/bitstreams/0ecc5bd1-129d-409b-a772-bf97ac64e30f. Acesso em 24 de novembro.

Maqableh, W., Alzyoud, F. Y., & Zraqou, J. (2023). The use of facial expressions in measuring students' interaction with distance learning environments during the COVID-19 crisis. *Visual informatics, 7*(1), 1-17.

Nogueira, Maria Francisca Magalhães; Faria, Claudia Sousa Oriente de. A comunicação não-verbal nas organizações: o corpo fala. Comunicologia: revista de comunicação e epistemologia da Universidade Católica de Brasília, v. 6, n.1, p. 1-1, 2014.

Nunes, M. H. V., Pacheco, A. D., & Wagatsuma, J. T. (2020). Reconhecimento e avaliação da dor em bovinos: Revisão. *Pubvet, 15*, 181. Disponível em https://pdfs.semanticscholar.org/3228/611ddee9274d8d18244e8eb1821a038c0076.pdf. Acesso em 24 de novembro.

Pxhere. Melhores fotos grátis em um só lugar. Disponível em <https://pxhere.com/pt/>. Acesso em 24 de novembro.

Radoux, J., & Bogaert, P. (2020). About the Pitfall of Erroneous Validation Data in the Estimation of Confusion Matrices.

Rossum, Guido; Drake Jr, Fred L. **Python tutorial**. Amsterdam, The Netherlands: Centrum voor Wiskunde en Informatica, 1995.

Shahriar, N. (2023). What is Convolutional Neural Network — CNN (Deep Learning). Disponível em https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5 Acesso em 24 de novembro de 2023.

Silva, J. A. da, & Silva, M. J. P. da. (1995). Expressões faciais e emoções humanas levantamento bibliográfico. *Revista Brasileira De Enfermagem*, *48*(2), 180–187. https://doi.org/10.1590/S0034-71671995000200013. Acesso em 24 de novembro.

Silva, L. M. G. da, Brasil, V. V., Guimarães, H. C. Q. C. P., Savonitti, B. H. R. de A., & Silva, M. J. P. da. (2000). Comunicação não-verbal: reflexões acerca da linguagem corporal. *Revista Latino-americana De Enfermagem*, *8*(4), 52–58. https://doi.org/10.1590/S0104-11692000000400008. Acesso em 24 de novembro.

Slimani, K., et al. (2018). Facial emotion recognition: A comparative analysis using 22 LBP variants. *Proceedings of the 2nd Mediterranean Conference on Pattern Recognition and Artificial Intelligence*, 88-94.

Vigostki, L. Pensamento e Linguagem. São Paulo: editora Martins Fontes, 2008.