UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO – UENF

LARISSA RIBEIRO SARDINHA

**RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA ANÁLISE SEQUENCIAL DE IMAGENS**

TRABALHO DE CONCLUSÃO DE CURSO

Campos dos Goytacazes, RJ

Julho de 2025.

**RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA ANÁLISE SEQUENCIAL DE IMAGENS**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Estadual do Norte Fluminense Darcy Ribeiro como requisito para a obtenção do título de Bacharel em Ciência da Computação, sob orientação de Prof. Luis Antonio Rivera Escriba.

Universidade Estadual do Norte Fluminense Darcy Ribeiro – UENF

Centro de Ciência e Tecnologia – CCT

Laboratório de Ciências Matemáticas – LCMAT

Curso de Ciência da Computação

*Este projeto é dedicado a Deus e a minha querida amiga Ediene Freitas Rodrigues da Costa(In Memoriam). Sem eles esse projeto não seria possível.*

**AGRADECIMENTOS**

A Deus, pela minha vida e por me permitir superar todos os obstáculos encontrados ao longo da realização deste trabalho.

Aos meus pais, Uilis e Alailza, por sempre acreditarem em mim e por me apoiarem na conquista de espaços que eles próprios não tiveram a oportunidade de conquistar. Vocês são minha inspiração e minha base.

À minha irmã Rayssa, por compartilhar comigo as experiências ao longo dessa caminhada que chamamos de vida, e aos meus sobrinhos Sophia e Lucca, por serem o meu refúgio e o meu lar.

Aos amigos que esta universidade me proporcionou, em especial ao Jhonatan, Juliana, Nicole e Pablo, que dividiram comigo o dia a dia e a jornada acadêmica, sempre com muito café, sorrisos e ombros para os momentos de desânimo. Vocês tornaram essa caminhada mais leve e significativa.

Ao corpo docente, por compartilhar conhecimento e sabedoria, inspirando-me a buscar sempre mais.

À universidade, por proporcionar a estrutura e o ambiente necessários para o meu crescimento acadêmico e pessoal, e ao seu corpo técnico, pelo apoio essencial em todas as etapas desta trajetória.

Por fim, mas não menos importante, à escola Leôncio Pereira Gomes e ao Instituto Federal Fluminense, que formaram a base do meu conhecimento; a Bolsa de Apoio Acadêmico e de Auxílio Moradia, essenciais para minha permanência neste curso.

# INTRODUÇÃO

Vivemos em um era altamente tecnológica, onde tarefas cotidianas foram digitalizadas visando um melhor funcionamento e precisão. Entre elas, destaca-se o reconhecimento de emoções, que busca interpretar expressões faciais humanas de forma automática através da análise de expressões faciais em vídeo, emergindo como um campo de pesquisa fundamental e desafiador.

Entende-se como reconhecimento de emoções a capacidade de analisar visualmente a configuração e os movimentos dos músculos faciais para identificar a emoção predominante em determinada expressão (WILHELM et al., 2014). Trata-se de uma tarefa desafiadora de ser realizada com precisão sem o auxílio de ferramentas computacionais.

Nesse contexto, a visão computacional, aliada aos avanços em aprendizado profundo, oferece ferramentas poderosas para abordar essa complexidade. Técnicas de análise de imagem e vídeo permitem a detecção e classificação automatizada de expressões faciais, transformando dados visuais brutos em interpretações emocionais significativas (Bhatt et al., 2020).

O Reconhecimento de Emoções Faciais (Facial Emotion Recognition – FER) consolidou-se como uma subárea vital da visão computacional, focada na identificação de categorias emocionais a partir de dados faciais (Cîrneanu et al., 2023). Abordagens baseadas em aprendizado profundo, principalmente as redes neurais artificiais têm demonstrado grande sucesso.

As redes neurais convolucionais (CNNs) provaram ser particularmente eficazes na extração de características espaciais relevantes dos pixels da face, capturando padrões visuais complexos. Enquanto isso, as redes neurais recorrentes (RNNs), como LSTMs ou GRUs, são frequentemente empregadas para modelar a dinâmica temporal das expressões, crucial para a análise de vídeos (AWARI, 2023).

Contudo, a eficácia desses modelos depende da qualidade das características extraídas. A vasta quantidade de informação presente em cada quadro de vídeo, somada à variabilidade inerente às expressões, exige métodos robustos de caracterização para garantir a acurácia.

Apesar dos avanços, a tarefa de reconhecimento emocional enfrenta diversos desafios, como condições ambientais desfavoráveis, oclusões, múltiplos rostos em uma mesma imagem, variações de escala, orientação e iluminação, além das diferenças culturais na manifestação das emoções (Cîrneanu et al., 2023).

Assim, torna-se imprescindível que os algoritmos empregados sejam robustos e apresentem alta confiabilidade, especialmente em aplicações que demandam respostas precisas com base no estado emocional do usuário.

## Formulação do problema

Considerando os desafios no reconhecimento de emoções em vídeo, como a combinação de descritores de características clássicos (HOG) e características aprendidas por redes neurais convolucionais (CNN), seguida por classificação com redes neurais recorrentes (RNN), pode aprimorar a eficiência e a acurácia na identificação de expressões faciais dinâmicas em comparação com abordagens que utilizam apenas CNNs para extração de características?

## Hipótese

A combinação de características extraídas pelo descritor HOG com as extraídas por uma Rede Neural Convolucional (CNNs), quando utilizada para alimentar Redes Neurais Recorrentes (RNNs) para classificação temporal, resulta em maior acurácia no reconhecimento de emoções faciais em vídeo. Acredita-se que essa abordagem captura as transições emocionais de forma mais eficaz do que métodos que dependem apenas de características de CNNs para a análise temporal.

## Objetivo

**Objetivo Geral:** Desenvolver e avaliar um sistema de reconhecimento de emoções faciais em vídeo, empregando uma abordagem híbrida de extração de características (HOG+CNN) e classificação temporal (RNN), visando aprimorar a precisão na identificação de expressões dinâmicas em comparação com métodos baseados apenas em CNN.

**Objetivos Específicos:**

1. Apresentar uma estrutura conceitual do reconhecimento facial de emoções.
2. Implementar uma etapa de pré-processamento para detecção facial e descritores HOG + uma rede neural convolucional com transfer learning para a extração de características faciais relevantes em cada frame..
3. Estabelecer um modelo eficiente de identificação de emoções mais comuns com base na análise sequencial de frames utilizando redes neurais recorrentes.
4. Avaliar quantitativamente o desempenho do sistema proposto (HOG+CNN+RNN) utilizando métricas padrão (acurácia, F1-score e matriz de confusão) em um dataset público.

## Justificação

A precisão e a confiabilidade em sistemas de reconhecimento de emoções possuem um alto índice de importância visto que essa técnica traz melhorias pra diversas aplicações como obter informações sobre o comportamento e as preferências de clientes para o marketing; auxiliar os profissionais de saúde a monitorarem e diagnosticarem problemas emocionais em pacientes.

Diante dessas limitações, a hipótese de que a combinação de descritores robustos como HOG com o poder de aprendizado das CNNs pode levar a uma representação de características mais completa e eficaz para a análise temporal com RNNs motiva este estudo, que busca contribuir para o avanço da área.

## Método

A combinação da CNN + RNN é baseada no trabalho de Rangulov;Fahum (2020) na qual se utiliza uma CNN para a extração de característica e uma RNN para a classificação, modelando assim a sequência temporal.

O banco de dados utilizado é o dataset KDEF-DYN, que possui vídeos de atores expressando determinadas emoções. Cada vídeo será dividido em frames, que serão analisados individualmente. Esses vídeos devem ser pré-processados e as características chaves devem ser extraídas.

Assim criaremos um vetor concatenado de extrações da CNN e de um Histograma de Gradiente Orientado (em Inglês *Histogram of Oriented Gradients*).

O vetor de características será direcionado para a entrada de um LSTM (Memória de Curto Longo Prazo do *Inglês Long Short Term Memory*), que realizará a classificação da emoção.

Finalmente, a validação do sistema com dados de banco de dados de referência, em função da precisão, acurácia e F1-score.

## Organização do trabalho

A presente pesquisa é estruturada em 5 capítulos. O primeiro capítulo é introdutório, contextualizando o que é o reconhecimento facial de emoções e qual a sua importância. O segundo capítulo oferece um referencial teórico com revisão bibliográfica dos conceitos abordados. O terceiro capítulo apresenta o modelo proposto do sistema, com detalhes sobre como foi estruturado e também sobre as tecnologias escolhidas para a implementação. O quarto capítulo apresenta o desenvolvimento e experimentação, dentre eles: captura de imagem, o pré-processamento, treinamento e teste do modelo proposto. Por fim, o quinto capítulo contém a conclusão da pesquisa e os trabalhos futuros.

# CÁPITULO II: RECONHECIMENTO FACIAL DE EMOÇÕES

As emoções são parte de nós, parte do nosso dia a dia e de nossa interação como sociedade; porém, nem sempre é fácil saber identificá-las, principalmente quando queremos decifrar qual emoção o outro está sentindo e o que exatamente essa emoção significa.

Porém, há algo que nos entrega: nossa expressão facial. O famoso dito popular “está escrito na sua face” é verdadeiro. Mesmo sem o nosso consentimento, diversas vezes as nossas expressões faciais entregam o nosso estado emocional.

A detecção visual de variações de expressões faciais que caracterizam uma emoção exige o conhecimento das diferentes configurações expressivas, o que torna o processo complexo. Com o avanço das técnicas de visão computacional, esse desafio tem sido enfrentado com maior eficiência. Em um mundo altamente visual, onde imagens e vídeos são constantemente produzidos, cresce a demanda por soluções capazes de interpretar essas informações.

Essa evolução foi impulsionada pelos avanços em técnicas de aprendizado de máquina e, particularmente, pelo desenvolvimento de arquiteturas de aprendizado profundo capazes de modelar as complexas relações entre configurações faciais e estados emocionais subjacentes.

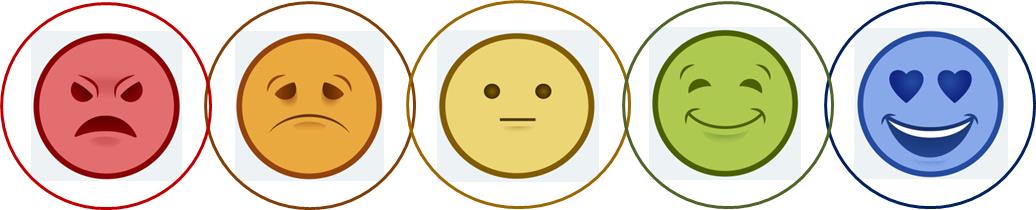
## Emoções

As emoções são manifestações que envolvem reações intensas e breves do organismo em resposta a um evento inesperado ou, por vezes, muito aguardado e fantasiado (Bock et al, 2008); são processos complexos que envolvem tanto aspectos psicológicos quanto fisiológicos.

 Segundo Damásio (1994), as emoções são respostas automáticas do organismo a estímulos ambientais, mediadas por estruturas cerebrais como o hipotálamo e o córtex pré-frontal. Essas estruturas trabalham em conjunto para gerar respostas emocionais que podem ser observadas em nossas reações.

Essas reações são importantes para a liberação de tensão, mas muitas das vezes saem do controle humano, tornando assim quase impossível não transparecer fisicamente quando algo nos afeta, mesmo quando o nosso desejo é simular uma neutralidade; assim, de acordo com Bock et al. (2008), as nossas feições acabam nos traindo e demonstrando a emoção que tanto lutamos para esconder.

Não se sabe ao certo a quantidade de emoções que existem, mas é possível observar na literatura que muitos autores determinam um conjunto de emoções básicas, sendo elas: alegria, medo, surpresa, tristeza, nojo e raiva (Miguel, 2015), tal como ilustra a Figura 2.1.

Figura 2.1. Cinco emoções básicas (raiva, medo, calma, alegria, surpresa).

## Expressões faciais

Em 1976 surgiu uma forte teoria da identificação de emoções a partir da análise facial com a publicação do artigo *Measuring Facial Movements* (em português, “Medindo o Movimento Facial”) por Ekman & Friensen (1976), em que os autores concluem que as expressões faciais são manifestações biológicas consistentes das emoções humanas.

Eles desenvolveram então um sistema de codificação de ação facial (*Facial Action Coding System* - FACS), que classifica movimentos musculares faciais possíveis em 46 unidades de ação (em inglês, *Actions Unit* (AU)), esse sistema é utilizado até hoje para o treinamento e avaliação de sistemas automáticos de reconhecimento de emoções.

A conclusão é a de que uma expressão emocional é resultado de unidades de ações constituintes em um determinado conjunto. Por tanto, uma expressão emocional é caracterizada através de uma série de movimentos.

As contrações musculares e movimentos de áreas do nosso rosto como olhos, pálpebras, testa, sobrancelhas e da parte inferior do rosto (ao redor da boca), contribuem principalmente para a manifestação dessas emoções, por tanto a detecção dessas contrações é possível apenas através da análise de uma sequência de imagens (segmento de vídeo) relativo ao possível gesto (Rangulov;Fahum, 2020).

Podemos observar na Figura 2.2, uma representação de expressões faciais das seis emoções básicas (1. nojo, 2. medo, 3. alegria, 4. surpresa, 5. tristeza, 6. raiva).

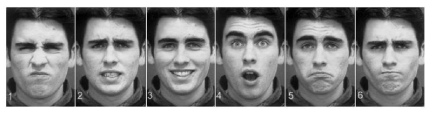


Figura 0.2 - Fonte (Schmidt;Cohn, 2001)

A configuração facial de uma emoção é a união de várias movimentações musculares faciais subjacentes, que refletem processos nossos processos mentais e afetivos.

Para Cacioppo et al. (2000), as expressões faciais não surgem de forma abrupta, mas sim por meio de uma transição suave, na qual diferentes grupos musculares são ativados em sequência ou paralelamente; assim em alguns momentos a configuração muscular de uma emoção converge temporariamente com a de outra, criando expressões ambíguas ou mistas.

## Visão Computacional e Reconhecimento Facial

A visão computacional é definida por Bhatt et al.(2020) como uma área da inteligência artificial que permite a interpretação automatizada de imagens e vídeos, possibilitando assim a identificação de objetos a partir de pixels de imagens brutas.

Conforme definido por Shanmugamani (2018), este campo engloba técnicas e algoritmos que permitem a extração, análise e compreensão de informações significativas a partir de imagens e vídeos. Assim, o principal objetivo da visão computacional é “fazer com que as máquinas vejam o mundo da mesma forma que os humanos” (Bhatt et al., 2020, p. 1-2).

Essa técnica desempenha um importante papel no processo de reconhecimento facial, pois possibilita a detecção facial automática em imagens, identificando e realizando a extração de características faciais que possibilitam a análise da emoção.

O reconhecimento facial através da visão computacional surgiu inicialmente como uma técnica para identificar indivíduos a partir de seus rostos, utilizando bancos de dados para autenticar a identidade (Bastos; Esteves, 2021). A partir dessa aplicação, desenvolveu-se o reconhecimento automatizado de emoções faciais, baseado na análise algorítmica de gestos e expressões humanas sequenciais com o objetivo de identificar sinais emocionais (Cîrneanu et al., 2023)

O processo do reconhecimento de emoções a partir da análise facial com visão computacional possui um certo conjunto de etapas para garantir a sua eficácia, as principais são: captura dos dados, pré-processamento, extração de características e a classificação final, ilustradas na Figura 2.3

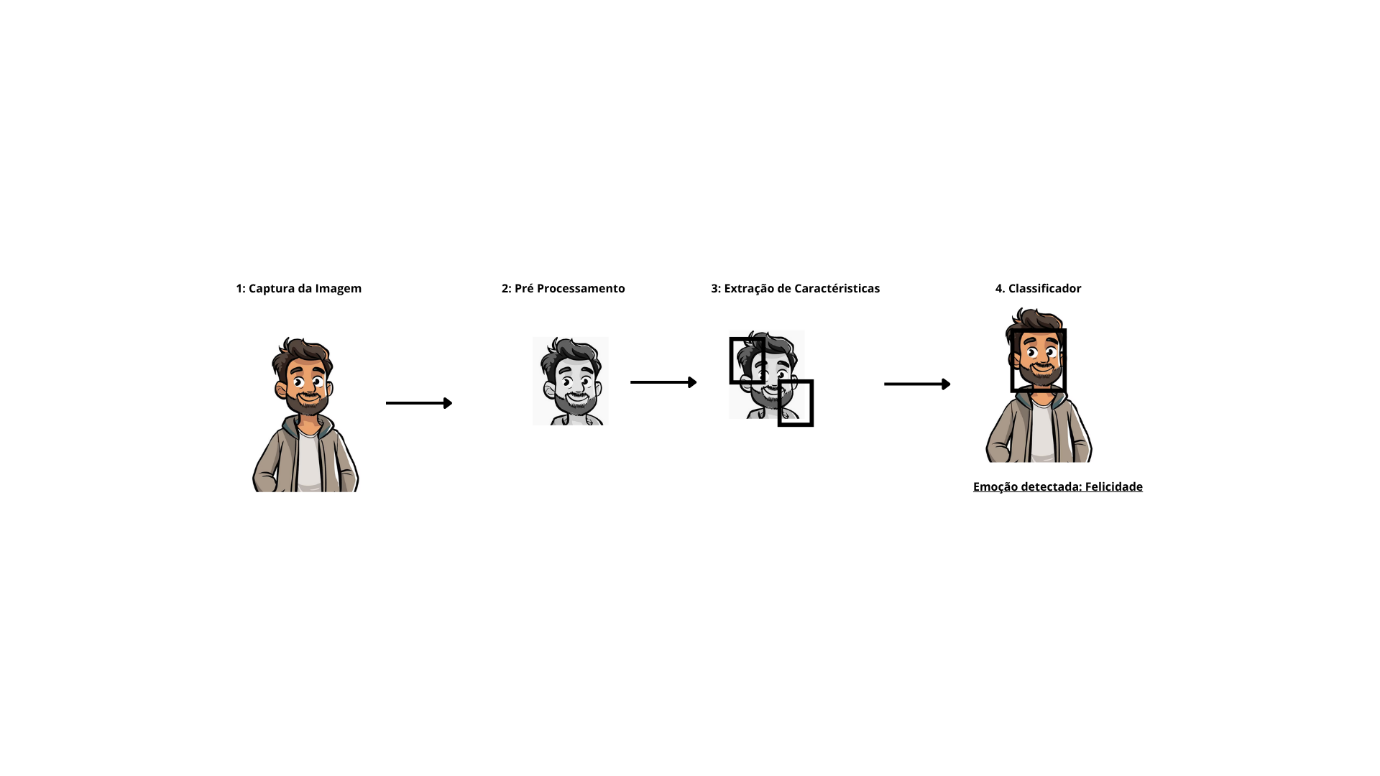


Figura 2.3. Etapas para reconhecimento de uma emoção

Para que o processo ocorra de maneira eficiente, o classificador deve ser treinado previamente com imagens já rotuladas, assim irá detectar padrões e aprender a reconhecer a emoção expressada. Desta forma, a Figura 2.4 ilustra o processo de reconhecimento de emoção facial automático da seguinte forma:

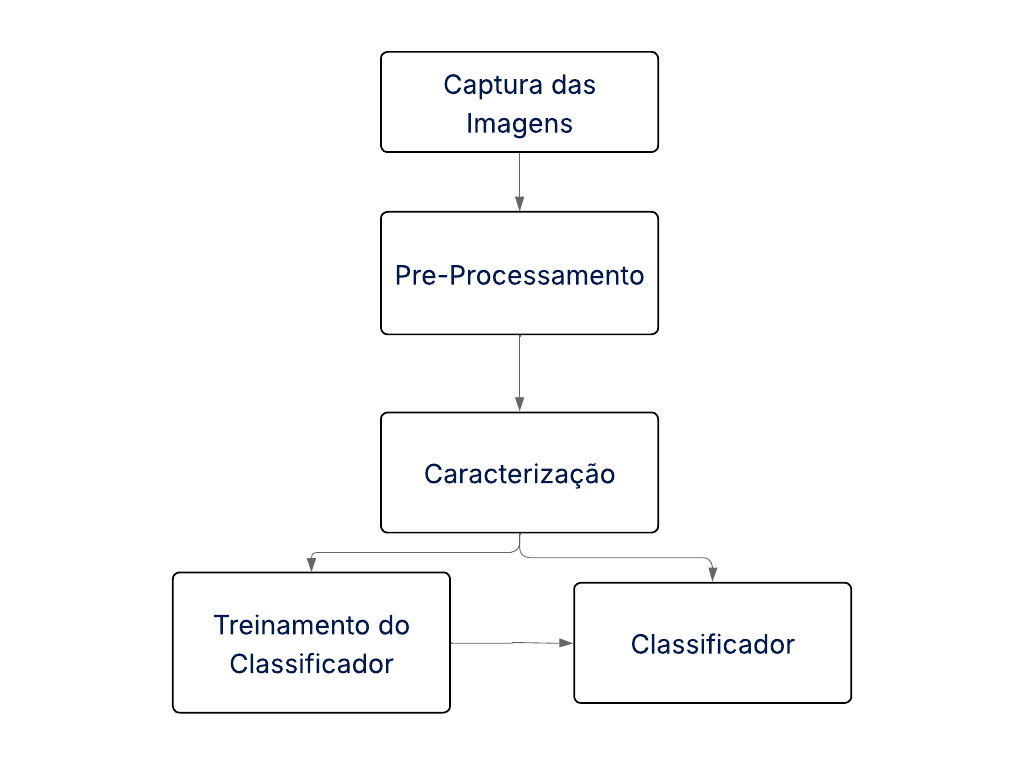


Figura 2.4: Processo de Reconhecimento Facial Automático

Assim, para entender o processo completo de um reconhecedor facial automático, é necessária uma revisão de todas as suas etapas.

* + 1. Captura das Imagens

O processo inicial de um sistema de visão computacional é a captura do dado, ou seja, a obtenção das imagens ou vídeos. Essa captação é realizada através de sensores como câmeras digitais, smartphones, webcam ou até mesmo a busca em bancos de dados específicos para cada aplicação.

Para a detecção sequencial de emoções, o dado a ser capturado é um vídeo, que nada mais é do que uma série de imagens exibidas sequencialmente, dessa forma, toda a série deve ser capturada e analisada de forma individual.

Essas imagens individuais, também conhecidas como frames, contêm dados que devem ser processados quadro por quadro (Figura 2.5), permitindo assim a detecção e a interpretação de padrões visuais dinâmicos (GONZÁLEZ; WOODS, 2010).

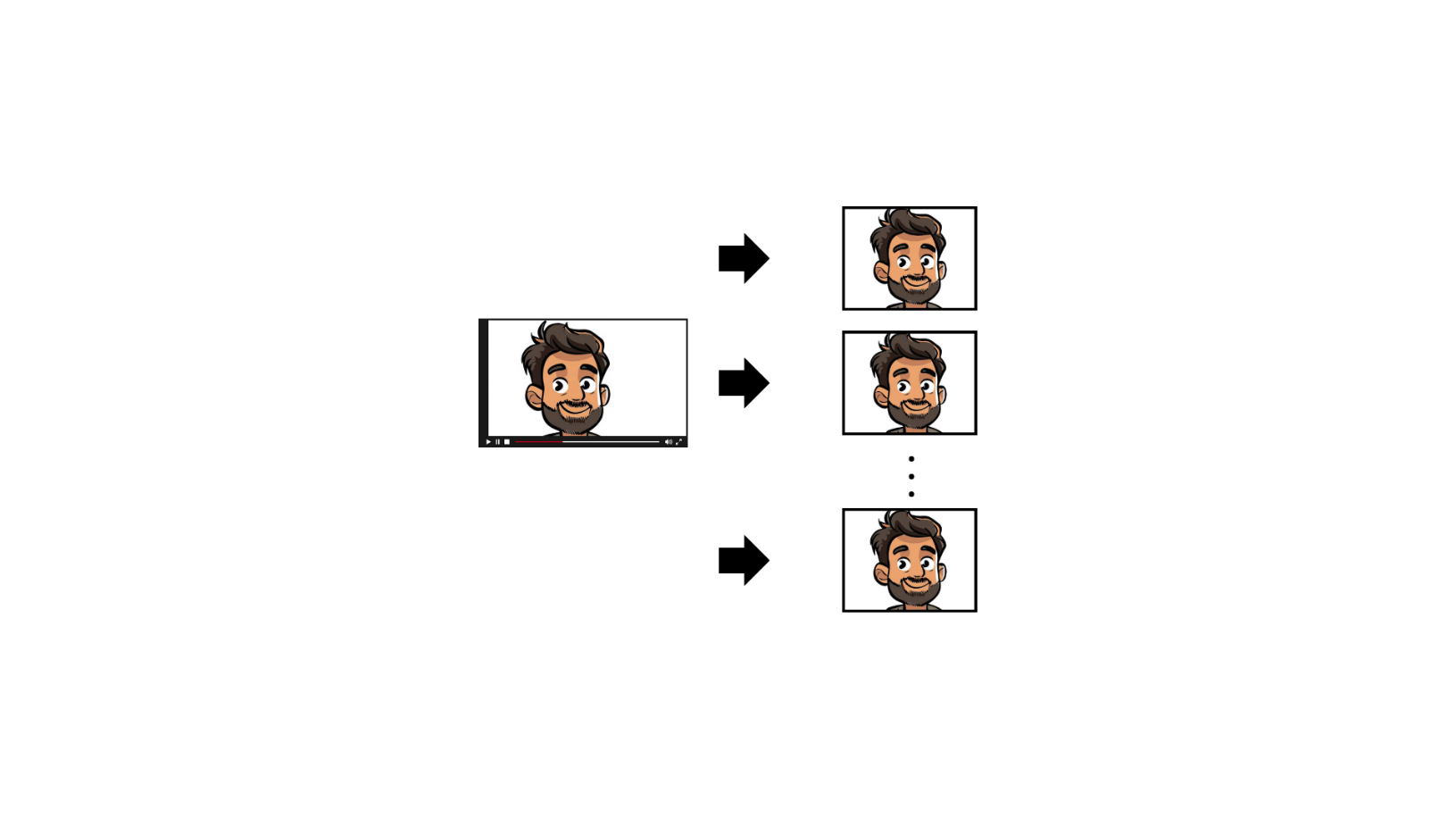


Figura 2.5: Divisão de vídeo em quadros(imagens).

Quanto maior a quantidade de frames extraídos, maior a quantidade de informação. Além disso, fatores como a qualidade da imagem e nitidez são extremamente importantes pois influenciam diretamente as próximas etapas.

Dessa forma, aspectos como resolução, taxa de quadros, iluminação, focalização e ruído devem ser considerados para garantir a integridade dos dados visuais coletados.

* + 1. Pré-Processamento

A próxima etapa é o pré-processamento, onde os dados brutos são preparados para a análise. Para isso, são aplicados um conjunto de ações visando facilitar a interpretação das características.

Algumas dessas técnicas são a aplicação de filtros para a eliminação de ruídos, a detecção da área de interesse, o redimensionamento das imagens para um tamanho padrão, rotação e etc.

Além disso, os filtros podem ser aplicados com o objetivo de alterar o nível de intensidade de cores, a luminosidade ou até mesmo a quantidade de canais de cores, visto que as informações necessárias na aplicação de reconhecimento facial de emoções independem das informações dadas pelas cores.

Outra ação é a detecção facial, ou seja, o processo de identificar e localizar rostos humanos em imagens ou vídeos, que é um dos pilares de todo esse processo. Os algoritmos criados para este fim buscam isolar a face para assim utiliza-la, dessa forma eles possuem um mesmo objetivo (Figura 2.6), mas os processos realizados são diferentes.

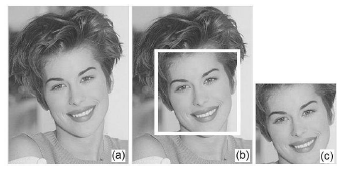


Figura 1.6: Detecção da face. (a) Imagem de entrada; (b) face detectada; (c) face isolada.

Fonte: Rowley et al (2020).

O método de detecção facial empregado por Rangulov e Fahim (2020) é o classificador Haar Cascade.

Esse algoritmo utiliza características básicas que analisam a diferença de intensidade entre regiões adjacentes da imagem, somando os valores dos pixels em áreas retangulares contíguas, com base na luminosidade dos pixels, demonstrado visualmente na Figura 2.7.

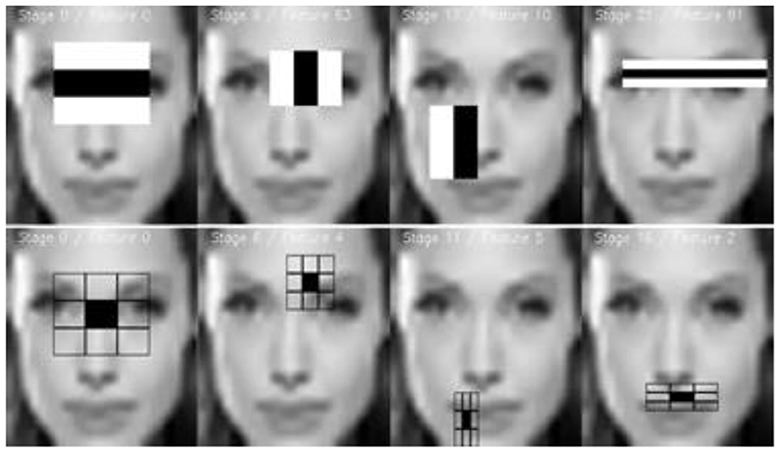


Figure 2.7: Método Haar Cascade

Fonte: Rangulov e Fahim (2020)

Outro algoritmo de detecção utilizado por Ballesteros et al. (2024) é o MTCNN (Multi-task Cascaded Convolutional Networks), uma abordagem baseada em redes neurais convolucionais organizadas em cascata que realiza a detecção de rostos e pontos faciais com alta precisão.

O MTCNN consiste em três etapas sequenciais: a Proposal Network (P-Net), que gera regiões candidatas a conter rostos; a Refine Network (R-Net), que filtra falsos positivos e ajusta as caixas delimitadoras; e a Output Network (O-Net), que aprimora a detecção e identifica pontos faciais-chave como olhos, nariz e boca.

* + 1. Caracterização

A caracterização é o processo de extrair informações relevantes ou atributos representativos de uma imagem para facilitar sua análise e reconhecimento.

Consiste basicamente em identificar e analisar detalhes específicos da face — como contornos, texturas, gradientes de intensidade e formas — que ajudam a diferenciar uma expressão facial ou uma emoção.

Essas características extraídas funcionam como uma “impressão digital” da imagem, permitindo que algoritmos posteriores possam classificar ou reconhecer o que está representado na imagem com maior precisão. Técnicas comuns para caracterização incluem o Histograma de Gradiente Orientado (HOG), LBP (Local Binary Patterns), entre outros.

* HOG (Histograma de Gradiente Orientado):

O histograma de gradiente é calculado pela distribuição das intensidades e direções dos gradientes (mudanças de intensidade) da imagem, detectando assim bordas e contornos (Anil e Padma Suresh, 2023).

As etapas do HOG podem ser vistas detalhadamente na Figura 2.8:

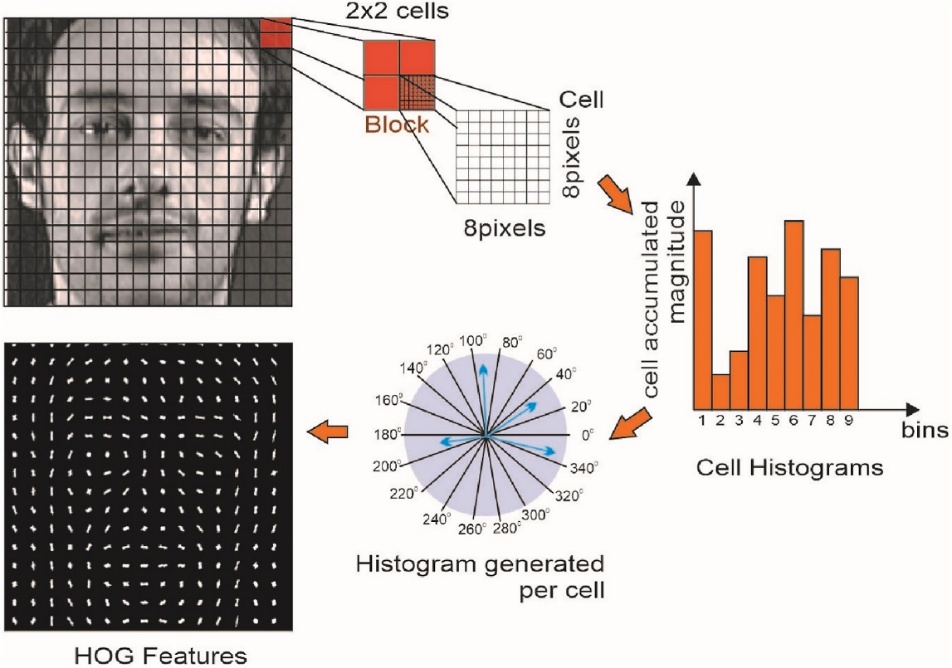


Figure 3.8: Processo de um Histograma de Gradiente Orientado

Fonte: (Anil e Padma Suresh, 2023)

A imagem de entrada é redimensionada para um tamanho padrão e convertida para tons de cinza. Em seguida, a imagem é dividida em pequenas células, e para cada pixel são calculadas a magnitude e a direção do gradiente, que geralmente é quantizada em ângulos discretos como 0°, 45°, 90° e 135°.

Para cada célula, é calculado um histograma que representa a distribuição das orientações dos gradientes, evidenciando as bordas presentes. Para minimizar os efeitos de variações na iluminação, os histogramas são normalizados, dividindo-se cada histograma pela norma do conjunto de histogramas da imagem.

Por fim, os histogramas normalizados de todas as células são concatenados em um único vetor de características que representa o objeto contido na imagem.

* LBP(Padrão Binário Local):

O padrão local binário realiza a caracterização ao comparar cada pixel com seus vizinhos para codificar informações de textura em padrões binários. O resultado forma números binários de 8 bits para cada pixel, que são transformados em um histograma que representa as frequências dos padrões. Esse histograma funciona como um vetor de características usado para a classificação de imagens. SEDAGHATJOO et al. (2024)

Na Figura 2.9 é possível observar como esse algoritmo funciona.

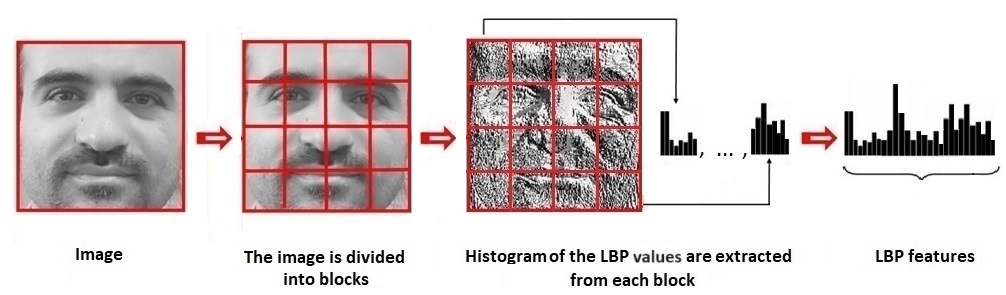


Figure 2.9: Processo de Funcionamento do LBP.

Fonte: SEDAGHATJOO et al. (2024)

* + 1. Classificação

Após extrair as características é necessário identifica-las, ou seja, realizar a

classificação; nesta etapa utiliza-se de algoritmos de aprendizado de máquina e redes neurais para categorizar as expressões faciais nas respectivas emoções.

Os métodos mais populares são:

* Redes neurais:

As redes neurais artificiais são capazes de aprender representações complexas a partir dos dados por meio de um processo de treinamento supervisionado ou não supervisionado (Goodfellow et al. 2016).

Dentre os tipos mais conhecidos, destacam-se as Redes Neurais Convolucionais (CNNs), utilizadas principalmente para reconhecimento de padrões espaciais em imagens), e as Redes Neurais Recorrentes (RNNs), voltadas para o processamento de sequências temporais, como vídeos.

.

* Máquinas de vetores de suporte( (SVM: Support Vector Machine):

O SVM é um algoritmo de aprendizado supervisionado muito eficaz para classificação e análise de padrões. Ele busca encontrar o hiperplano que melhor separa as classes de dados, maximizando a margem entre elas (Cortes & Vapnik, 1995)

* + 1. Treinamento do Classificador

Nesta fase, o classificador (independentemente do tipo escolhido) é alimentado com um conjunto de dados rotulado, ou seja, cada amostra de entrada está associada à sua respectiva classe. O treinamento consiste em ajustar os parâmetros do modelo para minimizar o erro entre a predição realizada e o rótulo verdadeiro.

Durante o treinamento, o conjunto de dados é dividido em subconjuntos de treino e validação, o primeiro é usado para ajustar o modelo e o segundo para monitorar o desempenho e evitar que o modelo se adapte demais e pare de analisar.

Após o treinamento, o classificador é avaliado com um conjunto de teste separado, verificando assim sua capacidade.

## Aprendizado de Máquina e Aprendizado Profundo

O Aprendizado de Máquina (ML: *Machine Learning*) é a utilização de um extenso conjunto de dados para melhorar a capacidade de análise de dados, realizando a identificação de padrões, para assim tomar uma decisão (Goodfellow et al. 2016).

O reconhecimento de padrões por meio do aprendizado de máquina consolidou-se como uma das técnicas mais avançadas na atualidade, uma vez que, segundo Prateek (2017), essa abordagem permite a realização de previsões com base em dados que seriam desconhecidos e até mesmo imperceptíveis ao ser humano.

Com o aumento constante da quantidade de dados disponíveis, surgiu o Aprendizado Profundo (Deep Learning), que possui a capacidade de processar grandes volumes de dados e se beneficia do poder e da velocidade computacionais das máquinas modernas (Cîrneanu et al., 2023), promovendo uma evolução exponencial da inteligência artificial e potencializando todos os seus campos de aplicação.

Dentre as técnicas mais importantes do ML e do DL estão as Redes Neurais Artificiais (ANN), que de acordo com Alves (2020), baseiam-se na arquitetura dos neurônios humanos, buscando assim reproduzir o aprendizado através do desenvolvimento de sistemas que aprendem ao serem treinadas com uma base de dados de treinamento.

Em sua estrutura é possível ver arranjos em nós, semelhantes a corpos celulares humanos, que se articulam com outros nós por meio de conexões programadas, ponderadas com base em sua capacidade de fornecer um resultado esperado(Goodfellow et al. 2016).

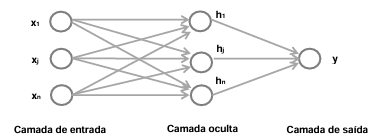


Figura 2.10. Diagrama de uma rede neural básica (Shanmugnami, 2018)

Existem diversos tipos de Redes Neurais Artificial com propostas e tamanhos diversos, variando até mesmo o número de neurónios na camada de entrada e na camada oculta, mas veremos com mais detalhes de duas delas: a Rede Neural Convolucional (CNN) e a Rede Neural Recorrente (RNN).

* + 1. **Rede Neural Convolucional**

A Rede Neural Convolucional (CNN) é uma arquitetura de rede neural artificial que incorpora operações de convoluções em diferentes camadas da rede. Esta rede é amplamente utilizada para tarefas de processamento de imagens e reconhecimento de padrões visuais (Awari, 2023).

As CNNs são projetadas para capturar padrões espaciais dentro de matrizes de entrada multidimensionais (Cîrneanu et al., 2023). Para isso, elas são treinadas através do uso de um conjunto de dados rotulados, onde cada entrada (por exemplo, uma imagem) está associada a uma saída desejada (por exemplo, a classe do objeto na imagem), assim ela é denominada uma rede *Feedforward*, ou seja, a entrada que considera apenas sua entrada atual, as entradas anteriores não são úteis durante o processo, não tendo assim noção de ordem no tempo.

Uma das camadas mais importantes de uma CNN é a camada de convolução, onde após transformar a imagem de entrada em uma matriz de valores para cada pixel, são aplicados filtros (ou kernels) para detectar padrões como cantos, bordas ou texturas, resultando em mapas de características. Esses mapas de características são passados por uma camada de ativação ReLU para introduzir não linearidade e aprender padrões mais complexos, facilitando assim a extração de características.

A Figura 2.9 ilustra uma rede CNN com um número n de camadas de convolução para extrair características locais, seguidas de funções de ativação como ReLU para introduzir não linearidade; após isso passa para camada de pooling para reduzir a dimensão dos mapas de características, tornando o processamento mais eficiente

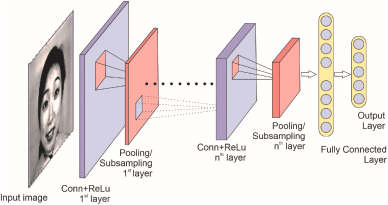


Figura 2.9: Camadas de uma Rede Neural Convolucional. Fonte: (Anil e Padma Suresh, 2023)

Depois, camadas totalmente conectadas geram pontuações para cada classe, convertidas em probabilidades, gerando assim o resultado da classificação como saída.

* + 1. **Rede Neural Recorrente**

De acordo com Hanafi et al. (2021), uma Rede Neural Recorrente (RNN) é um tipo de rede neural projetada para lidar com dados sequenciais, como séries temporais, textos, imagens e vídeos, em que a ordem dos elementos é importante. Isto é possível pois as RNNs possuem conexões recorrentes, ou seja, a saída de uma célula da rede pode ser reutilizada como entrada para a mesma célula em passos subsequentes.

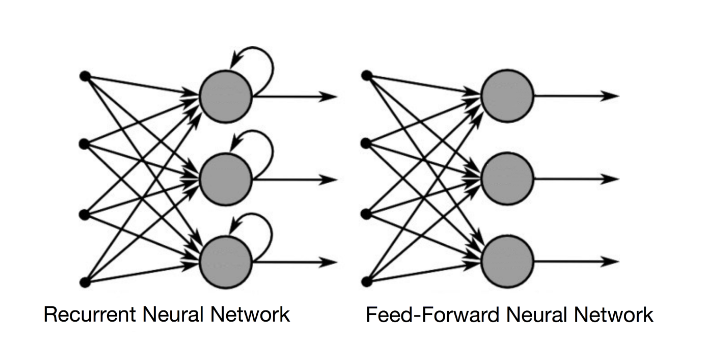


Figure 2.10: Redes Neurais Recorrentes e Feed-Forward Fonte: (Hanafi et al, 2021)

Um dos problemas das RNNs é que sua memória é rápida e curta, dessa forma ela lembra apenas o que aconteceu nas ultimas interações; assim surge a Memória de curto longo prazo, conhecida como LSTM (*Long Short Term Memory*) (Rangulov;Fahum, 2020).

A LSTM é um tipo de rede neural recorrente desenvolvida para solucionar o problema das dependências de longo prazo, visto que consegue "lembrar" de informações por longos períodos e processar até mesmo sequências de imagens (Cîrneanu et al., 2023).

Além da LSTM, existem outros tipos de Redes Neurais Recorrentes como as Redes neurais recorrentes bidirecionais (BRRNs), as Unidades recorrentes fechadas (GNUs), as RNN com codificador-decodificador e etc. (Hanafi et al., 2021)

* + 1. **Transferência de Aprendizado (Transfer Learning)**

Uma das técnicas de aprendizado de máquina muito utilizada para lidar com uma baixa quantidade de dados é a transferência de aprendizado (Cîrneanu et al., 2023).

Assim, se utiliza o conhecimento adquirido de uma rede já treinada com outros dados, aplicando-os em uma nova tarefa, melhorando assim a generalização, dessa forma as redes são treinadas não apenas para uma tarefa específica, mas de uma forma que possa ser aplicada em diferentes tarefas (Soria Olivas et al., 2009).

Na figura 2.12 podemos observar a diferença entre métodos de aprendizado tradicionais, criados para tarefas específicas, e métodos de transferência de aprendizado, que podem ser aplicados a diferentes tarefas.

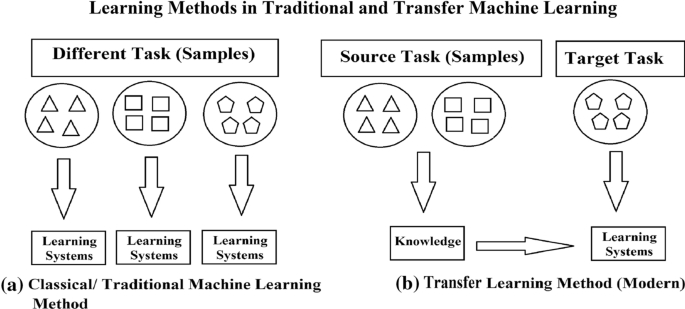


Figure 2.12: Aprendizado de Máquina e Aprendizado por transferência.

Fonte: (SORIA OLIVAS et al., 2009).

Além de eliminar a necessidade de grandes conjuntos de dados de treinamento, a transferência de aprendizado acaba reduzindo o custo computacional ao reutilizar um modelo já existente para resolver um novo problema, aproveitando características e pesos já aprendidos por modelos robustos, treinados geralmente com grandes datasets.

O uso de redes neurais convolucionais pré-treinadas têm sido muito explorados nos últimos anos, pois dessa forma as redes já aprenderam a identificar padrões gerais em imagens (Cîrneanu et al., 2023).

Exemplos notórios de arquiteturas frequentemente utilizadas para transfer learning incluem a VGG-16, a ConvNet e a MobileNetV2 (Ballesteros et al. 2024).

## Trabalhos relacionados

O reconhecimento facial de emoções é um campo de estudo que tem recebido crescente atenção devido às suas amplas aplicações em áreas como interação humano-computador, segurança e bem-estar emocional.

No artigo Ekman; Friesen (1976), os autores desenvolvem o Sistema de Codificação da Ação Facial (Facial Action Coding System - FACS), um método detalhado para descrever movimentos faciais específicos associados às expressões emocionais. A pesquisa de Ekman e Friesen foi fundamental para estabelecer uma base científica rigorosa para a interpretação das expressões faciais, ajudando a identificar padrões universais de expressão emocional que são reconhecidos globalmente

A introdução das Redes Neurais Convolucionais (CNNs) trouxe um avanço significativo na precisão do reconhecimento facial de emoções. Tang (2013), utilizou CNNs para reconhecimento de emoções em imagens faciais, e demonstrou a superioridade dessas redes em benchmarks como o FER-2013.

Em Anil e Padma Suresh (2023) é desenvolvido uma metodologia híbrida para reconhecimento facial, que combina o HOG e a CNN para extração e refinamento de características, seguido por classificação com um classificador não iterativo como o KELM. Essa técnica, denominada HOG-CKELM, alia o rápido tempo de treinamento do KELM à capacidade da CNN de extrair características profundas e invariantes, oferecendo alta precisão e rapidez quando comparada à abordagem HOG-CNN tradicional.

O artigo Cîrneanu et al. (2023) aborda uma visão abrangente das tendências recentes no reconhecimento de emoções utilizando redes neurais e análise de imagens. O artigo destaca a importância crescente desse campo em diversas aplicações, como educação, saúde e segurança pública. Eles comparam CNNs com outras arquiteturas, como redes neurais recorrentes (RNNs) e redes adversárias generativas (GANs), destacando os elementos-chave, desempenho, vantagens e limitações de cada modelo, além de concluir que o uso de aprendizado por transferência e o desenvolvimento de arquiteturas mais eficientes possam melhorar acurácia.

O artigo Ballesteros et al. (2024) tem como objetivo a criação de um software para reconhecer a emoção facial expressada, para isso utiliza técnicas de visão computacional e de inteligência artificial. O artigo utiliza um método de detecção facial para identificar a região de interesse e após isso, utiliza 2 redes neurais convolucionais, uma rede pré treinada, utilizada através do transfer learning, para a extração de características e uma nova rede neural convolucional criada para a classificação.

Em Huang et al. (2023), técnicas de visão computacional foram utilizadas para identificar pontos de referência faciais importantes para a detecção da emoção facial. Com esse propósito, foram feitos experimentos de validação com duas redes neurais já criadas, fazendo o uso do tranfer learning, destacando assim a importância de usar a aprendizagem por transferência para melhorar o desempenho dos algoritmos.

O estudo realizado por Sarvakar et al.(2023) propõe o desenvolvimento e implementação de um sistema de reconhecimento facial com base em redes neurais convolucionais (CNN). A escolha da arquitetura correta permitiu obter uma boa acurácia na classificação de rostos. Os autores apontam que ajustes, como a redução da taxa de aprendizado e a simplificação da arquitetura, podem aumentar a acurácia e reduzir o custo computacional e o tempo de treinamento.

Em Rangulov;Fahum (2020) é proposta uma abordagem híbrida para reconhecimento de emoções em vídeos, combinando redes convolucionais (CNNs) como extratoras de características e redes neurais recorrentes (RNNs) para modelar a dinâmica temporal. O método foi testado em um grande conjunto de dados de vídeo, com resultados promissores, superando abordagens tradicionais que consideram apenas quadros estáticos. A combinação do extrator convolucional com a RNN permite capturar tanto o conteúdo visual quanto a evolução temporal, essencial para o reconhecimento robusto de emoções.

# CAPÍTULO 3 - Sistema de Reconhecimento Facial de Emoções

O sistema de reconhecimento facial de emoções proposto busca analisar sequências de expressões faciais através de vídeos, visto que os vídeos na verdade são sequências de imagens(frames). Cada frame do vídeo deve ser analisado para detectar expressões faciais, ou seja, após a coleta, cada frame da sequência de expressão facial da emoção passa para as etapas de modo individual.

Assim, é implementado um pipeline completo que analisa os vídeos, processa cinco frames selecionados e detecta a emoção presente em cada um deles. Este pipeline integra múltiplas técnicas de visão computacional e aprendizado profundo, organizadas em quatro etapas principais: pré-processamento, extração de características, classificação e análise de resultados.

Dessa forma, a modelagem proposta na presente pesquisa é definida e detalhada na Figura 3, contendo assim as etapas necessárias e as técnicas utilizadas.

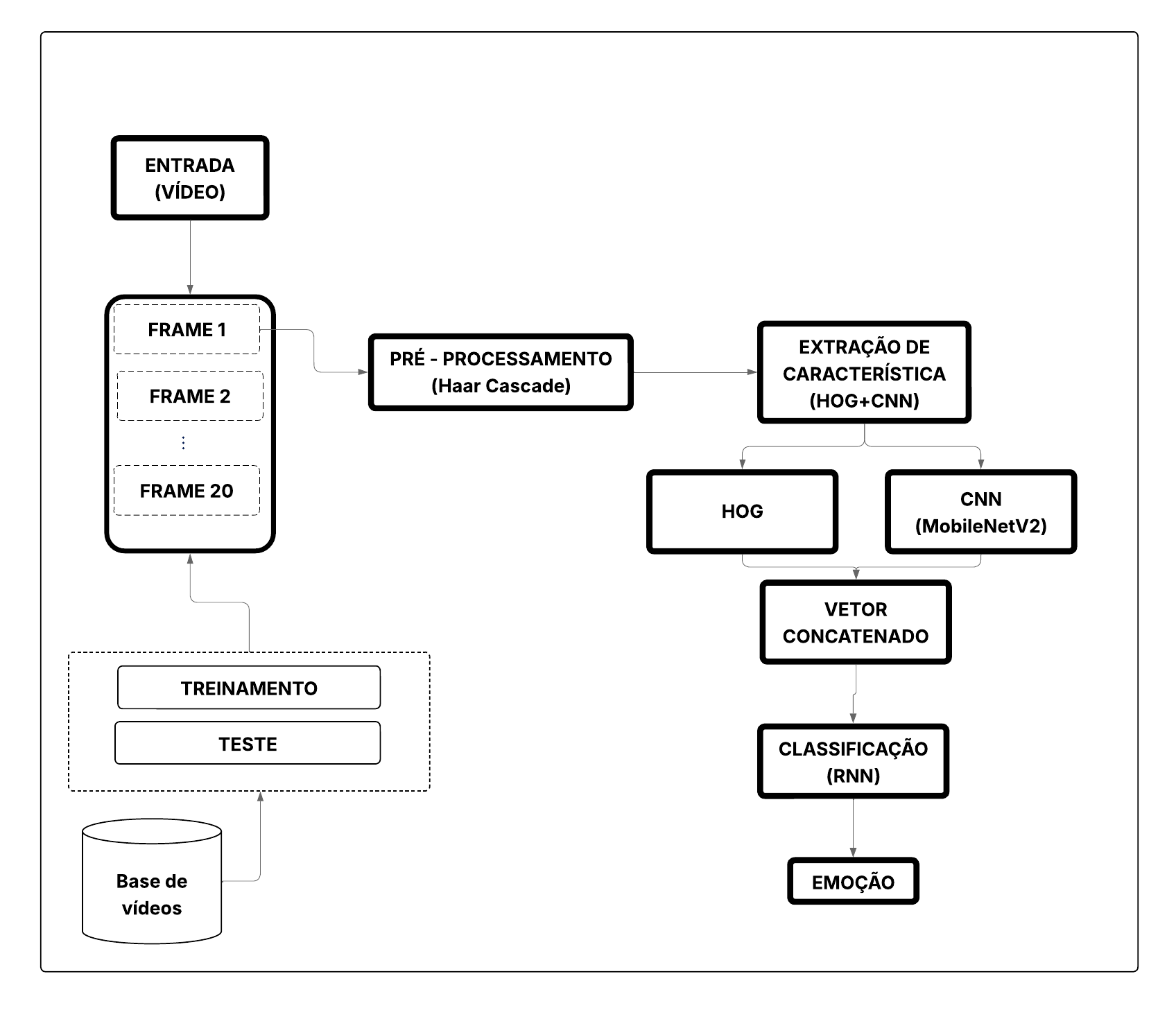


Figure 3:Modelagem Proposta

Fonte: Imagem criada pela autora.

## 3.1 Captura da Imagem

As imagens/frames serão utilizadas em três conjuntos nesse sistema: para treinamento, validação e teste.

Para isso, utiliza-se um banco de dados de vídeo faciais, o KDEF-dyn (Karolinska Directed Emotional Faces – dynamic version), composto por 240 vídeos que retratam atores expressando seis emoções(alegria, tristeza, raiva, medo, nojo e surpresa) de forma dinâmica, o que é fundamental para capturar as transições e nuances temporais das expressões faciais, demonstrando assim a evolução de um estado neutro para uma das emoções básica em vídeos curtos, com duração média de 1,033ms cada.

Cada um desses vídeos será dividido em 20 frames, que serão analisados individualmente, respeitando a sequência.

Desta forma, esses dados serão úteis para gerar o conhecimento para o classificador, que será aplicado posteriormente em vídeos gravados de forma particular com o intuito de testar o bom funcionamento do modelo proposto.

Os vídeos inseridos também serão divididos em 20 frames, que serão enviados para a próxima etapa de forma sequencial.

## Pré-Processamento das Imagens

No pré-processamento são aplicadas técnicas específicas para otimizar a qualidade e a eficiência do processamento dos frames de vídeo. Estas técnicas são fundamentais para garantir a precisão e o desempenho do sistema de reconhecimento facial.

Os frames são redimensionados para um formato padronizado de 48×48 pixels, garantindo consistência dimensional para o modelo e reduzindo significativamente a carga computacional.

A transformação de imagens coloridas (RGB) para escala de cinza reduz a dimensionalidade dos dados preservando as informações essenciais de iluminação que são críticas para a detecção de características faciais.

Além disso, é necessário utilizar algoritmos para identificar a região de interesse, um desses algoritmos é o Haar Cascade.

3.2.1. Haar Cascade

O Haar Cascade é um método que visa detectar objetos ou regiões de interesse, através do aprendizado de máquina (Viola;Jones, 2001). Assim, o algoritmo é treinado a partir de imagens que contém o objeto a ser identificado e também com imagens que não possuem o objeto.

O seu funcionamento consiste em 3 fases:

* Características Haar:

São realizados cálculos em regiões retangulares adjacentes em um local específico da imagem, visando buscar uma mudança de intensidade. Dessa forma, as características de haar são calculadas através da diferença da soma de pixel sob o retângulo branco e sob o retângulo preto(Ballsteros et al. 2024).

*Valor da característica = Σ(pixels na região branca) - Σ(pixels na região preta)*

Estas características são particularmente eficazes para detectar bordas, linhas e estruturas centro-circundantes que são comuns em faces humanas (Lienhart & Maydt, 2002). A Figura 3.1 ilustra os principais tipos de características Haar.

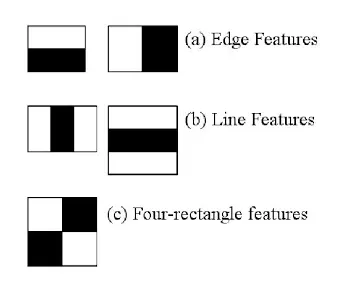


Figure 3.1: Tipos de características de Haar

Fonte: ResearchGate.

Capturando assim características essenciais para a detecção da região de interesse, visto que a região dos olhos geralmente é mais escura que as bochechas, ou a ponte nasal é mais clara que os olhos adjacentes (Ballesteros et al., 2024).

* Imagem Integral:

O processo de calcular cada retângulo pode ser longo, assim utiliza-se uma imagem intermediária para diminuir o tempo de processamento. Essa imagem intermediária é na verdade a imagem integral, dividida normalmente como na Figura abaixo (Ballsteros et al. 2024).

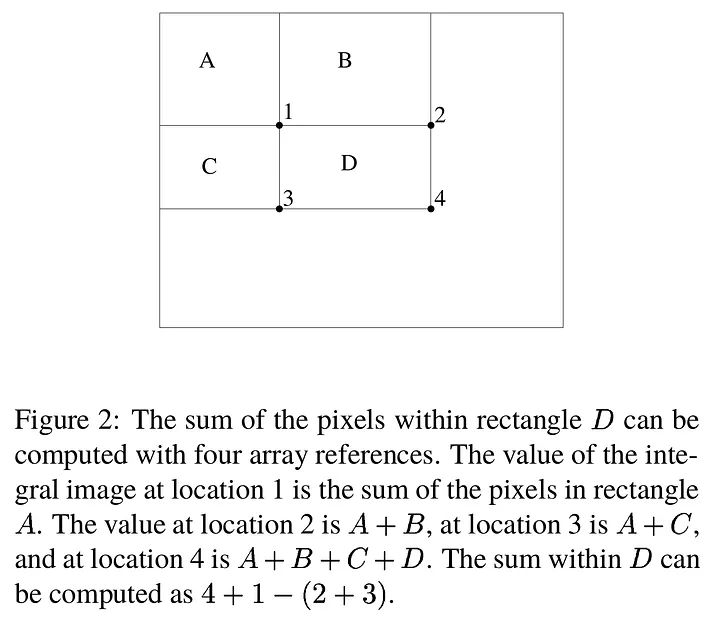


Figure 4: Integral da imagem

Fonte: (Viola;Jones, 2001)

Dessa forma, a imagem integral na posição (x,y) é a soma dos pixels acima e à esquerda de (x,y), como podemos ver na Equação (1):

(1)

Onde:

é a integral da coordenada,

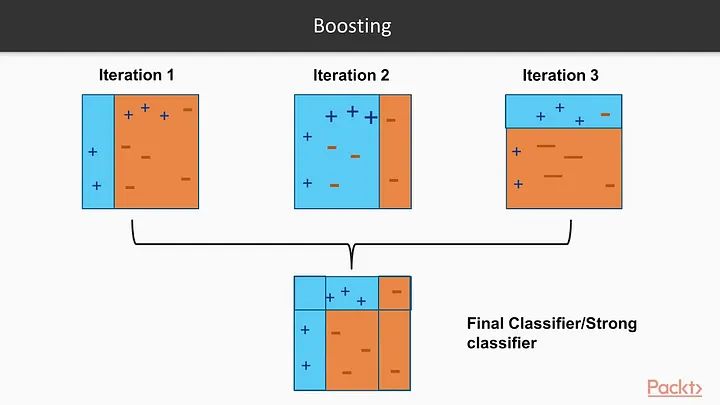
é imagem original.

Reduzindo assim o grau de complexidade ao lidar com as imagens, calculando a soma dos pixels em qualquer região retangular usando apenas quatro referências na imagem integral.

* Treinamento Adaboost:

No treinamento Adaboost, apenas os melhores recursos obtidos são escolhidos e direcionados para diversos classificadores que são treinados separadamente, focando assim em corrigir os erros do classificador anterior (Ballsteros et al. 2024).

A cada iteração, o algoritmo escolhe o classificador e as características com menor taxa de erro. Quando um classificador identifica uma região da imagem, essa região é passada para o próximo, e um novo classificador, focado em outra característica, é escolhido. Cada classificador se concentra em uma única característica, e o processo se repete a cada iteração, criando assim um classificador final mais forte, como podemos ver na Figura abaixo.



Além disso, o método combina a aplicação dessas características com um classificador em cascata e o uso de uma janela deslizante para identificar rostos de forma eficiente (VIOLA; JONES, 2001; RANGULOV; FAHIM, 2020).  
  
A biblioteca OpenCV disponibiliza classificadores Haar Cascade pré-treinados para diversas aplicações, incluindo detecção facial, basta fornecer o caminho do arquivo e chamar a função correspondente para utilizá-los.

## Extração de Características

A extração de características é a etapa onde apenas informações relevantes são extraídas das imagens faciais e tudo o que não serve para a análise é descartado.

No modelo proposto, utilizaremos a concatenação de dois recursos de extração: um extrator HOG e a CNN pré-treinada, a MobileNetV2.

### 3.3.1. Histograma de Gradiente Orientado (HOG)

Uma técnica de extração de características muito conhecida é o histograma de gradientes orientados (HOG), técnica que calcula os gradientes locais de intensidade da imagem e gera histogramas que descrevem as direções predominantes das bordas, criando assim uma representação robusta das estruturas faciais (Anil e Padma Suresh, 2023).

Segundo Shu et. al (2011), o processo de extração do HOG pode ser dividido nas seguintes etapas:

* Cálculo do gradiente:

São calculados os gradientes horizontal e vertical de cada pixel da imagem através da aplicação de máscaras como o operador Sobel, muito utilizado para detecção de bordas em imagens ao identificar as variações na direção horizontal(2) e na direção vertical(3) de forma individual.

**[ ]** (2)

**[ ](**3)

Após obter os gradientes horizontal e vertical, podemos calcular a magnitude e a orientação do gradiente de cada pixel.

A magnitude de cada gradiente é calculada através da raiz quadrada da soma dos quadrados do gradiente horizontal e do gradiente vertical através da expressão (4) abaixo:

(4)

A orientação do gradiente, ou seja, o seu ângulo, é calculada através do arco tangente do gradiente vertical ao gradiente horizontal, como podemos ver na fórmula (5).

(5)

* Divisão em Células:

A imagem é dividida em pequenas regiões espaciais conectadas, denominadas "células". Para cada célula, constrói-se um histograma das orientações dos gradientes. As orientações dos pixels dentro da célula calculadas anteriormente são quantizadas em um número pré-definido de "bins", ou seja, são definidos em categorias angulares. Cada pixel contribui para um bin do histograma com um peso proporcional à magnitude do seu gradiente.

Dessa forma, a distribuição das orientações das bordas dentro da célula é obtida a partir do histograma.

* Agrupamento em Blocos e Normalização:

Para aumentar a robustez a variações locais de iluminação e contraste, as células são agrupadas espacialmente em blocos maiores e sobrepostos. Um vetor de características é formado pela concatenação dos histogramas de todas as células dentro de um bloco. Este vetor concatenado é então normalizado por exemplo, usando a norma L2, equação (6).

​. (6)

* Formação do Vetor de Características HOG:

Finalmente, os vetores de características normalizados de todos os blocos da imagem são concatenados para formar o descritor HOG final, que representa a distribuição das orientações dos gradientes na região analisada.

Este descritor HOG resultante pode então ser utilizado como entrada para algoritmos de aprendizado de máquina para tarefas de classificação e detecção de objetos.

### 3.3.2. MobileNetV2

A MobileNetV2 é uma rede neural Convolucional (CNN) desenvolvida por pesquisadores da Google como uma nova versão da MobileNet, criada para lidar com aplicações de visão computacional em dispositivos com poucos recursos de memória e processamento, além de sua agilidade(Sharma, 2023). Ela se destaca por seu equilíbrio entre eficiência computacional e acurácia na classificação de imagens.

Sua estrutura possui as camadas descritas na Figura abaixo:

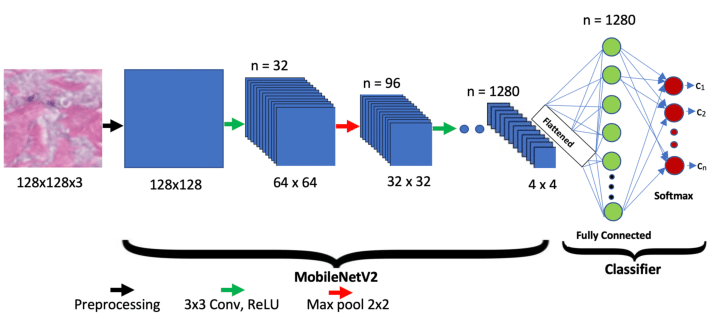


Figura: Fonte (Sharma, 2023)

Entre seus principais componentes estão as convoluções separáveis em profundidade, que dividem a operação padrão em duas etapas, a convolução em profundidade e a convolução pontual, diminuindo significativamente os cálculos. A estrutura também utiliza resíduos invertidos e projetos de gargalo, que expandem e comprimem os canais de forma estratégica, permitindo ao modelo extrair características mais complexas com menos parâmetros (SANDLER et al., 2019).

Outro diferencial é o uso de gargalos lineares, que evitam a perda de informações ao empregar ativações lineares ao final dos blocos, e dos blocos SE (Squeeze-and-Excitation), que recalibram os canais para destacar os recursos mais relevantes. Esses elementos tornam o MobileNetV2 altamente eficaz para tarefas de visão computacional, mantendo um equilíbrio entre desempenho e leveza (SANDLER et al., 2019).

Após essa camada Convolucional, ocorre uma camada de Max Pooling para agregar as características espaciais. Depois de passar por várias sequencias de Convolução + Max Pool, é gerado um vetor de característica que é então direcionado para a camada de classificação.

Esse modelo é muito utilizado em diversas aplicações através do aprendizado por transferência (transfer learning). Assim, os pesos pré-treinados com o conjunto de dados ImageNet são utilizados no novo modelo, aproveitando assim os conhecimentos obtidos durante o processamento do dataset original. A camada de classificação original é descartada, dessa forma será possível obter apenas o vetor de características gerado.

É realizado um ajuste para que o modelo se adeque a nova tarefa, adaptando as camadas finais ou novamente treinando de maneira progressiva mais camadas para otimizar o desempenho no dataset alvo (Sharma, 2023).

## 3.4. Classificação

O vetor concatenado de características gerado pela extração do HOG e da MobileV2, será direcionado para a camada de entrada de uma Rede Neural Recorrente, mais especificadamente uma LSTM.

A Long Short-Term Memory(LSTM), em português Memória de curto longo prazo, é uma rede neural recorrente capaz de manter e atualizar informações por um período de tempo.

Sua estrutura também é composta por uma camada de entrada, diversas camadas ocultas e uma camada de saída, porém em suas camadas ocultas estão diversas unidades de memória, também conhecidas como células.

Essas células são compostas por três portas principais: a porta de entrada(Input Gate), a porta de esquecimento(Forget Gate) e a porta de saída(Output Gate). São essas portas que controlam o fluxo de informações dentro da célula de memória,

A estrutura de uma célula da LSTM possui a definição demonstrada na Figura 2.11.

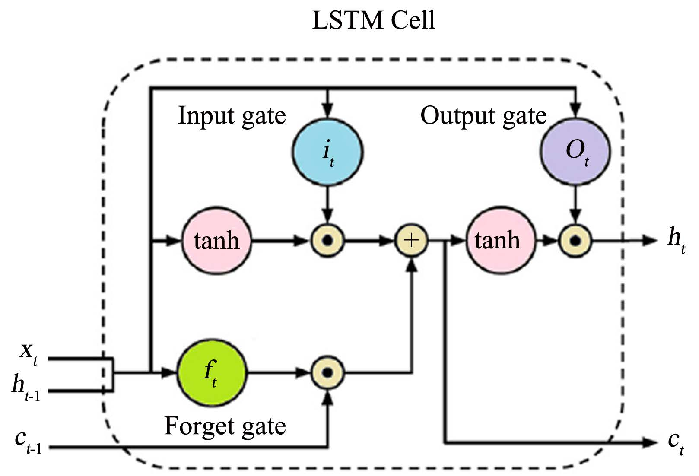


Figure 2.11 Estrutura de uma LSTM

Fonte:(Basiri et al., 2021)

Essa estrutura complexa de portões permite que a LSTM controle o fluxo de informações dentro da célula de memória, aprendendo quais informações são importantes para manter, quais devem ser esquecidas e quais devem ser usadas para fazer previsões ou classificações em cada ponto da sequência. Isso a torna particularmente eficaz em tarefas onde o contexto de longo prazo é essencial (Basiri et al., 2021).

Dessa forma a LSTM permite que o sistema identifique não apenas os traços faciais de um único frame, mas também como eles evoluem ao longo do tempo, o que é crucial para a detecção de emoções em vídeos.

## 3.5. Ferramentas e Ambiente de Desenvolvimento

O desenvolvimento do sistema foi realizado na linguagem **Python**, utilizando principalmente as bibliotecas **TensorFlow** e **Keras** para a implementação das redes neurais, e as bibliotecas **Numpy e** OpenCV para o pré-processamento das imagens. O treinamento do modelo foi executado em uma máquina com GPU para acelerar o processo.

# CAPÍTULO 4 - IMPLEMENTAÇÃO E RESULTADOS

Este capítulo descreve em detalhes a implementação do sistema de reconhecimento de emoções faciais em vídeo proposto neste trabalho, bem como os resultados experimentais obtidos.

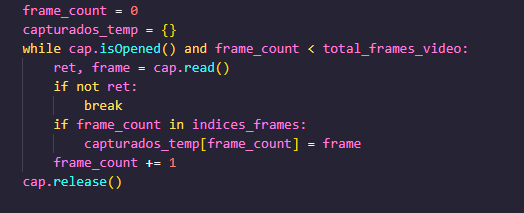
As etapas incluem a aquisição e preparação dos dados, o pré-processamento das imagens e extração de características, o processo de treinamento e, por fim, a avaliação e análise dos resultados alcançados.

## Captura das Vídeos

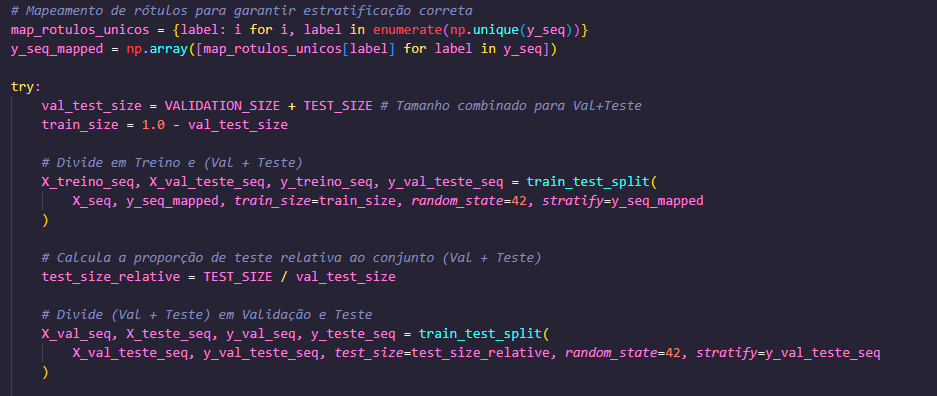
A etapa inicial da implementação consistiu na obtenção e preparação do conjunto de dados para treinar e avaliar o modelo de reconhecimento de emoções. Para este propósito, utilizou-se o banco de dados KDEF-dyn (Karolinska Directed Emotional Faces – dynamic version) como fonte.

Este dataset é composto por 240 vídeos que retratam atores expressando seis emoções de forma dinâmica, o que é fundamental para capturar as transições e nuances temporais das expressões faciais, demonstrando assim a evolução de um estado neutro para uma das emoções básicas

Cada um desses vídeos passou por um processo (Código 1) para captação de 20 frames de forma espaçada, tornando assim os 240 vídeos em 240 sequências, onde cada sequência era composta por 20 frames.



O conjunto de dados foi dividido (Código 2) em 60% para o treinamento(144 sequências), 20% para a validação(48 sequências) e 20% para o teste(48 sequências), garantindo uma separação adequada entre as fases de aprendizado e validação do modelo, buscando evitar o overfitting - quando o modelo se adapta demais aos dados que viu e não consegue lidar com novas informações que não estavam no conjunto de treinamento.

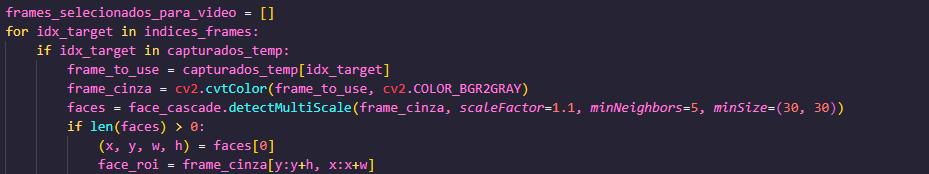
)

O parâmetro *stratify=rotulos\_videos* garante que 20% das sequências de cada uma das emoções irá para o conjunto de teste e de validação, evitando assim que uma emoção fique desbalanceada no conjunto de teste.

## Pré-Processamento

Como dito antes, para cada vídeo do dataset, foram extraídos uma sequência fixa de 20 frames por vídeo. Desta forma, é necessário que cada frame extraído passe por um pré-processamento básico essencial: conversão para escala de cinza, eliminando informações de cor e focando nas variações de intensidade, e o redimensionado para um tamanho padrão de 48x48 pixels.

Esse processo é descrito no Código 3.



Além disso, foi necessário utilizar o Haar Cascade, para realizar a detecção facial determinando a região de interesse.

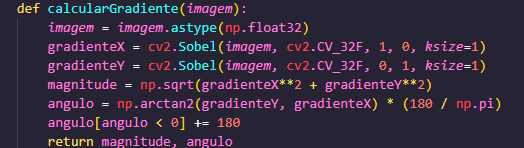
Essa padronização é necessária para garantir que todas as entradas para os extratores de características tenham dimensões consistentes.

## Extração de Características

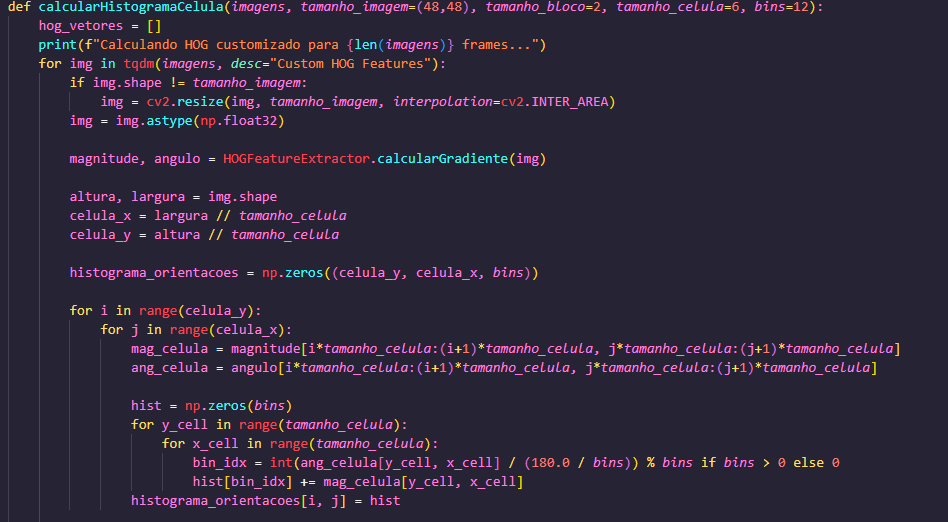
A etapa seguinte foi a extração de características. O sistema implementado empregou uma abordagem híbrida, combinando dois métodos para capturar diferentes níveis de informação da imagem: o Histograma de Gradientes Orientados (HOG) e uma Rede Neural Convolucional (CNN) pré-treinada (MobileNetV2).

* + 1. **Histograma de Gradientes Orientados (HOG)**

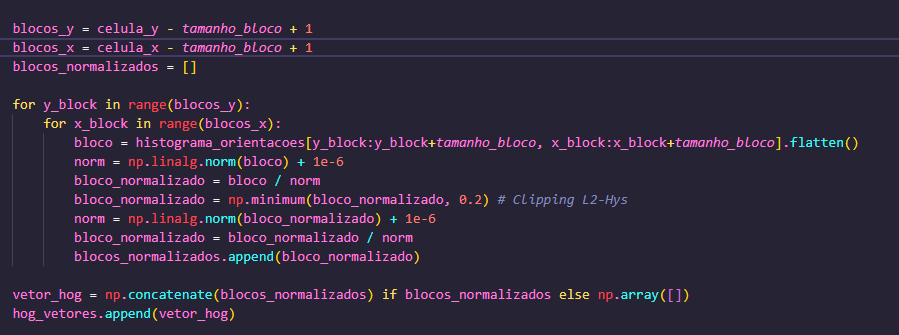
A extração de características HOG foi realizada para cada frame já pré-processado, o processo(Código 4) se inicia com o cálculo dos gradientes de intensidade nas direções horizontal e vertical utilizando o operador Sobel e a partir destes, a magnitude e a orientação do gradiente eram computadas para cada pixel.



A imagem era então dividida em células de 6x6 pixels. Dentro de cada célula, um histograma das orientações dos gradientes (distribuídas em 12 bins) era construído, ponderado pelas magnitudes dos gradientes, processo descrito no Código 5.



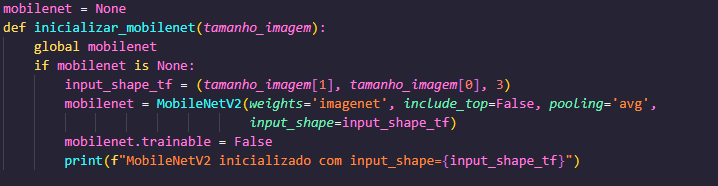
Subsequentemente (Código 6), os blocos de 2x2 células eram formados, deslizando sobre a imagem com sobreposição. Os histogramas das células dentro de cada bloco eram concatenados e normalizados utilizando a técnica L2-Hys (normalização L2, seguida de clipping a 0.2 e re-normalização L2) para conferir robustez a variações de iluminação.



Finalmente, os vetores normalizados de todos os blocos eram concatenados para formar o descritor HOG final para aquele frame. Este descritor captura informações sobre as formas e texturas locais presentes na expressão facial.

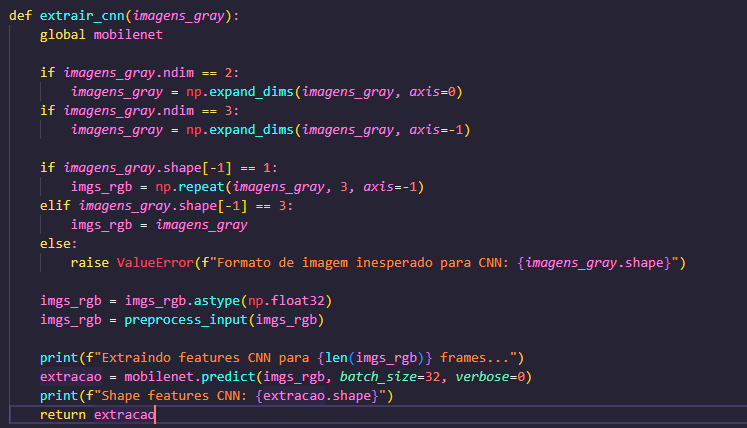
* + 1. **Características via CNN (MobileNetV2)**

Paralelamente à extração HOG, características de mais alto nível foram extraídas utilizando a arquitetura MobileNetV2, pré-treinada na vasta base de dados ImageNet (Código 7).



A função inicializar\_mobilenet foi responsável por carregar o modelo MobileNetV2 através da biblioteca Keras, configurando-o para atuar como um extrator de características: a camada de classificação final foi removida (include\_top=False), uma camada de Global Average Pooling foi adicionada (pooling='avg') para obter um vetor de características de tamanho fixo, e os pesos das camadas convolucionais foram congelados (mobilenet.trainable = False), caracterizando assim que o modelo será utilizado apenas através do transfer learning.

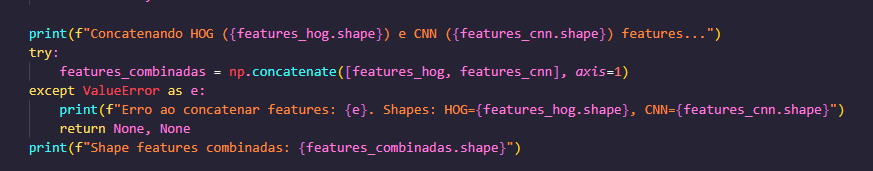
A função extrair\_cnn (Código 8) processava os frames em escala de cinza, convertendo-os para um formato RGB ao replicar o canal de cinza três vezes e aplicando o pré-processamento específico da MobileNetV2 (preprocess\_input).



Os frames processados eram então alimentados ao modelo mobilenet, e os vetores de características resultantes da camada de pooling eram coletados.

* + 1. **Combinação de Características**

A etapa final da extração consistiu em combinar as informações obtidas pelos dois métodos (Código 9). Para cada frame individual, o vetor de características HOG e o vetor de características CNN foram concatenados.

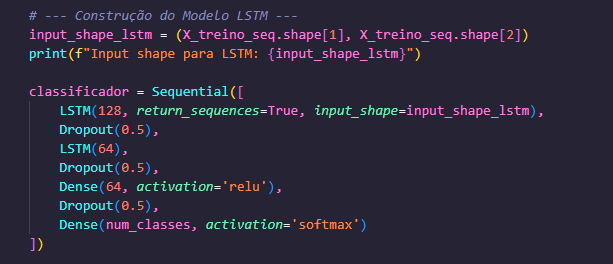


Isso resultou em um único vetor de características combinado por frame, que integra tanto as informações de textura e forma local do HOG quanto as representações da CNN, fornecendo uma descrição mais rica e robusta da expressão facial naquele instante para a modelagem temporal subsequente.

* 1. **Estrutura do Modelo Sequencial (LSTM)**

Para analisar as sequências temporais das características combinadas (HOG+CNN) extraídas dos frames e classificar a emoção expressa no vídeo, foi implementada uma Rede Neural Recorrente (RNN) do tipo Long Short-Term Memory (LSTM).

A rede foi construída (Código 10) como uma pilha de camadas. A entrada para o modelo consiste nas sequências de características combinadas, ou seja, é a dimensão do vetor resultante da concatenação das características HOG e CNN.



A primeira camada foi uma LSTM com 128 unidades, configurada com *return\_sequences=True*, significando que esta camada processa a sequência de entrada e produz uma sequência de saída de mesmo comprimento, permitindo que a informação temporal seja passada para a camada seguinte. Seguiu-se uma camada Dropout com taxa de 0.5, uma técnica de regularização que desativa aleatoriamente metade das conexões durante o treinamento para mitigar o risco de overfitting.

A segunda camada LSTM possuía 64 unidades e foi configurada com *return\_sequences=False,* fazendo com que a camada processe a sequência de entrada vinda da camada anterior, mas retorne apenas o vetor de saída correspondente ao último passo temporal. Este vetor final encapsula a informação relevante de toda a sequência processada. Outra camada Dropout com taxa de 0.5 foi adicionada após a segunda LSTM.

Após as camadas recorrentes, uma camada Densa (totalmente conectada) com 364 neurônios e função de ativação ReLU foi incluída para processar a representação final da sequência.

Finalmente, a camada de saída foi uma camada Densa com um número de neurônios igual ao número de classes de emoção identificadas no dataset, ou seja, 6 neurônios.

A função de ativação softmax foi utilizada nesta última camada para produzir uma distribuição de probabilidade sobre as classes de emoção, onde a classe com a maior probabilidade é selecionada como a predição do modelo para a sequência de vídeo.

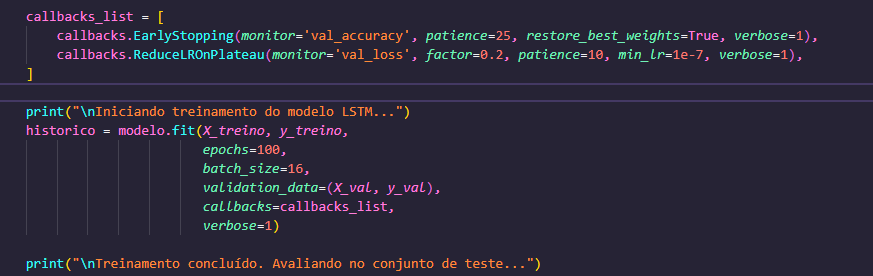
**4.4.1 Treino**

O processo de treinamento (Código 11) do modelo LSTM foi configurado e executado utilizando as funcionalidades da biblioteca Keras. Antes do treinamento, o modelo foi compilado especificando-se os componentes essenciais para o aprendizado.



O otimizador escolhido foi o adam, um algoritmo de otimização estocástica baseado em gradiente amplamente utilizado e eficaz para problemas de aprendizado profundo.

O treinamento foi realizado através da chamada ao método model.fit (Código 12). Este método recebeu como entrada os dados de treinamento (X\_treino, y\_treino), que consistem nas sequências de características combinadas e seus respectivos rótulos numéricos de emoção.



O número máximo de épocas foi definido como 100 e o tamanho do lote foi estabelecido em 16, indicando que os pesos do modelo seriam atualizados após o processamento de 16 sequências de vídeo.

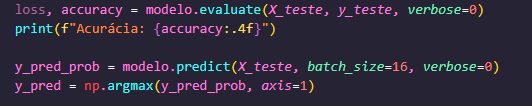
Para monitorar o desempenho e controlar o processo de treinamento, foram utilizados callbacks do Keras. O EarlyStopping foi empregado para interromper o treinamento caso a perda de validação não apresentasse melhora por um número consecutivo de épocas e carregasse os pesos do melhor modelo salvo pelo ModelCheckpoint ao final.

Além disso, o ReduceLROnPlateau foi utilizado para reduzir a taxa de aprendizado do otimizador sempre que a perda de validação estagnasse por um determinado número de épocas, auxiliando na convergência para um mínimo melhor.

O conjunto de validação separado inicialmente (20% dos dados), foi fornecido ao modelo treinado, permitindo que o modelo fosse avaliado ao final de cada época, fornecendo informações sobre a generalização do modelo e servindo como base para a atuação dos call-backs.

**4.4.2 Teste e Classificação**

Após a conclusão do treinamento ao atingir o EarlyStopping na época 42, o modelo treinado foi utilizado para realizar predições no conjunto de teste, vistos no código 13.



Para obter a classificação final para cada sequência, aplicou-se a função np.argmax sobre as probabilidades preditas ao longo do eixo das classes. Isso seleciona o índice (correspondente à classe de emoção) com a maior probabilidade como a predição do modelo para aquela sequência.

Com os rótulos preditos e os rótulos verdadeiros para o conjunto de teste, foi possível calcular diversas métricas de desempenho utilizando funções da biblioteca Scikit-learn.

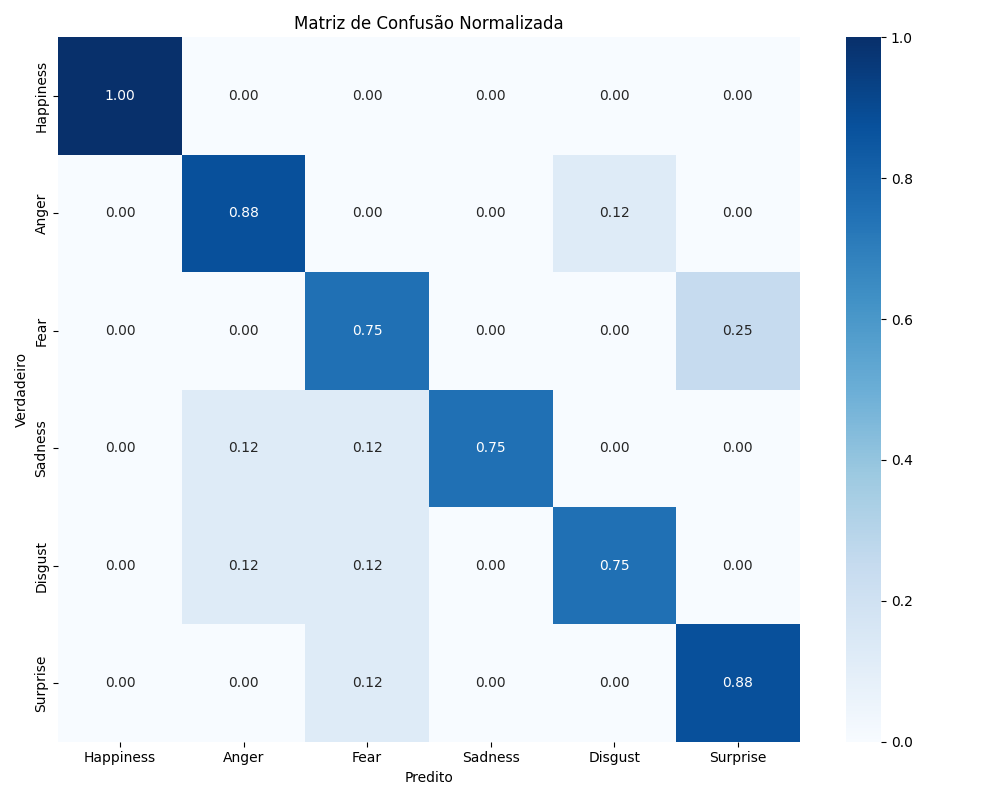
**4.5 Resultados**

A avaliação do sistema de reconhecimento de emoções foi realizada com o conjunto de teste separado, composto por 20% dos dados totais e que não foram utilizados durante o treinamento, fornecendo assim diversas informações sobre o desempenho do modelo proposto.

As métricas quantitativas calculadas resumem a capacidade do classificador LSTM, alimentado pelas características HOG e CNN combinadas, em generalizar para dados não vistos.

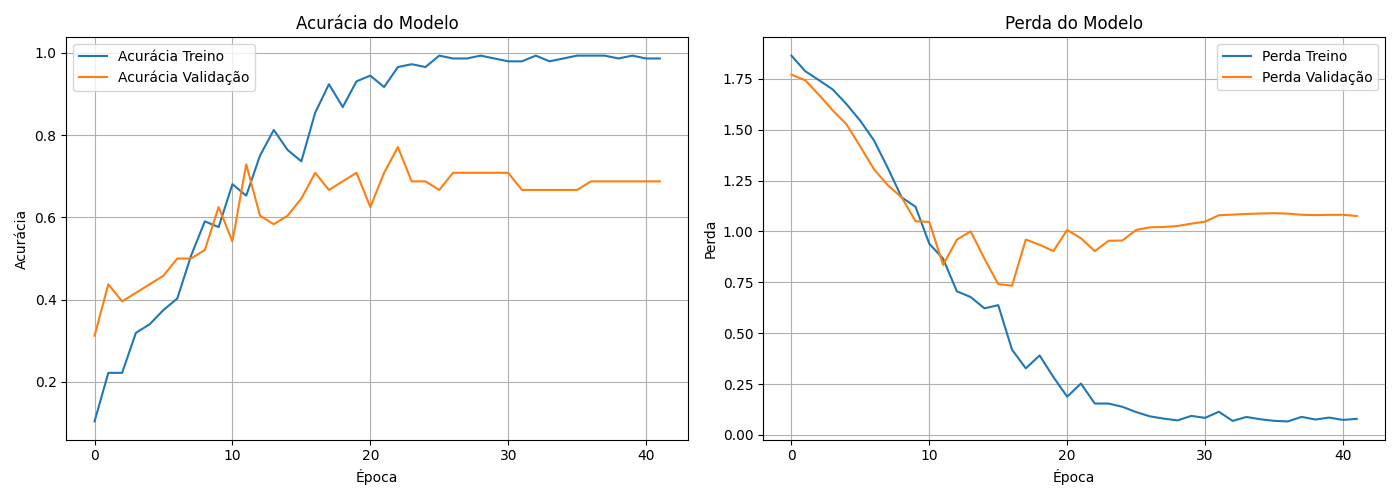
A acurácia média alcançada no conjunto de testem, ou seja, a porcentagem geral de acerto do modelo foi de 83.0%. A precisão, que mede a exatidão de previsões verdadeiras foi de 84.7% e o F1-score, uma média ponderada entre Precisão e Recall (capacidade de encontrar todos os positivos) atingiu 83.5%.

A análise da matriz de confusão (Figura abaixo) permitiu uma compreensão mais aprofundada do desempenho por classe.



Observou-se que a emoção de Felicidade (Happiness) foi classificada com uma alta precisão (100%). Por outro lado, apresentou uma maior confusão entre as emoções Medo (Fear)/Surpresa(Surprise) e Raiva(Anger)/Nojo(Disgust), indicando uma dificuldade do modelo em distingui-las com base nas características aprendidas.

Na Figura abaixo podemos observar as curvas de aprendizado, onde são plotados a acurácia e a perda nos conjuntos de treinamento e validação ao longo das épocas, forneceram informações sobre o processo de treinamento.



A diferença entre as curvas de treino e validação sugere um grau considerável de overfitting, indicando que o modelo se especializou nos dados de treinamento, mas não generaliza tão bem para dados novos.

A análise visual das amostras(Figura abaixo) confirmou a capacidade do modelo em classificar corretamente diversas expressões. No entanto, também foram observados erros de classificação que confirmam os dados da matriz de confusão, como a classificação incorreta de Raiva como Nojo e de Medo como Surpresa.



Para validar ainda mais o sistema, foram utilizados 2 vídeos espontâneos para testar a aplicabilidade prática do sistema. Os resultados obtidos nessa validação indicam a capacidade de generalização do modelo em cenários do mundo real.

Em conjunto, os resultados demonstram a viabilidade do modelo para a identificação de emoções, alcançando um desempenho notável. Porém, os resultados também evidenciam desafios significativos, principalmente na distinção entre emoções com expressões faciais semelhantes.

# CONCLUSÃO

O presente trabalho teve como principal objetivo a implementação de um sistema de reconhecimento automático de emoções por meio da análise sequencial, utilizando técnicas de visão computacional e aprendizado de máquina.

Para isso foram aplicadas técnicas para facilitar a análise, incluindo a detecção facial com classificadores em cascata baseados em características Haar (HAAR cascade), a combinação de características extraídas por Histograma de Gradiente Orientado (HOG) com as extraídas por uma Rede Neural Convolucional(CNN); e uma Long Short-Term Memory (LSTM), com o objetivo de classificar a sequência de frames em uma emoção.

A avaliação do sistema, utilizando a base de dados KDEF-dyn e dados reais, apresentou um bom desempenho geral. Destacou-se a excelente precisão para a emoção “Felicidade” e taxas de acerto entre 75% e 88% para outras emoções.

Entretanto, foram identificadas limitações significativas, visto que o sistema apresentou dificuldade em distinguir algumas emoções. Além disso, a análise do treinamento revelou a ocorrência de overfitting no modelo, comprometendo sua capacidade de generalização para novos dados.

Para trabalhos futuros, recomenda-se a aplicação de técnicas de regularização mais eficazes. Também é crucial ampliar e diversificar os conjuntos de dados utilizados no treinamento e na validação, de modo a incluir maior representatividade de etnias, faixas etárias, condições de iluminação e espontaneidade das expressões.

Como sugestão adicional, propõe-se a adaptação do sistema para operação em tempo real, com captura dos vídeos via webcam. Além disso, a integração multimodal, combinando a análise visual com informações de áudio, poderá contribuir significativamente para a robustez do sistema.

Conclui-se que a arquitetura proposta é tecnicamente viável e promissora. As limitações identificadas e as sugestões para trabalhos futuros abrem caminhos para novas pesquisas que busquem superar os desafios atuais e avançar no desenvolvimento de sistemas de reconhecimento automático de emoções mais precisos e eficazes.

# REFERÊNCIAS

ALVES, Priscila Mello. **Inteligência Artificial e Redes Neurais**. Brasília: IPEA, 2020. Disponível em: https://www.ipea.gov.br/cts/pt/central-de-conteudo/artigos/artigos/106-inteligencia-artificial-e-redes-neurais. Acesso em: 18 jun. 2024.

ANIL, J.; PADMA SURESH, L. A novel fast hybrid face recognition approach using convolutional Kernel extreme learning machine with HOG feature extractor. **Sensors**, [S.l.], v. 30, 2023. DOI: <https://doi.org/10.1016/j.measen.2023.100907>.

AWARI. **Redes neurais convolucionais: como elas funcionam**. 2023. Disponível em: <https://awari.com.br/redes-neurais-convolucionais/>. Acesso em: 18 jun. 2024.

BALLESTEROS, J.; Ramirez, G.; Moreira, F.; Solano, A.; Palaez, C. (2024). Facial emotion recognition through artificial intelligence. Fronties in Computar Science, <https://doi.org/10.3389/fcomp.2024.1359471>

Basiri, M.E., Nemati, S., Abdar, M., Cambria, E. and Acharya, U.R. (2021) ABCDM: An Attention-Based Bidirectional CNN-RNN Deep Model for Sentiment Analysis. Future Generation Computer Systems, 115, 279-294. https://doi.org/10.1016/j.future.2020.08.005

BASTOS, Elísio Augusto Velloso; ESTEVES, Vitória Barros. Tecnologias de reconhecimento facial: um estudo a partir do contexto de vigilância digital e sutil. **Direitos Democráticos & Estado Moderno**, São Paulo, v. 3, n. 1, p. 91-107, jan./jun. 2021. Disponível em: <https://revistas.pucsp.br/index.php/DDEM/article/view/53875>. Acesso em: 15 abr. 2025.

BHATT, D. et al. Computer vision and deep learning for emotion recognition in images: a survey. **Journal of Artificial Intelligence Research**, [S.l.], v. 1, p. 1-25, 2020.

BOCK, Ana Mercês Bahia; FURTADO, Odair; TEIXEIRA, Maria de Lourdes Trassi. **Psicologias: uma introdução ao estudo de psicologia**. 14. ed. São Paulo: Saraiva, 2008.

CACIOPPO, J. T. et al. The psychophysiology of emotion. In: LEWIS, M.; HAVILAND-JONES, J. M. (Ed.). **Handbook of emotions**. 2. ed. New York: Guilford Press, 2000. p. 173-191.

CÎRNEANU, Andrada-Livia; POPESCU, Dan; IORDACHE, Dragoş. Novas tendências no reconhecimento de emoções usando análise de imagens por redes neurais, uma revisão sistemática. **Sensors**, [S.l.], v. 23, n. 16, p. 1-20, 2023. DOI: 10.3390/s23167092.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273-297, 1995.

DAMÁSIO, António R. **O erro de Descartes: emoção, razão e o cérebro humano**. São Paulo: Companhia das Letras, 1994.

EKMAN, Paul; FRIESEN, Wallace V. Measuring facial movement. **Environmental Psychology and Nonverbal Behavior**, [S.l.], v. 1, n. 1, p. 56-75, 1976.

GONZÁLEZ, Rafael C.; WOODS, Richard E. **Processamento digital de imagens**. 3. ed. São Paulo: Pearson Prentice Hall, 2010.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Disponível em: <https://www.deeplearningbook.org/>HOCHREITER, S.;

HANAFI, Anasse &amp; BOUHORMA, Mohammed &amp; LOFTI, Elaachak. (2021). Building a Deep Learning Model to Generate Human Readable Text Using Recurrent Neural Networks and LSTM. 10.21203/rs.3.rs-753724/v1.

HUANG, Z.; CHIANG, C.; CHEN, J.; CHUNG, H.; CAI, Y.; HSU, H. A (2023). study on computer vision for facial emotion recognition. Scientific Reports, (2023) 13:8425, <https://doi.org/10.1038/s41598-023-35446-4>

Implementation of Haar Cascade Classifier and Eye Aspect Ratio for Driver Drowsiness Detection Using Raspberry Pi - Scientific Figure on ResearchGate. Disponível em: https://www.researchgate.net/figure/a-Edge-feature-b-Line-feature-and-c-Four-Triangle-feature\_fig5\_339448897

**KAROLINSKA INSTITUTET.** Karolinska Directed Emotional Faces – dynamic version (KDEF-dyn) [banco de dados]. Disponível em: <https://kdef.se/>. Acesso em: 20 maio 2025.

Lienhart, R., & Maydt, J. (2002). An extended set of Haar-like features for rapid object detection. In Proceedings of the International Conference on Image Processing (Vol. 1, pp. I-I). IEEE.

MIGUEL, Fabiano Koich. Psicologia das emoções: uma proposta integrativa para compreender a expressão emocional. **Psico-USF**, Itatiba, v. 20, n. 1, p. 153-162, jan./abr. 2015. DOI: 10.1590/1413-82712015200114.

PRATEEK, Joshi. **Artificial Intelligence with Python**. Mumbai: Packt, 2017.

RANGULOV, Denis; FAHIM, Muhammad. Emotion Recognition on Large Video Dataset Based on Convolutional Feature Extractor and Recurrent Neural Network. In: IEEE International Conference on Image Processing, Applications and Systems (IPAS 2020), 4., 2020, Genova, Itália. Anais [...]. Genova: IEEE, 2021. p. 14–20. DOI: 10.1109/IPAS50080.2020.9334935.

SANDLER, Mark; HOWARD, Andrew; ZHU, Menglong; ZHMOGINOV, Andrey; CHEN, Liang-Chieh. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. [S.l.]: Google Inc., 2019. Disponível em: <https://arxiv.org/abs/1801.04381>.

SARVAKAR, K.; SENKAMALAVALLI, R.; RAGHAVENDRA, S.; KUMAR, j.; MANJUNATH, R.; JAISWAL, S. (2023). Facial emotion recognition using convolutional neural networks. Materials Today: Proceedings 80 (2023) 3560–3564. <https://doi.org/10.1016/j.matpr.2021.07.297>

SCHMIDT, K. L.; COHN, J. F. Human facial expressions as adaptations: evolutionary questions in facial expression. **American Journal of Physical Anthropology**, [S.l.], v. 44, n. S33, p. 3-24, 2001.

SEDAGHATJOO, Zeinab; HOSSEINZADEH, Hossein; BIGHAM, Bahram Sadeghi. Local Binary Pattern (LBP) Optimization for Feature Extraction. arXiv preprint arXiv:2407.18665v1, 26 jul. 2024. Disponível em: https://arxiv.org/abs/2407.18665. Acesso em: 24 maio 2025

SHANMUGAMANI, Rajalingappaa. **Deep Learning for Computer Vision**. Mumbai: Packt, 2018.

**SHARMA, N.** What is MobileNetV2? Features, Architecture, Application and More. *Analytics Vidhya*, [S. l.], dez. 2023. Disponível em: https://www.analyticsvidhya.com/blog/2023/12/what-is-mobilenetv2-features-architecture-application-and-more/. Acesso em: 3 jun. 2025.

SHU, Chang; DING, Xiaoqing; FANG, Chi. *Histogram of the Oriented Gradient for Face Recognition*. **Tsinghua Science and Technology**, Beijing, v. 16, n. 2, p. 216–224, abr. 2011. ISSN 1007-0214.

SORIA OLIVAS, Emilio; MARTIN GUERRERO, Jose David; MARTINEZ SOBER, Marcelino; MAGDALENA BENEDITO, Jose Rafael; SERRANO LOPEZ, Antonio Jose. *Manual de pesquisa sobre aplicações e tendências de aprendizado de máquina: algoritmos, métodos e técnicas*. Referência em Ciência da Informação, 2009.

**VIOLA, Paul; JONES, Michael.** Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, Kauai, HI. Anais [...]. IEEE, 2001. p. I-511-I-518.

WILHELM, O. et al. Test battery for measuring the perception and recognition of facial expressions of emotion. **Frontiers in Psychology**, [S.l.], v. 5, p. 1-23, 2014. DOI: 10.3389/fpsyg.2014.00404.