

Identificação de emoções em tempo real utilizando Redes Neurais Convolucionais

Luis F. Bertuol Ulisses L. da Silva

Instituto Tecnológico de Aeronáutica
São José dos Campos/SP – 2025

Resumo—Este trabalho apresenta o desenvolvimento de um sistema para reconhecimento de emoções faciais em tempo real a partir de imagens capturadas por webcam. O pipeline proposto combina a detecção de rostos por meio de classificadores em cascata do tipo Haar (*HaarCascade*), implementados em OpenCV, com a classificação das expressões faciais realizada por uma rede neural convolucional (CNN). Para o treinamento, foi utilizada a base Augmented FER2013, composta por aproximadamente 215 mil imagens em escala de cinza, balanceadas em sete classes emocionais. Três arquiteturas convolucionais foram avaliadas e a rede escolhida, de inspiração VGG, alcançou acurácia de validação de 62,49% e acurácia de 58,3% no conjunto de teste, com melhor desempenho nas classes *happy* e *surprise*. O sistema resultante foi integrado a uma aplicação em tempo real, capaz de detectar a face e sobrepor a emoção prevista diretamente no fluxo de vídeo, com taxa de quadros adequada ao uso interativo. Por fim, discutem-se limitações relacionadas ao desbalanceamento entre classes e apontam-se melhorias futuras, como o uso de *transfer learning* e arquiteturas mais profundas que possam captar melhor os padrões e, consequentemente, ter melhor desempenho na classificação.

I. INTRODUÇÃO

As Redes Neurais Convolucionais (CNNs) têm se consolidado como abordagem dominante para processamento e classificação de imagens. Por meio da extração hierárquica de características, essas redes são capazes de reconhecer padrões complexos como bordas, texturas e formas humanas. Aplicadas ao reconhecimento de expressões faciais, as CNNs permitem identificar emoções com desempenho superior a abordagens clássicas de visão computacional.

Neste trabalho propõe-se o desenvolvimento de um sistema para detecção e classificação de emoções faciais em tempo real, utilizando uma CNN treinada sobre a base de dados Augmented FER2013. A detecção dos rostos é realizada por meio de classificadores Haar (*HaarCascade*), técnica eficiente e amplamente difundida em aplicações vídeo em tempo real utilizando OpenCV, que oferece bom desempenho na detecção de faces com um baixo custo computacional. Após detectar a face, um recorte é extraído, redimensionado para 48×48 pixels e enviado à CNN para inferência.

A. Problema

Este trabalho aborda o problema do reconhecimento de emoções a partir de imagens capturadas por webcam, com classificação entre sete categorias: *angry*, *disgust*, *fear*, *happy*, *sad*, *surprise* e *neutral*. O pipeline inclui detecção facial, pré-processamento, classificação e visualização em tempo real.

B. Motivação

O reconhecimento automático de emoções faciais constitui um dos desafios mais instigantes da visão computacional e da inteligência artificial contemporânea. A capacidade de um sistema identificar e interpretar estados emocionais humanos em tempo real abre possibilidades em múltiplos domínios, como educação personalizada, monitoramento de saúde, interação homem-máquina mais natural e sistemas de transporte inteligentes com detecção de sinais de fadiga ou estresse em motoristas. Nesse contexto, estudar e aprimorar soluções computacionais para detecção de emoções faciais é de grande relevância científica e social.

C. Objetivos

Objetivo geral: desenvolver um sistema funcional para detecção e classificação de emoções faciais em tempo real.

Objetivos específicos:

- treinar e validar uma rede convolucional usando a base Augmented FER2013;
- implementar e testar um pipeline completo (captura de vídeo, detecção facial e classificação);
- avaliar o desempenho utilizando métricas padrão (acurácia, F1, matriz de confusão).

II. METODOLOGIA

A. Base de dados

Foi utilizada a base *Augmented FER2013*, uma versão expandida da FER2013 original. Na versão utilizada, obtida no Kaggle, o conjunto contém aproximadamente 215 mil imagens em escala de cinza com resolução 48×48 , distribuídas em sete classes emocionais.

O diretório original encontra-se organizado em duas pastas principais: `train` e `test`. No total, foram contabilizadas:

- 172.254 imagens na pasta de treinamento;
- 43.068 imagens na pasta de teste.

Durante o treinamento, foi utilizada divisão interna de validação com `validation_split = 0.1`, resultando em:

- ≈ 155.030 imagens para treino efetivo;
- ≈ 17.224 imagens para validação;
- 43.068 imagens para teste independente.

As imagens foram carregadas em formato $1 \times 48 \times 48$ (um canal), com normalização dos pixels para o intervalo $[0, 1]$.

B. Pré-processamento e aumento de dados

Foi utilizada a classe `ImageDataGenerator` da biblioteca Keras, com normalização de pixel e *data augmentation*. Foram aplicadas:

- rotações aleatórias ($\pm 15^\circ$);
- *zoom* aleatório;
- deslocamentos horizontais e verticais;
- inversão horizontal (*horizontal flip*).

Esse conjunto de transformações visa reduzir sobreajuste e aumentar a variabilidade do conjunto de treinamento.

C. Comparação de arquiteturas

Três arquiteturas foram testadas:

- **Modelo 1 – Light**: rede rasa com três blocos Conv2D + MaxPooling.
- **Modelo 2 – Base (Vencedor)**: rede estilo *VGG*, com quatro blocos convolucionais e camada densa de 512 neurônios com ReLU.
- **Modelo 3 – Heavy**: rede mais profunda com ativações ELU e regularização extra.

Cada modelo foi treinado com Adam, categorical cross-entropy, early stopping, ReduceLROnPlateau e ModelCheckpoint. A escolha final baseou-se na **melhor acurácia de validação**, sendo selecionado o **Modelo 2 (Base)**, com *val accuracy* final de:

$$\text{val_accuracy} = 0.6249$$

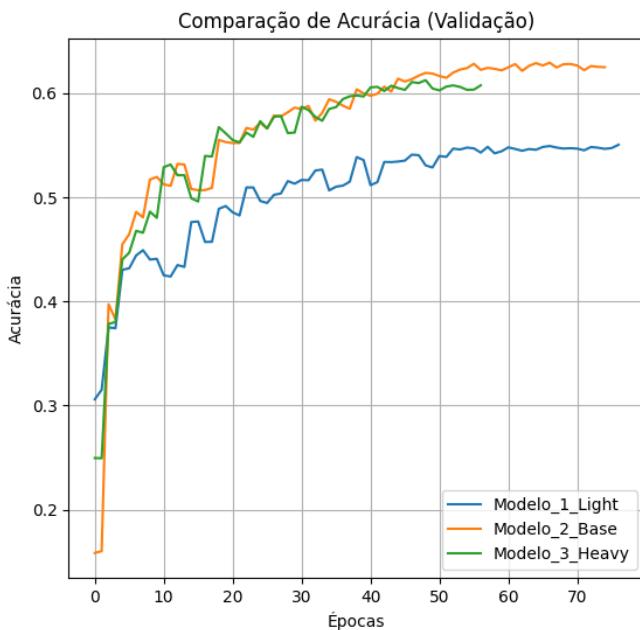


Figura 1. Curvas de acurácia dos três modelos testados.

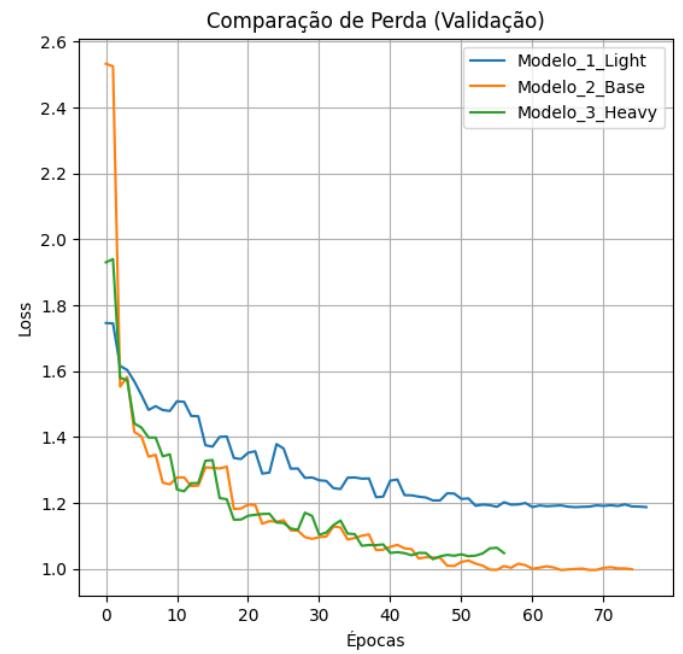


Figura 2. Curvas de loss dos três modelos testados.

D. Detecção facial em tempo real

A detecção facial foi realizada com o classificador *Haar-Cascade* distribuído com o OpenCV, aplicando janelas escalonadas ao quadro da webcam e retornando coordenadas da região facial. Posteriormente, o recorte é convertido para escala de cinza, redimensionado para 48×48 e passado à CNN para predição.

E. Treinamento

Hiperparâmetros:

- Optimizador: Adam;
- Função de perda: *categorical cross-entropy* com *class_weights*;
- Número de épocas: entre 100 e 200 (com early stopping);
- Lote: 64 imagens;
- Callbacks: ModelCheckpoint, EarlyStopping, ReduceLROnPlateau.

F. Avaliação e métricas

No conjunto de teste foram calculados:

- acurácia global;
- F1 *macro* e *micro*;
- matriz de confusão;
- relatório por classe (precisão, recall e F1-score);
- curvas de acurácia e perda por época.

III. REVISÃO TEÓRICA

A. Redes Neurais Convolucionais (CNNs)

As redes neurais convolucionais (CNNs) são arquiteturas projetadas para lidar com dados estruturados espacialmente,

como imagens. Elas são compostas por camadas *Convolutional*, responsáveis por extrair características locais da imagem utilizando filtros; camadas de *Pooling*, empregadas para reduzir dimensionalidade e controle de sobreajuste; e camadas totalmente conectadas (*Dense*), responsáveis pela classificação final.

As CNNs exploram três propriedades fundamentais:

- **Compartilhamento de pesos:** um mesmo filtro é aplicado a diferentes regiões da imagem.
- **Conektividade local:** os filtros analisam pequenas regiões (*patches*) ao invés da imagem inteira.
- **Extração hierárquica de características:** camadas iniciais capturam bordas e texturas, enquanto camadas profundas capturam características mais complexas como formas faciais.

A combinação dessas propriedades permite uma representação eficiente de imagens, tornando CNNs o estado da arte em tarefas de visão computacional, como reconhecimento de objetos, detecção facial e classificação de emoções.

B. Filtro HaarCascade para detecção facial

A detecção facial realizada neste trabalho utiliza classificadores *HaarCascade*, baseados no algoritmo de Viola e Jones. Os classificadores Haar extraem padrões retangulares de contraste entre regiões da imagem, permitindo identificar estruturas que caracterizam rostos humanos, como olhos, nariz, boca e contornos faciais.

O processo intercala múltiplos classificadores em cascata, onde cada estágio decide se uma região da imagem deve ser rejeitada ou analisada em etapas subsequentes. Essa estrutura torna o algoritmo eficiente, com baixa complexidade computacional, ideal para aplicações em tempo real.

No presente trabalho, o classificador HaarCascade foi utilizado para localizar e recortar regiões faciais em quadros capturados pela webcam. Esses recortes foram redimensionados para 48×48 pixels em escala de cinza e enviados à CNN final para predição da emoção.

IV. DESENVOLVIMENTO

A. Preparação do dataset

O conjunto de dados utilizado consiste em 172.254 imagens de treino e 43.068 imagens de teste, organizadas em pastas nomeadas com suas respectivas emoções. O carregamento foi realizado utilizando o método `flow_from_directory` do Keras, permitindo particionar automaticamente um conjunto de validação correspondente a 10% das imagens de treino.

Data augmentation foi aplicado durante a geração de batches de treino, a fim de aumentar a robustez e a capacidade de generalização do modelo.

B. Construção da Rede Neural Convolucional

O modelo final adotado foi o nomeado como **Modelo 2 – Base**. Ele é composto por quatro blocos convolucionais com filtros 3×3 e funções de ativação ReLU, intercalados

por camadas MaxPooling2D, seguidos de uma camada totalmente conectada de 512 neurônios com *dropout* de 0.5.

O treinamento foi realizado com o otimizador Adam, taxa de aprendizado adaptativa e função de perda *categorical cross-entropy*. O uso de EarlyStopping permitiu interromper automaticamente o processo quando não havia melhora do desempenho em validação por um número pré-definido de épocas.

C. Aplicação do HaarCascade

A lógica de detecção facial em tempo real foi desenvolvida utilizando a biblioteca OpenCV. O fluxo de processamento por quadro foi desenhado da seguinte maneira:

- 1) Captura de quadro da webcam;
- 2) Conversão para escala de cinza;
- 3) Aplicação do filtro HaarCascade para detecção facial;
- 4) Recorte e redimensionamento para 48×48 ;
- 5) Normalização e passagem pela CNN para inferência;
- 6) Sobreposição da emoção detectada na tela.

V. RESULTADOS

A. Desempenho durante o treinamento

A Figura 3 mostra as curvas de acurácia por época durante o treinamento e validação. Observa-se aumento progressivo da acurácia nas etapas iniciais, seguido por estabilização. A validação atingiu valor máximo próximo de **0.6249**, com acurácia final de treino de **0.5585**, resultando em um *generalization gap* negativo de aproximadamente -6.64% , indicando que o modelo aprendeu a generalizar melhor do que memorizar.

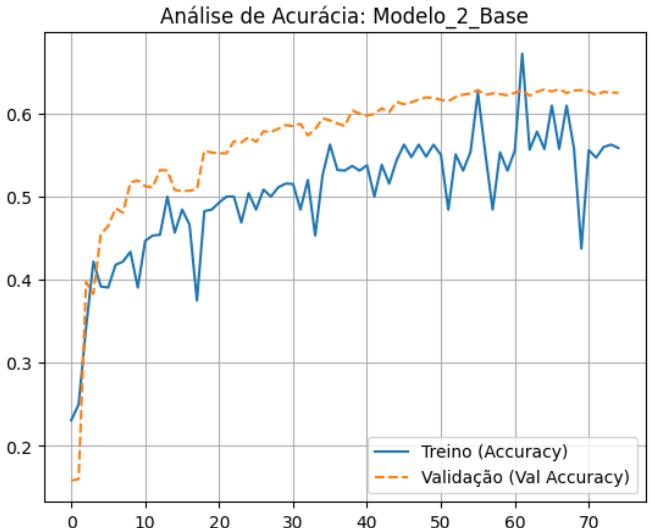


Figura 3. Curvas de acurácia (treinamento e validação) por época do modelo 2.

A Figura 4 apresenta a perda do modelo ao longo das épocas. Nota-se que a perda de validação, após inicialmente diminuir, passou a oscilar, indicando início de sobreajuste.

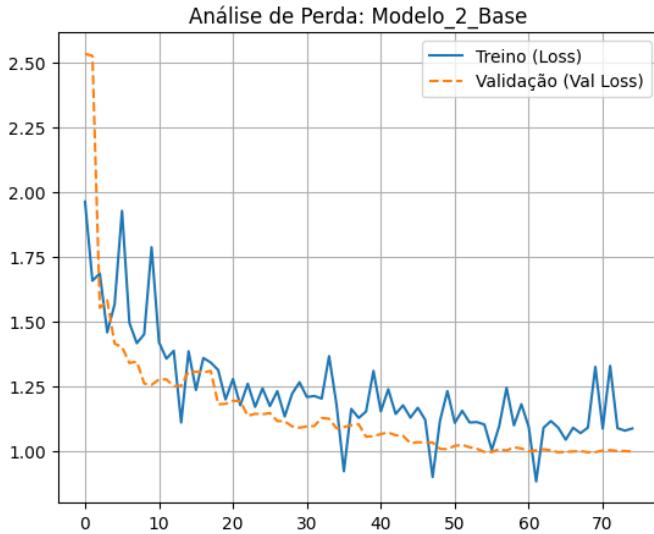


Figura 4. Curvas de loss (treinamento e validação) por época do modelo 2.

B. Avaliação por classe

A Tabela I apresenta o relatório de classificação sobre o conjunto de teste contendo 43.068 imagens. Esses valores foram gerados utilizando a função `classification_report` do `sklearn.metrics`.

Tabela I
RELATÓRIO DE CLASSIFICAÇÃO POR CLASSE (TESTE, 100 ÉPOCAS)

Classe	Precision	Recall	F1-score	Suporte
Angry	0.53	0.49	0.50	5748
Disgust	0.20	0.80	0.33	666
Fear	0.48	0.31	0.37	6144
Happy	0.82	0.80	0.81	10644
Neutral	0.52	0.60	0.56	7398
Sad	0.48	0.40	0.44	7482
Surprise	0.64	0.79	0.71	4986
Acurácia			0.58	43068
Média Macro	0.53	0.60	0.53	43068
Média Ponderada	0.59	0.58	0.58	43068

C. Matriz de Confusão

A matriz de confusão obtida no teste é apresentada na Figura 5. As classes *happy*, *disgust* e *surprise* foram as mais bem identificadas, enquanto *fear* e *neutral* apresentaram desbalanceamento considerável.

D. Simulações em tempo real

Para validar o funcionamento prático do modelo treinado, foi realizada uma simulação utilizando a webcam do computador. A inferência em tempo real apresentou boa responsividade. A detecção facial foi realizada em média a 25–30 FPS. As Figuras 6 a 12 exemplificam os quadros capturados com a emoção prevista sobreposta ao rosto detectado.

Os resultados visuais reforçam a capacidade do sistema em identificar corretamente diferentes expressões faciais em

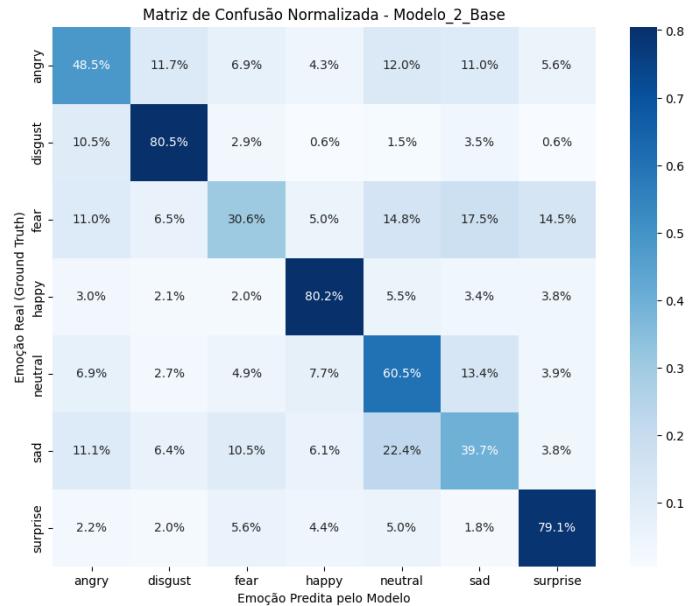


Figura 5. Matriz de confusão da predição sobre o conjunto de teste.

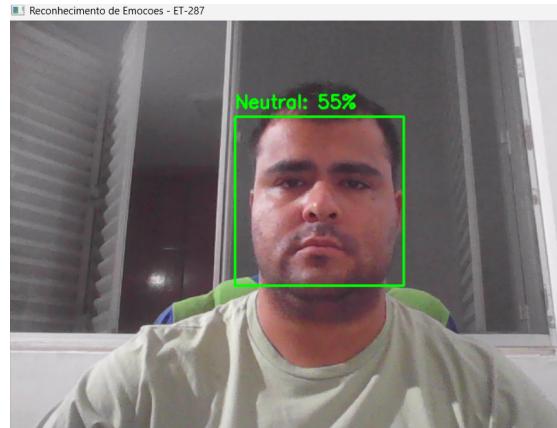


Figura 6. Exemplo de detecção facial e predição da emoção em tempo real.

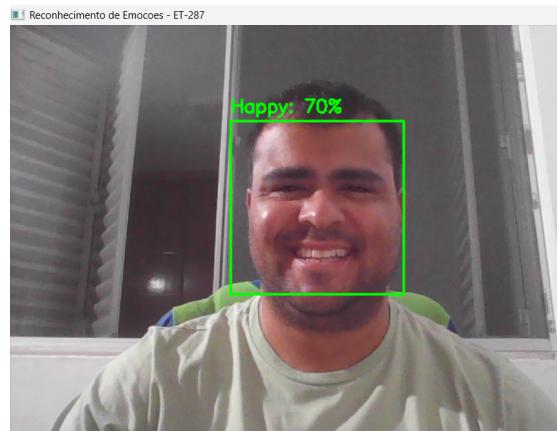


Figura 7. Exemplo de detecção facial e predição da emoção em tempo real.

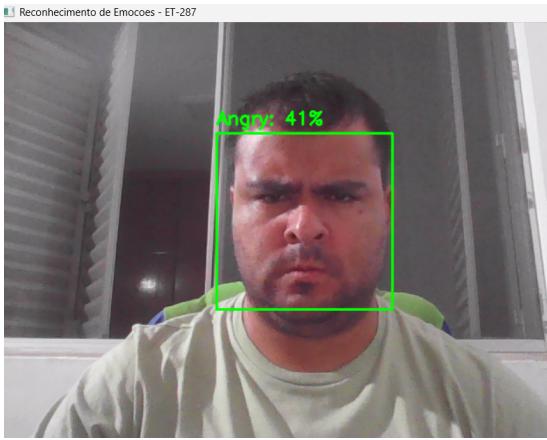


Figura 8. Exemplo de detecção facial e predição da emoção em tempo real.

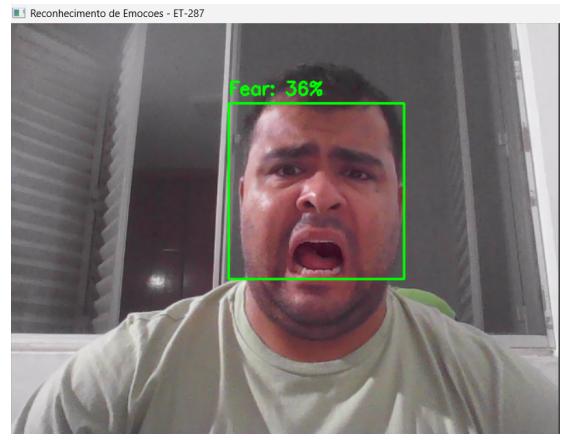


Figura 11. Exemplo de detecção facial e predição da emoção em tempo real.

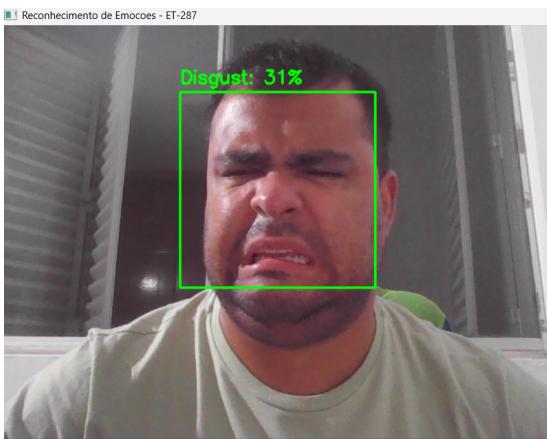


Figura 9. Exemplo de detecção facial e predição da emoção em tempo real.

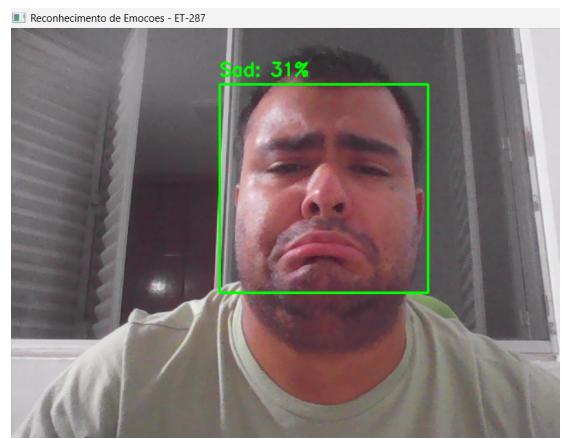


Figura 12. Exemplo de detecção facial e predição da emoção em tempo real.

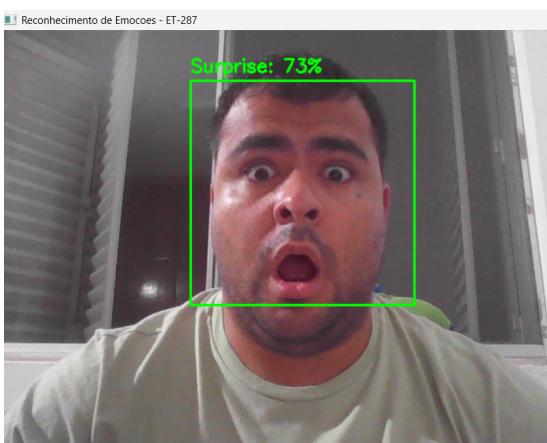


Figura 10. Exemplo de detecção facial e predição da emoção em tempo real.

tempo real, confirmando a razoável eficácia da solução desenvolvida para aplicações práticas. Não obstante, o modelo está plenamente apto a receber melhorias futuras.

VI. DISCUSSÃO DOS RESULTADOS

A classificação atingiu acurácia final de 0.5834 no conjunto de teste, com *macro F1* de 0.53 e *weighted F1* de 0.58. Esses valores indicam desempenho relativamente competitivo frente ao estado da arte em CNNs simples aplicadas à FER2013, especialmente considerando a execução em tempo real.

O melhor desempenho foi observado para as classes *happy* (0.82) e *surprise* (0.64), alinhado com a grande representatividade dessas classes na base de dados e maior distinção visual. Por outro lado, o caso de *disgust* é notavelmente peculiar: Apesar de apresentar *recall* elevado (0.80), a *precision* baixa (0.20) indica que o modelo continua classificando erroneamente imagens de outras classes como *disgust*, reforçando o impacto do desbalanceamento.

Os resultados apontam que abordagens complementares como *class weights*, *oversampling* direcionado (por exemplo, em classes com menos amostras como *Disgusting*) e uso de arquiteturas mais profundas (e.g., ResNet, MobileNet) podem ajudar a corrigir esses efeitos.

VII. CONCLUSÃO

Este trabalho apresentou o desenvolvimento de um sistema de reconhecimento de emoções faciais em tempo real utili-

zando CNNs e filtros HaarCascade. Os experimentos, conduzidos sobre uma versão aumentada da base FER2013, demonstraram desempenho satisfatório considerando a complexidade do problema e a variabilidade intrínseca das expressões faciais.

Embora o modelo tenha alcançado acurácia razoável, os resultados apontam desafios particulares relacionados ao desbalanceamento entre classes, especialmente para emoções com baixa representatividade. Melhorias futuras podem incluir o uso de técnicas de *transfer learning*, arquiteturas mais robustas, aumentos artificiais direcionados e detectores faciais mais precisos.

REPOSITÓRIOS

Os códigos-fontes desenvolvidos, incluindo as implementações da rede neural, scripts de treinamento e simulação, bem como também as imagens e arquivos de texto relativos aos testes e validações realizadas, estão disponíveis no repositório público, na subpasta "Projeto Final":

- <https://github.com/ulissesuls/ET-287>.

REFERÊNCIAS

- [1] I. Goodfellow et al., "Challenges in Representation Learning: A report on three machine learning contests", Unicer, 2013. Available: <http://arxiv.org/abs/1307.0414>.
- [2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. Available: <http://www.deeplearningbook.org>.
- [3] D. P. Lopez, F. Z. Canal, G. G. Scotton, E. Pozzebon, and A. C. Sobieranski, "Uma abordagem computacional em tempo real para reconhecimento de expressão facial humana baseada na extração de recursos de referência", *SEVENED – Revista de Saúde, Educação e Meio Ambiente*, vol. 4, no. 1, pp. 4–6, 2024. Available: <https://doi.org/10.56238/sevened2024.004-006>.
- [4] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)", *SN Applied Sciences*, vol. 2, no. 3, pp. 1–8, Feb. 2020. doi: <https://doi.org/10.1007/s42452-020-2234-1>.
- [5] A. Deshpande, "The 9 deep learning papers you need to know about (Understanding CNNs Part 3)", University of California, Los Angeles (UCLA).
- [6] R. Ratan, "Building an Emotion Detector with LittleVGG", GitHub repository. [Online]. Available: <https://github.com/rajeevratan84/DeepLearningCV/blob/master/18.2%20Building%20an%20Emotion%20Detector%20with%20LittleVGG.ipynb>. Accessed: Jul. 2025.
- [7] F. Z. Canal, "Método para Reconhecimento em Tempo Real de Expressões Faciais em Grupos utilizando Redes Neurais Convolucionais", M.S. thesis, Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2024.