Reconhecimento de expressões faciais utilizando redes neurais convolucionais

Anabel Marinho Soares[1], **Nicolas Emanuel Alves Costa**[1], **Thiago Luan Moreira Souza**[1], **Rosana C. B. Rego**[2]

[1]Tecnologia da Informação
Universidade Federal Rural do Semi-Árido – Brasil

[2]Departamento de Engenharias e Tecnologia
Universidade Federal Rural do Semi-Árido – Brasil

{anabel.soares, nicolas.costa, thiago.souza}@alunos.ufersa.edu.br

Abstract. *Emotional recognition can help identify signs of stress, anxiety, or other emotional issues, allowing for early interventions to ensure students' mental well-being. This study addresses facial expression recognition through convolutional neural networks (CNNs). Using the FER-2013 dataset from Kaggle, facial images are classified into seven distinct emotions. The process includes data preparation steps such as loading, preprocessing, and visualization. Subsequently, a CNN model is implemented and trained to recognize facial expressions. The results reveal an average accuracy of 86%, demonstrating notable effectiveness in real-time recognition. This study highlights the viability and accuracy of CNNs in facial expression recognition across various applications.*

Resumo. *O reconhecimento emocional pode ajudar a identificar sinais de estresse, ansiedade ou outras questões emocionais, permitindo intervenções precoces para garantir o bem-estar mental dos estudantes. Este estudo aborda o reconhecimento de expressões faciais por meio de redes neurais convolucionais (CNNs). Utilizando o conjunto de dados FER-2013 da Kaggle, as imagens de rostos são classificadas em sete emoções distintas. O processo inclui etapas de preparação de dados, como carregamento, pré-processamento e visualização. Posteriormente, um modelo CNN é implementado e treinado para reconhecer as expressões faciais. Os resultados revelam uma acurácia média de 86%, demonstrando uma eficácia notável no reconhecimento em tempo real. Este estudo destaca a viabilidade e a precisão das CNNs no reconhecimento de expressões faciais em diversas aplicações.*

1. Introdução

A expressão facial é uma maneira não-verbal pela qual os seres humanos se comunicam. É por meio de expressões faciais que o homem revela quem ele é, contribuem desde suas relações interpessoais até a construção da sua identidade emocional [Duarte 2005]. A cultura tem influência nas expressões faciais e pode moldá-la. Embora isso, segundo o psicólogo Paul Ekman, pioneiro no estudo de emoções e expressões faciais, em suas pesquisas transculturais, incluindo com a tribo Fore na Papua Nova Guiné [Ekman , Chemin 2019], confirma a universalidade das expressões faciais. Com esses

estudos, pode-se destacar que esse conhecimento adquirido por Ekman trouxe um avanço fundamental no desenvolvimento de algoritmos voltados à detecção de expressões faciais, ele concluiu que as expressões faciais mais usadas entre os humanos eram seis: raiva, alegria, aversão, tristeza, medo e surpresa [Chemin 2019], anos depois ele adicionou desprezo, totalizando sete. Com relação a isso, a conexão entre psicologia e tecnologia fica mais evidente com a evolução que a inteligência artificial vem tendo recentemente.

Nos últimos anos, foi visto uma notável intensificação no campo da inteligência artificial, por causa dos diversos avanços em diversas áreas tecnológicas, como as redes neurais e o aprendizado de máquinas [Kaufman 2019]. Contudo, apesar dos grandes avanços tecnológicos, ainda são enfrentados desafios para captar expressões faciais para algumas pessoas. De acordo com a psicóloga Cristina Rocca, transtornos psiquiátricos podem afetar na capacidade de identificação de expressões faciais [Gramado 2010]. Crianças bipolares, por exemplo, não conseguem compreender expressões de raiva.

Nessa situação, a análise e reconhecimento de expressões faciais é um fator importante que será trabalhado neste artigo por meio do uso de redes neurais artificias. Uma rede neural é um método de inteligência artificial que ensina computadores a processar dados de forma inspirada pelo cérebro humano [Hinton et al. 1992]. Uma sub área das redes neurais é o aprendizado profundo (*deep learning*), um tipo de algoritmo de machine learning, usa nós ou neurônios interconectados em uma estrutura em camadas, semelhantes ao cérebro humano.

Utilizando a técnica de *machine learning*, é possível captar emoções universais como raiva, alegria, tristeza, entre outras. Portanto, as contribuições do projeto são:

i) Precisão na captação de expressões;
ii) Aplicabilidade em diversas áreas, como segurança e saúde mental;
iii) Aprofundamento na compreensão das emoções humanas.

2. Trabalhos Relacionados

A análise e identificação de expressões faciais pode ser explorada em diversas pesquisas e estudos. Na pesquisa de [Hens 2021], são utilizadas redes neurais convolucionais (CNNs) em sistemas com recursos limitados a hardware. As CNNs aprendem por padrões locais em imagens através de filtros convolucionais e hierarquizam padrões espaciais durante o processo de aprendizado. Nessa abordagem hierárquica, a CNN aprende características simples nas primeiras camadas e depois identifica detalhes complexos, possibilitando o reconhecimento eficaz de padrões em imagens. No artigo [Chinchanikar 2019], os autores apresentam uma revisão sobre expressões faciais usando deep learning, além de também destacar CNNs. A pesquisa apresenta várias abordagens baseadas em 2D e 3D e na problemática em obter conjuntos de dados diversificados que representam condições do mundo real, como diferentes poses faciais, idade, iluminação, inclinação da cabeça etc. A conclusão do artigo é que o deep learning por meio de CNNs demonstrou um melhor desempenho no sistema para a identificação de expressões faciais, tendo eficácia tanto em imagens 2D como em 3D. O trabalho de Pramerdorfer e Kampel [Pramerdorfer and Kampel 2016] revisa o estado da arte nesse campo, discutindo abordagens eficazes e áreas de pesquisa em desenvolvimento. Em um trabalho relacionado, Liu et al. [Liu et al. 2018] apresentam uma revisão abrangente sobre o reconhecimento de expressões faciais em tempo real utilizando deep learning, com foco especial em CNNs.

Seu estudo, intitulado ”Deep learning-based real-time facial expression recognition: A survey”, destaca as técnicas e os desafios enfrentados nessa área. Além disso, Barsoum et al. [Barsoum et al. 2016] propõem uma arquitetura de rede neural profunda para o reconhecimento de expressões faciais em ”Deep Learning Approach for Facial Expression Recognition”. Eles investigam a eficácia dessa abordagem em comparação com outras técnicas utilizadas no campo do reconhecimento de expressões faciais. Esses trabalhos fornecem insights valiosos sobre o estado atual da pesquisa em reconhecimento de expressões faciais usando CNNs e deep learning.

3. Coleta de dados

Para o desenvolvimento deste trabalho, fez-se o uso do dataset FER-2013 disponível na plataforma Kaggle. O conjunto de dados consiste em imagens de 48×48 pixels de faces em uma escala de cinza. As faces foram automaticamente registradas para que o rosto estivesse mais ou menos centralizado e ocupa quase a mesma quantidade espaço em cada uma das imagens. A Figura 1 mostra exemplos das imagens do *dataset*.

Figure 1. Exemplos de expressões faciais do conjunto de imagens utilizadas.

Foi realizado processamento nas imagens e categorizadas com base em sete emoções. A Tabela 1 mostra esse mapeamento da categorização e a respectiva emoção. Todas as faces devem ser categorizadas baseadas em uma das sete emoções.

Índice	Emoção
0	Raiva
1	Nojo
2	Medo
3	Feliz
4	Triste
5	Surpreso
6	Neutro

Table 1. Mapeamento dos índices das emoções.

3.1. Tratamento de dados

O tratamento de dados utilizando esse conjunto envolve várias etapas comuns de pré-processamento de dados.

1. Carregamento dos dados: As imagens são carregadas do Dataset e são separadas em ArrayLists de treino e teste, dentro dessas ArrayLists se encontram as classes dessas imagens, nesse caso, as emoções.
2. Pré-processamento do Conjunto de Dados: Os pixels das imagens são convertidos em arrays e os valores dos pixels são normalizados.

3. Divisão em Conjuntos de treino e teste: Os dados são divididos em conjuntos de treinamento 80% e teste 20%.
4. Normalização dos dados: Os dados passam por um processo de normalização, para que todas as classes de imagens tenham a mesma quantidade de exemplos e faixa de valores.
5. Definição do modelo CNN (Rede Neural Convolucional): O modelo é definido usando a API Sequential do Keras. O modelo consiste em camadas densas intercaladas com camadas de dropout. A última camada usa a função softmax para gerar probabilidades para as diferentes emoções.
6. Compilação do modelo: O modelo é compilado com o otimizador Adam e uma taxa de aprendizado de 0.001%.
7. Treinamento do modelo: Durante o treinamento, um callback é configurado para salvar o melhor modelo com base na acurácia de validação. O modelo é treinado com 109 épocas, até que atinge a parada antecipada.

4. Modelo de aprendizado de máquina

Para desenvolvimento do algoritmo, foi realizado a implementação de uma arquitetura CNN. As CNNs são uma classe especializada de redes neurais profundas, projetadas principalmente para processar dados em forma grade, como imagens. Elas são capazes de aprender automaticamente características úteis das imagens, como bordas, texturas e padrões mais complexos [Goodfellow et al.]. Essas redes são compostas por uma série de camadas, incluindo camadas convolucionais, de *pooling* e totalmente conectadas. Durante a fase de feedforward, os dados de entrada são propagados pela rede, passando por cada camada, enquanto durante a fase de treinamento, a rede é ajustada usando o algoritmo de backpropagation para minimizar a diferença entre as previsões e os rótulos reais [Goodfellow et al. , Lecun et al. 1998].

As CNNs são amplamente aplicadas em diversas áreas da visão computacional, incluindo reconhecimento de objetos, classificação de imagens, detecção e reconhecimento de rostos, segmentação de imagens e processamento de vídeo. A Figura 2 mostra a arquitetura de rede CNN desenvolvida para o reconhecimento de expressões.

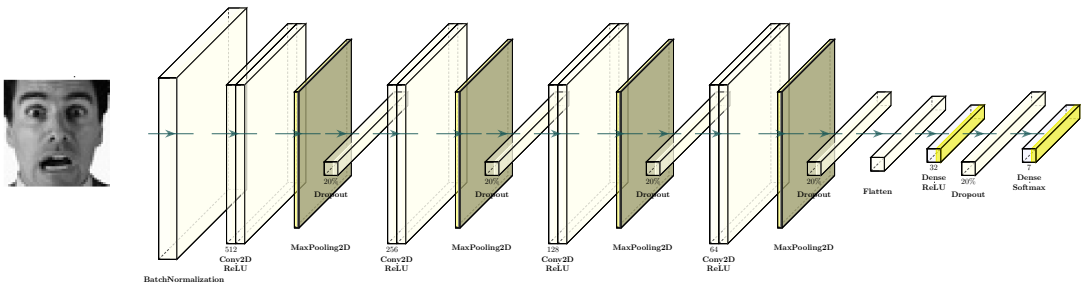


Figure 2. Arquitetura da Rede Neural Convolucional implementada.

Conforme a Figura 2, a rede recebe a imagem de entrada no formato 48×48, a imagem passa pelas três camadas convolucionais, respectivamente com 512, 128 e 64 neurônios. Além disso, foi adicionado camadas para a operação *maxpooling*. Nas camadas convolucionais, a função de ativação ReLu $f(x)$ dada pela equação

$$f(x) = max(0, x) \tag{1}$$

foi utilizada. Para evitar *overfitting* no modelo, a operação de *dropout* foi aplicada. Por fim, foi adicionado uma camada totalmente conectada com 32 neurônios e função de ativação ReLu. Na camada final, a função softmax dada pela equação,

$$f_{softmax}(s)_i = \left(\frac{e_i^s}{\sum_j^C e_j^s} \right) \tag{2}$$

foi aplicada, em que s é a saída da rede neural associada à classe i.

A Figura 3 mostra um exemplo de funcionamento da CNN para reconhecimento de expressões faciais. A rede tenta capturar cada detalhe da imagem para realizar o reconhecimento da expressão.

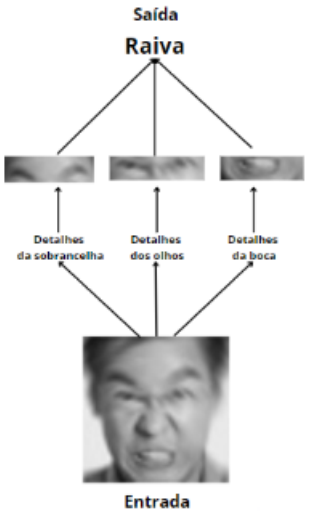


Figure 3. Exemplificação simplificada do funcionamento do reconhecimento de expressões faciais.

Essas redes revolucionaram a área de visão computacional, alcançando resultados impressionantes em uma variedade de tarefas. Sua capacidade de aprender representações hierárquicas de dados de entrada torna-as extremamente poderosas em lidar com problemas complexos de análise de imagens [Goodfellow et al. , Lecun et al. 1998].

4.1. Treinamento

Os modelos de *deep learning* precisam de uma função de perda para calcular o desempenho do modelo durante o treinamento. Como o problema, neste trabalho é uma classificação multiclasses (7 classes possíveis, ou 7 expressões para serem reconhecidas), a função de perda entropia cruzada foi selecionada. A função é dada pela equação

$$Loss(y, s_p) = -\sum_i^N y \cdot \left[log \left(\frac{e^{s_p}}{\sum_j^C e_j^s} \right) \right] \tag{3}$$

onde s_p é a saída da rede neural e y é a probabilidade/saída verdadeira.

Para treinamento do modelo a divisão do dataset de 80/20 foi adotada, alocando 80% dos dados para treinamento e reservando 20% para teste. Após isso, os dados de

treinamento passaram por uma normalização substancial para que todas as classes de imagem tivessem a mesma quantidade de exemplos, fazendo com que todas as classes ficassem com 7000 exemplos. Isso garantiu que nenhuma classe ficasse desbalanceada, visto que, antes disso, existia uma discrepância no número de exemplos por classe.

5. Resultados e discussões

O conjunto de treino consiste em 28,709 exemplos e o conjunto de testes consiste em 3,589 exemplos. Após a implementação do treinamento do modelo CNN proposto no conjunto de dados FER-2013, os resultados revelaram importantes *insights* sobre a eficácia do modelo de reconhecimento de expressões faciais.

O treinamento do modelo foi iniciado com a meta de 150 épocas, para garantir um bom resultado, cada uma representando uma passagem completa pelo conjunto de treinamento. Devido a função de parada antecipada previamente definida, após 5 épocas seguidas de treinamento desfavorável, a função foi finalizada, totalizando 109 épocas de treinamento no total. Esse treinamento foi capaz de garantir um modelo que provém os resultados mostrados na Tabela 2.

Métricas	**Valor**
Acurácia (%)	86.71
Precisão (%)	86.58
Recall (%)	87.05
F1-Score (%)	86.80

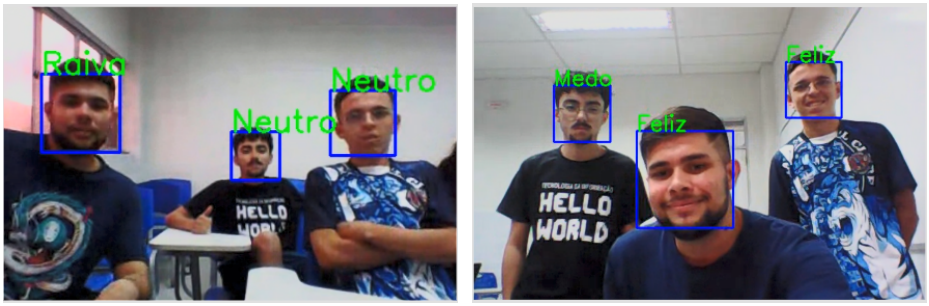
Table 2. Resultados das métricas dos modelos.

A Figura 4 mostra os resultados de reconhecimento das expressões após a aplicação do modelo no dados de teste. Já a Figura 5 mostra a aplicação do modelo em uma sala de aula. O modelo faz o reconhecimento das expressões em tempo real. O modelo foi implantado na plataforma *streamlit* para ser aplicado em tempo real. A análise desses resultados sugere que o sistema já é capaz de reconhecer as expressões faciais treinadas na maior parte do tempo. Apesar de não ser perfeito, há espaço para melhorias, onde é possível ver um cenário em que um sistema utilizando essa tecnologia ajudaria no tratamento diário das pessoas, com relação às suas emoções em situações cotidianas.

A Figura 6 ilustra a matriz de confusão obtida após o treinamento e teste do modelo. A matriz de confusão é uma tabela usada para descrever o desempenho de um modelo de classificação em um conjunto de dados. Ela permite visualizar a performance do modelo ao comparar as predições feitas pelo modelo com os valores reais conhecidos. De acordo com a matriz, pode-se observar que o modelo erra mais o reconhecimento da expressão de nojo e acerta mais a expressão feliz e triste.

A Figura 7 (a) ilustra o gráfico da função perda durante o treinamento e validação. Como uma forma de representação de que um modelo está aprendendo e se ajustando aos dados de treinamento, a função perda diminui no decorrer das épocas. Já a Figura 7 (b) ilustra o gráfico de acurácia adquirido durante o treinamento e validação do modelo. Assim como com a função de perda, a acurácia aumenta no decorrer do treinamento e validação, indicando que o modelo está melhorando suas previsões à medida que mais dados são apresentados a ele.

Figure 4. Imagens dos Resultados dos testes do modelo treinado.



(a) (b)

Figure 5. Sistema de reconhecimento de expressões sendo utilizado através do *Streamlit* em uma sala de aula: (a) reconhecimento das expressões raiva e neutro, e (b) reconhecimento das expressões medo e feliz.

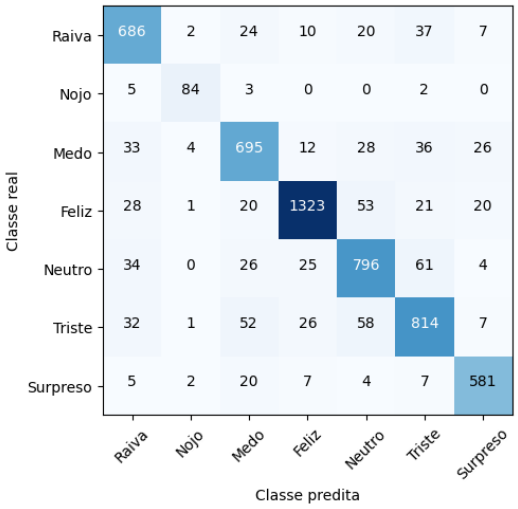
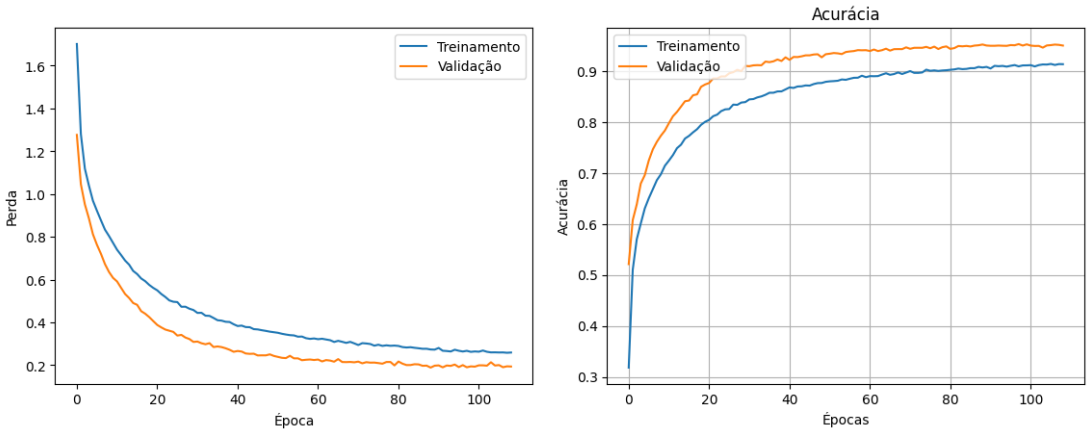


Figure 6. Matriz de Confusão do modelo treinado .



(a) Função Perda (b) Acurácia

Figure 7. Gráficos da função perda e acurácia do modelo durante as épocas.

6. Conclusão

Este estudo objetivou desenvolver um sistema para o reconhecimento de expressões faciais a partir de imagens, empregando um modelo de redes neurais convolucionais. Foi utilizado um conjunto de dados FER-2013 da plataforma KAGGLE, submetendo-o a várias etapas de pré-processamento antes do treinamento do modelo. Os resultados obtidos demonstraram uma acurácia média de 86,71% nos testes, indicando a capacidade do modelo em reconhecer expressões faciais com precisão em tempo real. Tal resultado é promissor, considerando a relevância das expressões faciais na comunicação não verbal e seu impacto em áreas como segurança e saúde mental.

References

Barsoum, E., Zhang, C., Canton Ferrer, C., and Zhang, Z. (2016). Deep learning approach for facial expression recognition. *IEEE Transactions on Affective Computing*, 7:469–480.

Chemin, A. N. (2019). Paul ekman e sua pesquisa sobre expressões faciais. LinkedIn.

Chinchanikar, N. A. (2019). Facial expression recognition using deep learning: a review. *IRJET*, 6:1–8.

Duarte, G. (2005). A página da educação. Edição 149, Outubro.

Ekman, P. Paul ekman e a descoberta das microexpressões faciais. InBody Language.

Goodfellow, I., Bengio, Y., and Courville, A. Deep learning. `https://www.deeplearningbook.org/`.

Gramado, L. F. (2010). Transtorno psiquiátrico pode afetar capacidade de identificar expressões faciais. Tilt UOL.

Hens, L. G. (2021). Detecção de emoções utilizando redes neurais convolucionais em sistemas com recursos limitados de hardware. Master's thesis, UFRGS.

Hinton, G. E. et al. (1992). *How neural networks learn from experience*. na.

Kaufman, D. (2019). *A inteligência artificial irá suplantar a inteligência humana?* Estação das letras e cores EDI.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

Liu, Z., Li, P., Li, J., and Chen, Q. (2018). Deep learning-based real-time facial expression recognition: A survey. *Journal of Visual Communication and Image Representation*, 55:539–555.

Pramerdorfer, C. and Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art. *IEEE Transactions on Affective Computing*, 7:97–108.