Aprimorando o conjunto de dados de reconhecimento de emoções faciais para análise de interação humano-robô usando aprendizado por transferência.

Bolsista Neelakshi Joshi (CTI) njoshi@cti.gov.br
Orientador: Dr. Josué JG Ramos

Resumo

Na interação humana, as emoções desempenham um papel importante. Quando aplicações de máquinas virtuais ou robôs são incorporados com capacidades de reconhecimento emocional, os humanos tendem a interagir mais com eles em comparação com outras formas de interação. A expressão facial é uma das modalidades para avaliar emoções humanas que tem atraído grande atenção no campo da visão computacional. Os esforços se concentram na construção de um modelo robusto e preciso para identificar emoções a partir de expressões faciais, permitindo a incorporação de capacidades de reconhecimento emocional em máquinas ou robôs. Obter dados de imagem ou vídeo rotulados de expressões faciais é um desafio. Para mitigar essa dificuldade, as técnicas de aprendizado por transferência podem ser úteis. Existem conjuntos de dados de imagens muito grandes e modelos profundos treinados com eles. O aprendizado por transferência permite utilizar o conhecimento adquirido por esses modelos para adaptá-los a novos conjuntos de dados, ajustando-os para aprender as sutilezas das expressões faciais. Este trabalho explora como dois desses modelos profundos podem ser treinados com dados de expressão facial para identificar emoções humanas, além de abordar as limitações e sugerir possíveis trabalhos futuros.

chave : Reconhecimento de Emoções Faciais, Aprendizado por Transferência, Aprendizado Profundo, Interação Humano-Computador, Reconhecimento de Emoções Faciais .

1. Introdução

A emoção é um elemento crucial nas interações humanas, moldando a percepção que temos das interações sociais. Com a capacidade de reconhecer emoções humanas, os robôs podem adaptar sua interação. Além disso, robôs com habilidades de expressão emocional permitem que os humanos interpretem sua interação associando-a a sinais sociais (RUTH, 2022). Portanto, para a interação prática entre humanos e robôs, a atenção se concentra na construção de robôs emocionalmente inteligentes.

A expressão facial (EF) é uma das modalidades para o estudo das emoções humanas, e seus conjuntos de dados (como imagens ou vídeos) são elementos essenciais para a construção de modelos. Um modelo de aprendizado de máquina é orientado por dados. A obtenção de grandes conjuntos de dados rotulados para reconhecimento de expressões faciais (REF) apresenta muitos desafios, como disponibilidade, acesso aberto e responsabilidades éticas.

Uma rede neural convolucional (CNN) é a arquitetura de última geração para reconhecimento de expressões faciais (FER), capaz de extrair características de alto nível de imagens e proporcionar uma alta taxa de reconhecimento. Diversos modelos de visão computacional

(CV) muito profundos estão disponíveis, treinados em conjuntos de dados de imagens gigantescos, como o ImageNet, que contém mais de 14 milhões de imagens de objetos. É possível aproveitar o conhecimento adquirido com esses modelos por meio da transferência de aprendizado (pesos dos modelos), um processo conhecido como aprendizado por transferência (TL). Vários modelos profundos diferentes estão disponíveis para TL, com arquiteturas, complexidades e tempos de inferência variados. Esses modelos são supervisionados, ou seja, treinados com dados rotulados. O TL pode ser implementado com um conjunto de dados que possua rótulos totalmente diferentes, como classes de objetos no ImageNet ou classes de emoções no FER, por meio de ajustes finos e da configuração de hiperparâmetros conforme a necessidade.

Essencialmente, aqui, os modelos pré-treinados funcionam como extratores de características, incorporando representações de baixo nível aprendidas de padrões visuais, que são transferíveis entre domínios visuais, como bordas, formas, gradientes e texturas. À medida que a rede se aprofunda, as camadas profundas aprendem diferentes características de alto nível e padrões específicos da tarefa, essenciais para distinguir diferentes classes. Para aprender variações sutis na classificação de expressões faciais, congelar a maioria das camadas pré-treinadas e reotimizar e treinar os últimos blocos, camadas profundas ou apenas a camada classificadora geralmente proporciona uma adaptação eficaz. O ajuste/calibração de hiperparâmetros, como taxa de aprendizado, otimizador, decaimento de peso e taxa de dropout, é igualmente crucial durante essa adaptação. Uma taxa de aprendizado menor pode conservar os pesos pré-treinados. O decaimento de peso e o dropout ajudam a reduzir ou evitar o sobreajuste. O aumento de dados pode aprimorar a generalização. Assim, incorporar o ajuste fino das camadas profundas e dos hiperparâmetros no aprendizado por transferência proporciona uma adaptação eficiente de grandes modelos pré-treinados a novos e diferentes conjuntos de dados rotulados.

Para tarefas de visão computacional como FER (Reconhecimento de Expressões de Expressões), o uso de TL (Aprendizado por Tempo) pode economizar muito esforço computacional e fornecer melhores resultados de classificação em comparação com o projeto e treinamento de um modelo próprio (FENG & CHASPARI, 2020). Consequentemente, as aplicações de TL estão aumentando (AKHAND et al., 2021; ENGEL et al., 2025; ALSUBAI ET AL., 2024; PAULCHAMY et al., 2025).

Este trabalho descreve como a Aprendizagem Temporal (TL) é utilizada para construir modelos de Reconhecimento de Expressões Falsas (FER) ajustados aos conjuntos de dados selecionados neste estudo. Este artigo está organizado da seguinte forma: a Seção 2 descreve o conjunto de dados, os modelos e a metodologia utilizada para construir o modelo FER. A Seção 3 apresenta a análise e seus resultados. A Seção 4 conclui o artigo discutindo as descobertas.

2. Materiais e Métodos

Esta seção apresenta os modelos pré-treinados utilizados para os bancos de dados TL e FER empregados neste trabalho, bem como a estratégia seguida para combiná-los e realizar as análises.

- O modelo de última geração *You Only Look Once* (YOLO) foi proposto por Redmon et al. (2016). Trata-se de um modelo simples e rápido, pois uma única CNN é treinada em imagens completas. A arquitetura do modelo foi inspirada no GoogLeNet

(Szegedy, 2014), onde o módulo Inception foi substituído por camadas de redução 1 × 1 seguidas por camadas convolucionais 3 × 3. Os modelos da família YOLO continuam a evoluir. Com treinamento mínimo (ajuste fino), eles se adaptam mais rapidamente a novos conjuntos de dados com rótulos totalmente diferentes. Devido à sua velocidade, precisão e tamanho, são uma escolha muito prática para inferência em tempo real. O modelo *Yolov11-cls utilizado* é um modelo de classificação desenvolvido por Jocher & Qui (2024) e disponível no Ultralytics .

- *As Redes Neurais Convolucionais Densamente Conectadas* (DenseNet; HUANG et al., 2017) aprimoraram o conceito de conexões de salto introduzido na ResNet (HE et al., 2016). Elas consistem em uma série de blocos densos. Cada bloco possui múltiplas camadas convolucionais interconectadas entre si, criando um padrão de conectividade densa. Assim, todos os blocos obtêm mapas de características concatenados reutilizáveis. Isso facilita um fluxo de gradiente mais suave durante a retropropagação, mitigando o problema do desaparecimento do gradiente e aumentando a estabilidade do treinamento. A DenseNet possui diversas versões. A *DenseNet161 utilizada* contém 4 blocos densos e 157 camadas convolucionais. Ela foi obtida a partir de modelos PyTorch com pesos ImageNet.

Neste trabalho, foram utilizadas sete bases de dados, descritas a seguir. As duas primeiras bases de dados são menores e do tipo encenado, gravadas em ambiente controlado. A primeira contém expressões faciais brasileiras, enquanto a segunda contém vistas laterais de rostos. As três seguintes são do tipo real, com imagens faciais coletadas da internet. As duas últimas são bases de dados de vídeo gravadas em ambiente controlado, e os frames de imagem foram considerados apenas para duas emoções, a fim de equilibrar o conjunto de dados combinado.

- Um banco de dados transcultural de expressões faciais (TEJADA et al., 2022) foi desenvolvido com voluntários brasileiros (60) e colombianos (50) não atores enquanto realizavam tarefas assistidas por computador. Ele contém 806 imagens para oito emoções: raiva, nojo, medo, felicidade, tristeza, surpresa e neutro, que são as sete emoções básicas, além da emoção de devassidão (ou zombaria).

- O banco de dados Karolinska Directed Emotional Faces (*KDEF*) [LUNDQVIST et al., 1998] contém expressões faciais de 70 participantes para sete emoções básicas, registradas a partir de 5 ângulos diferentes (±90°). Este conjunto de dados possui duas sessões gravadas, totalizando 4900 imagens, mas 10 imagens estão em branco, o que está documentado no banco de dados.

- MOLLAHOSSEINI et al. (2017) criaram um banco de dados *AffectNet* coletando imagens da internet. Cerca de 420 mil imagens foram anotadas manualmente, e o restante foi anotado pela rede neural ResNext, treinada com as imagens rotuladas. Um total de 291.650 imagens foram categorizadas entre as sete emoções básicas e o desprezo.

- A equipe do Kaggle coletou imagens de expressões faciais da internet para criar o banco de dados FER2013 (GOODFELLOW, 2013). Ele contém um total de 35.887 imagens para as sete emoções básicas. Posteriormente, BARSOUM et al. (2016) aprimoraram os rótulos usando métodos de crowdsourcing. A versão aprimorada, *FER+,* contém 35.272 imagens para as sete emoções básicas. Além dessas, outras categorias incluem desprezo, emoção desconhecida e ausência de expressão facial.

- Real-world Affective Faces (*RAF*) (LI et al., 2017) foi construído coletando imagens da internet. Ele possui dois conjuntos, um rotulado com sete emoções básicas e outro com 12 emoções compostas. Neste trabalho, utilizamos apenas o conjunto com as emoções básicas.
- *RAVDESS* (LIVINGSTONE & RUSSO, 2018) é um banco de dados audiovisual. *RAVDESS* foi gravado com 24 atores profissionais para as sete emoções básicas e para a expressão de emoções calmas. Neste trabalho, utiliza-se um conjunto de dados exclusivamente em vídeo, contendo fala com as sete emoções básicas.
- *O CREMA-D (* Crowd-sourced Emotional Multimodal Actors) (KEUTMANN, 2015) é um banco de dados audiovisual. O *CREMA-D* foi gravado com 91 atores de diversas origens étnicas, expressando emoções como raiva, nojo, medo, tristeza e neutras, com intensidades baixa, média, alta e não especificada.

De todos os conjuntos de dados, foram consideradas apenas as imagens com as sete emoções básicas. Começando pelos cinco primeiros bancos de dados, a Tabela 1 lista o número de imagens por emoção e mostra o desequilíbrio entre eles.

Para combinar esses bancos de dados, inicialmente cada banco de dados é distribuído em três conjuntos: treinamento (80%), validação (10%) e teste (10%). Posteriormente, eles são reorganizados de forma que o conjunto de treinamento contenha 20.645 imagens para cada uma das emoções: raiva, felicidade, neutro, tristeza e surpresa. Para isso, as imagens foram escolhidas aleatoriamente quando o conjunto de treinamento correspondente continha mais de 20.645 imagens; caso contrário, o número necessário de imagens foi transferido dos conjuntos de validação e teste para o conjunto de treinamento. Por exemplo, no AffectNet, 20.645 imagens foram escolhidas aleatoriamente para a emoção de felicidade, enquanto para a emoção de surpresa, 100 imagens foram transferidas dos conjuntos de validação e teste para o conjunto de treinamento. Uma atenção especial foi dada ao banco de dados KDEF. Para evitar vazamento de dados, as imagens de ambas as sessões do mesmo ator foram mantidas juntas em um único conjunto. Assim, os atores foram escolhidos aleatoriamente em vez das imagens para serem distribuídas entre os três conjuntos.

	raiva	nojo	medor	felicidade	neutro	tristeza	surpresa
Transcultural	91	98	96	106	110	99	103
KDEF	700	694	699	700	698	700	699
AfectNet	25382	4303	6878	134915	75374	25959	14590
FER+	3111	248	819	9355	12906	4371	4462
RAF	867	877	355	5957	3204	2460	1619
total	*30151*	*6220*	*8847*	*151033*	*92292*	*33589*	*21473*

Fonte: Produção do autor

Tabela 1 – Total de imagens de expressões faciais por emoção nos bancos de dados

A quantidade de imagens para as emoções de nojo e medo é tão baixa que, para atingir o número necessário, as imagens foram extraídas dos bancos de dados RAVDESS e CREMA-D, selecionando-se frames com frequência de um décimo de frames por segundo dos vídeos. De cada vídeo, os seis primeiros e os quatro últimos frames foram descartados para garantir a captura apenas da expressão emocional desejada. Para balancear o conjunto de

validação, um número necessário de imagens foi transferido do conjunto de teste para o conjunto de validação. Com esse procedimento, obtivemos o conjunto de dados combinado balanceado com 20.645, 700 e 128 imagens para cada emoção nos conjuntos de treinamento, validação e teste, respectivamente.

3. Análise e Resultados

Para a análise, foi utilizado o modelo nano Yolov11-cls com pesos pré-treinados no ImageNet, executado em um servidor com sistema operacional Ubuntu 24.04.2 LTS e GPU (CUDA 12.9). O ambiente foi construído em Python 3.11. Duas estratégias de otimização foram avaliadas, ajustando-se as camadas profundas e variando-se os hiperparâmetros. Os melhores resultados foram obtidos com

(a) Otimizador de descida de gradiente estocástico (SGD) com momento de 0,93, taxa de aprendizado inicial de 0,003, decaimento de peso de 1×10^{-5} $^{-5}$e dropout = 0,5. A estratégia de aumento de dados padrão do algoritmo foi mantida. Apenas a última camada (décima) foi descongelada para treinamento, a fim de aprender novas classes de emoção.

(b) Otimizador AdamW com taxa de aprendizado de 0,001, decaimento de peso de 1×10^{-5} $^{-6}$, dropout = 0,5 e estratégia de aumento de dados padrão. As três últimas camadas do modelo foram...

Both experiments were run for 50 epochs with a batch size of 64. All images were resized to 224✕224 pixels.

Com o modelo DenseNet161, a análise foi realizada em um servidor com sistema operacional Ubuntu 20.04.6 LTS e GPU (CUDA 11.4). O Docker foi construído em Python 3.7 e PyTorch. O modelo foi pré-treinado no ImageNet e adotado como extrator de características. Para otimizá-lo para a tarefa de reconhecimento de expressões faciais (FER), sua camada de classificação final foi substituída por uma nova camada de cabeçalho de duas camadas, composta por uma camada de dropout (p=0,4) para reduzir o sobreajuste, seguida por uma camada totalmente conectada mapeando vetores de características de 2208 dimensões para 7 classes de emoção.

O desempenho ideal foi alcançado usando o otimizador SGD com momentum de 0,9, uma taxa de aprendizado inicial de 1×10^{-5} $^{-4}$e uma taxa de decaimento de peso de 5×10^{-5} $^{-5}$. O treinamento empregou a função de perda de Entropia Cruzada e o agendador de taxa de aprendizado por recozimento de cosseno. Os dois últimos blocos densos foram ajustados por 30 épocas com um tamanho de lote de 64. Todas as imagens de entrada foram redimensionadas para 224×224 pixels para atender aos requisitos de entrada do modelo. A técnica de aumento de dados incluiu inversão horizontal aleatória e uma pequena variação de cor para melhorar a generalização e reduzir o sobreajuste. A inversão horizontal aplica uma transformação geométrica que inverte a imagem ao longo de seu eixo vertical com uma probabilidade dada (usada 0,5). A variação de cor modifica o brilho, o contraste, a saturação e a tonalidade aleatoriamente com uma probabilidade dada (usada 0,2).

modelo	otimizador	taxa de aprendizagem inicial	impulso	perda de peso	cair fora	camadas/blocos descongelados
Yolov11	Adam W	0,001	--	1e-6	0,5	últimas três camadas
Yolov11	SGD	0,003	0,93	1e-5	0,5	última camada
DenseNet161	SGD	0,0001	0,90	5e-5	0,4	últimos dois quarteirões

Fonte: Criação do autor

Tabela 2 – configuração dos modelos

Com todas essas configurações de modelos, apresentadas na Tabela 2, o treinamento foi realizado. O conjunto de teste foi avaliado com o melhor modelo salvo. A Tabela 2 apresenta as métricas da análise e a avaliação do conjunto de teste. Além disso, as Figuras 1 e 3 apresentam o desempenho do treinamento dos modelos Yolov11 com (a) otimizador SGD, (b) otimizador AdamW e modelo DenseNet161 com otimizador SGD, respectivamente. As Figuras 2 e 4 mostram a avaliação do conjunto de teste com matriz de confusão: (a) Yolov11 SGD, (b) Yolov11 AdamW e com modelo DenseNet161.

modelo	otimizador	val_acc	teste_acc	pontuação test_F1	tempo_de_inferência/imagem
Yolov11	Adam W	70,90%	71,00%	70,00%	0,0013 segundos
Yolov11	SGD	66,17%	67,00%	67,00%	0,0011 segundos
DenseNet161	SGD	69,00%	69,00%	69,00%	0,0100 segundos

Fonte: Criação do autor

Tabela 3 – Resultados da avaliação dos modelos nos conjuntos de validação e teste

(a) (b)

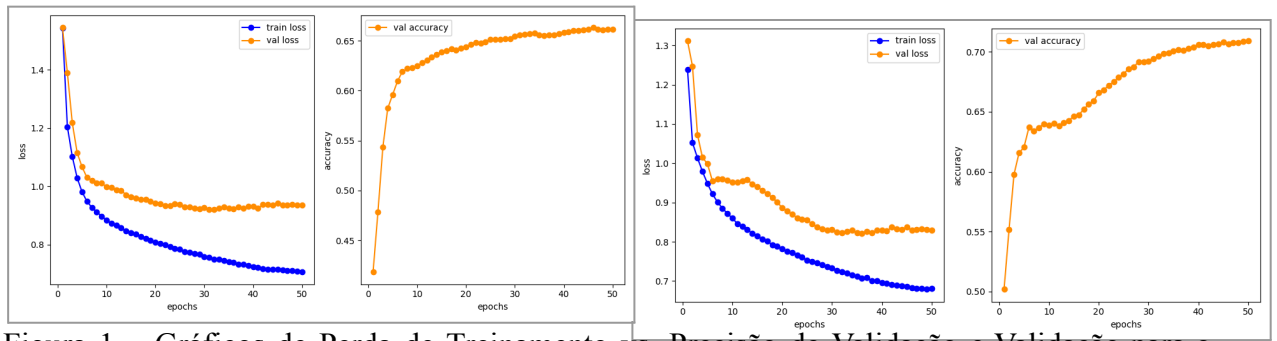


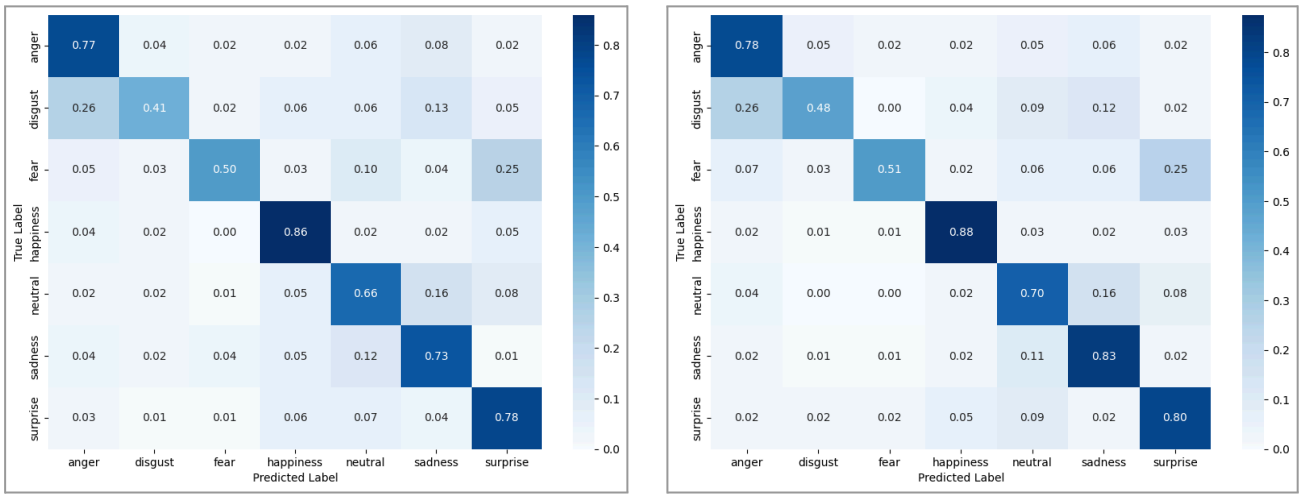
Figura 1 – Gráficos de Perda de Treinamento vs. Precisão de Validação e Validação para o modelo Yolov11 com otimizadores (a) SGD e (b) AdamW

Os gráficos de perda e acurácia do treinamento do modelo apresentados nas Figuras 1 e 3 mostram que a perda diminuiu gradualmente enquanto a acurácia aumentou. Ao final do treinamento, atingiu-se um platô e, como a perda de validação seguiu a perda de treinamento

com um valor ligeiramente maior, isso sugere que os modelos foram treinados de forma otimizada.

A maior pontuação F1 nos testes, de 70%, foi obtida com o modelo Yolov11 com o otimizador AdamW. O segundo melhor resultado foi obtido com o modelo DenseNet, com 69%, e com o Yolov11 SGD, com 67%. O tempo de inferência demonstra a rapidez do modelo Yolo. Ele levou, em média, apenas 1,1 a 1,3 ms por imagem, enquanto o modelo DenseNet161 levou 10 ms para avaliar a mesma imagem.

(a) (b)



igura 2 – Matriz de Confusão da avaliação do conjunto de teste com o modelo Yolov11 com otimizadores (a) SGD e (b) AdamW

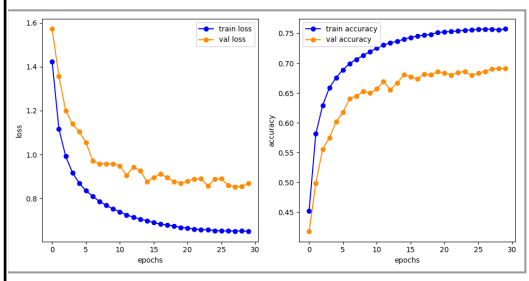


Figura 3 – Gráfico de Perda e Precisão de Treinamento vs. Validação com

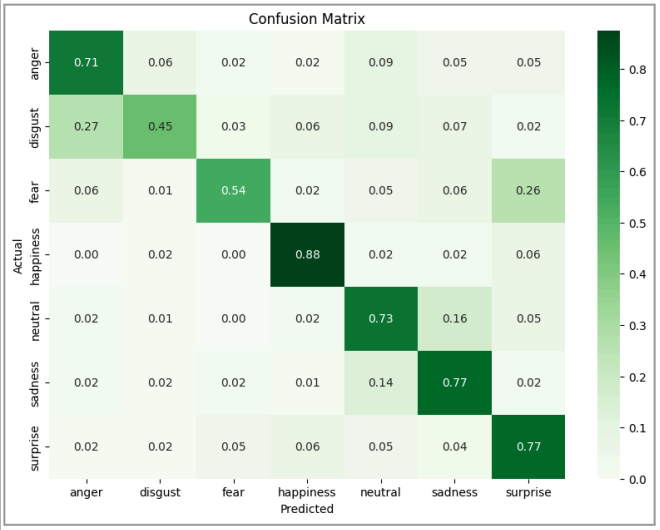


Figura 4 – Matriz de Confusão da avaliação do conjunto de teste com o modelo DenseNet161 (SGD)

A matriz de confusão mostra que todas as três configurações de modelos conseguiram reconhecer a emoção de felicidade com mais frequência (>86%). Em seguida, a taxa de

reconhecimento diminui gradualmente para as emoções de tristeza, surpresa, raiva e neutra. A menor taxa de reconhecimento foi obtida para as emoções de nojo e medo.

4. Discussão e Conclusão

Este trabalho apresenta experimentos de modelagem para reconhecimento de expressões faciais, análises realizadas e resultados obtidos. Para a análise, encontramos apenas um conjunto de dados de reconhecimento de expressões faciais (FER) com expressões faciais brasileiras, e este é muito pequeno. Identificar expressões a partir da vista lateral (ou de perfil) é um desafio e constitui um campo de pesquisa de interesse. Assim, o conjunto de dados KDEF foi considerado, embora também seja pequeno e do tipo simulado. Portanto, para obter uma variedade de características faciais, foram escolhidos conjuntos de dados do mundo real. A Tabela 1 lista o número de imagens por emoção incluídas nos bancos de dados utilizados. O conjunto de dados apresenta um alto grau de desbalanceamento. Para evitar o viés de desbalanceamento de classes, a estratégia de distribuição de dados entre os três conjuntos e o procedimento de aumento de dados implementado, que consiste na adição de imagens do conjunto de dados de vídeo, são descritos na seção 2.

Para as análises, a técnica de Aprendizado por Tempo (TL) foi utilizada com dois modelos de rede profunda: Yolov11 e DenseNet161. O ajuste fino e os hiperparâmetros escolhidos são descritos na seção 3 e listados na Tabela 2. A pontuação F1 da avaliação do conjunto de teste com ambos os modelos é próxima: 70% para o Yolov11 com o otimizador AdamW e 69% para o modelo DenseNet161. Os modelos da série Yolo são reconhecidos por sua velocidade e precisão, o que se reflete nos tempos de inferência apresentados na Tabela 3. O tempo médio de inferência com o modelo Yolov11 é de 1,2 ms, enquanto o modelo DenseNet requer 10 ms para avaliar a mesma imagem nova e não vista anteriormente. Isso sugere que o modelo Yolov11 é mais adequado para avaliação e aplicação de Reconhecimento de Expressões Forenses (FER) em tempo real.

As Figuras 2 e 4 apresentam uma matriz de confusão da avaliação do conjunto de teste. A maior taxa de reconhecimento foi observada para a emoção de felicidade, onde 88% das imagens identificaram corretamente a emoção. Embora tenhamos tentado reduzir o viés devido ao desequilíbrio de classes no conjunto de dados, a matriz de confusão mostra um cenário diferente. A taxa de reconhecimento para as emoções de nojo e medo é muito baixa. Entre as cinco emoções – raiva, felicidade, neutro, tristeza e surpresa – cujas imagens foram obtidas de cinco bancos de dados diferentes, a menor taxa de reconhecimento foi de 70%.

Agora, passando para as duas emoções restantes, nojo e medo, onde o equilíbrio de classes foi alcançado pela extração de frames dos conjuntos de dados de vídeo, a taxa de reconhecimento é muito baixa em comparação com as outras cinco emoções. Isso implica que, embora o equilíbrio de classes seja alcançado pela extração de frames do conjunto de dados de vídeo, ele introduz diferentes tipos de vieses. Embora cada vídeo tenha fornecido muitas imagens, estas podem ter apresentado falta de diversidade em aparências faciais, intensidade de expressão, pose e iluminação. Isso pode ter causado sobreajuste no modelo e, portanto, resultado em uma baixa taxa de reconhecimento. Outra observação é a classificação incorreta de emoções: medo é classificado erroneamente como surpresa e nojo como raiva. Isso é atribuído à similaridade entre as classes, visto que essas emoções compartilham algumas características comuns, como unidades de ação (DU et al., 2014).

Para estudar a associação entre proximidade social e emoções humanas durante a interação humano-robô, este modelo pode servir como ponto de partida, especialmente para cinco

emoções, com exceção de nojo e medo. Trabalhos futuros investigarão possíveis maneiras de superar as limitações do presente estudo. Isso inclui obter mais imagens de expressões faciais de brasileiros, optar por emoções mais complexas ou menos intensas em vez de usar emoções básicas e aprimorá-las.

Referências

AKHAND, MAH. ROY, S. SIDDIQUE, N. KAMAL, MAS. & SHIMAMURA, T. *Reconhecimento de emoções faciais usando aprendizado por transferência na CNN profunda* . Electronics 2021, 10, 1036.

ALSUBAI, S. ALQAHTANI, A. ALANAZI, A. SHA, M. e GUMAEI A. *Reconhecimento de emoções faciais usando aprendizado profundo quântico e mecanismo avançado de transferência de aprendizado* . Front. Comput. Neurosci. 18:1435956, 2024.

BARSOUM, E. ZHANG, C. CANTON FC. & ZHANG, Z. *Treinamento de redes profundas para reconhecimento de expressões faciais com distribuição de rótulos colaborativa* . Conferência Internacional da ACM sobre Interação Multimodal, pp. 279-283, 2016.

D'ANGELO, F. ANDRIUSHCHENKO, M. VARRE, AV. e FLAMMARION, N. *Por que precisamos de decaimento de peso no aprendizado profundo moderno?* Advances in Neural Information Processing Systems, 37, 23191-23223, 2024.

DU, S. TAO, Y. & MARTINEZ, AM. *Expressões faciais compostas de emoção* . Anais da academia nacional de ciências, 111(15), E1454-E1462, 2014.

ENGEL, E. LI, L. HUDY, C. & SCHLEUSNER, R. *Aprendizagem por transferência multimodal para reconhecimento dinâmico de emoções faciais em ambientes reais.* arXiv preprint arXiv:2504.21248, 2025.

FENG, K. & CHASPARI, T. *Uma revisão da aprendizagem por transferência generalizável no reconhecimento automático de emoções* . Frontiers in Comp. Sc. Vol 2, artigo 9, 2020. doi: 10.3389/fcomp.2020.00009

GOODFELLOW, IJ. ERHAN, D. & CARRIER, PL. *Desafios na aprendizagem de representações: um relatório sobre três competições de aprendizado de máquina* . In: Anais da Conferência Internacional de Processamento de Informação Neural. Berlim, Alemanha: Springer, pp. 117–124, 2013.

HE, K. ZHANG, X. REN, S. SUN, J. *Aprendizado Residual Profundo para Reconhecimento de Imagens* . In Anais da Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões (CVPR) de 2016, pp. 770–778, 2016.

HUANG, G. LIU, Z. VAN DER MAATEN, L. WEINBERGER, KQ. *Redes convolucionais densamente conectadas* . In Anais da Conferência IEEE sobre Visão Computacional e Reconhecimento de Padrões (CVPR), pp. 2261–2269, 2017.

JOCHER, G. & QIU, J. *Ultralytics YOLO11* . Versão 11.0.0, 2024. https://github.com/ultralytics/ultralytics

KEUTMANN, MK. MOORE, SL. SAVITT, A. & GUR, RC. *Geração de um banco de itens para pesquisa translacional em cognição social: metodologia e validação inicial* . Behavior research methods, 47(1), 228-234, 2015. https://www.ncbi.nlm.nih.gov/pubmed/24719265

LI, S. DENG, W. & DU, J. *Aprendizado colaborativo confiável e preservação de localidade profunda para reconhecimento de expressões em ambientes reais* . Computer Vision and Pattern Recognition (CVPR), 2584--2593, 2017.

LIVINGSTONE, SR. & RUSSO FA. *O Banco de Dados Audiovisual Ryerson de Fala Emocional e Canto (RAVDESS): Um conjunto dinâmico e multimodal de expressões faciais e vocais no inglês norte-americano* . PLoS ONE 13(5): e0196391, 2018. doi:10.1371/journal.pone.0196391

LUNDQVIST, D. FLYKT, A. & ÖHMAN, A. *Karolinska Directed Emotional Faces - KDEF* , CD-ROM do Departamento de Neurociência Clínica, Seção de Psicologia, Instituto Karolinska, 1998 ISBN 91-630-7164-9.

MOLLAHOSSEINI, A. HASANI, B. & MAHOOR, MH. *AffectNet: Um banco de dados para computação de expressão facial, valência e ativação em situações reais.* Transações IEEE em Computação Afetiva PP:99-1 , 20 17.

PAULCHAMY, B. YAHYA, A. CHINNASAMY, N. & KASILINGAM, K. *Reconhecimento de expressão facial por meio de aprendizado por transferência: integração de VGG16, ResNet e AlexNet com um classificador multiclasse* . Acadlore Trans AI Mach Learn, 4(1), 25-39, 2025.

REDMON, J. DIVVALA, S. GIRSHICK, R. & FARHADI, A. *Você só olha uma vez: Detecção unificada de objetos em tempo real.* Em Anais da conferência IEEE sobre visão computacional e reconhecimento de padrões, pp. 779-788, 2016.

RUTH, SH. *Levantamento de emoções em interações humano-robô: perspectivas da psicologia robótica sobre 20 anos de pesquisa* . International Journal of Social Robotics 14:389–41, 2022, doi:10.1007/s12369-021-00778-6

SZEGEDY, C. LIU, W. JIA, Y. SERMANET, P. REED, S. ANGUELOV, D. ERHAN, D. VANHOUCKE, V. & RABINOVICH, A. *Indo mais fundo nas convoluções* . CoRR, abs/1409.4842, 2014.

TEJADA, J. FREITAG, RMK. PINHEIRO, BFM. CARDOSO, P.B. SOUZA, VRA. & SILVA, LS. *Construção e validação de um conjunto de imagens de expressões faciais para detecção de emoções: um estudo transcultural* . Pesquisa Psicológica, vol. 86, nº 6, pp. 1996–2006, 2022, doi:10.1007/s00426-021-01605-3