Reconhecimento de Expressões Faciais em Tempo Real com YOLO: Estudo Comparativo em Cenários Não Controlados

Paulo Ricardo[1], **Tiago de Melo**[1]

Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brazil

{prfds.eng21,tmelo}@uea.edu.br

Abstract. *Real-time facial expression recognition in unconstrained (in-the-wild) scenarios remains challenging due to illumination, head pose, and occlusions. This paper compares YOLOv8n, YOLOv10n, YOLOv11n, and YOLOv12n, all in the nano variant, for facial expression recognition using the public "9 Facial Expressions for YOLO" dataset under a unified experimental protocol. The results reveal clear mAP–speed trade-offs: YOLOv12n achieved the highest mAP, YOLOv10n the fastest inference, and YOLOv8n the best balance for real-time applications under computational constraints.*

Resumo. *O reconhecimento de expressões faciais em tempo real em cenários não controlados (in-the-wild) ainda é desafiador devido a variações de iluminação, pose e oclusões. Este trabalho compara YOLOv8n, YOLOv10n, YOLOv11n e YOLOv12n, todas na variante nano, para reconhecimento de expressões faciais usando o dataset público "9 Facial Expressions for YOLO" sob um protocolo experimental unificado. Os resultados evidenciam trade-offs entre mAP e velocidade: o YOLOv12n obteve maior mAP, o YOLOv10n maior FPS e o YOLOv8n o melhor equilíbrio para aplicações em tempo real sob restrições computacionais.*

1. Introdução

O reconhecimento de expressões faciais é uma área central em visão computacional e computação afetiva, com aplicações que vão desde a interação humano-computador e sistemas assistivos até contextos de segurança, saúde e pesquisa psicológica [Alshammari and Alshammari 2024]. Em cenários práticos, a interpretação automática de emoções a partir do rosto pode apoiar interfaces mais naturais e sistemas de monitoramento voltados à análise de atenção, estresse e comportamento. Entretanto, apesar dos avanços recentes, alcançar um desempenho consistente fora de ambientes controlados ainda é um desafio relevante.

Abordagens tradicionais para reconhecimento de expressões faciais baseavam-se na extração manual de características, explorando informações pré-definidas, como contornos e padrões de textura [Kumari et al. 2015]. A transição para métodos capazes de aprender representações diretamente a partir de exemplos visuais, abordagem conhecida como aprendizado profundo (*deep learning*), marcou um avanço significativo. Mesmo assim, a utilização robusta em cenários do mundo real ainda impõe obstáculos consideráveis, principalmente devido a variações de iluminação, pose de cabeça, expressões sutis e oclusões parciais, que exigem alta capacidade de generalização dos modelos [Li and Deng 2018].

Um dos fatores que historicamente limitou o progresso nessa área foi a disponibilidade de bases de dados de treinamento em escala suficiente para capturar a diversidade encontrada em ambientes não controlados [Mollahosseini et al. 2017]. Nesse contexto, a disponibilização recente de *datasets* com dezenas de milhares de imagens coletadas em cenários naturais representa uma mudança importante. Ainda assim, dados em grande escala não eliminam por completo as dificuldades de implantação, como a variabilidade real das imagens e as restrições computacionais das aplicações interativas, que impõem um segundo desafio ligado à eficiência.

De fato, mesmo quando o modelo apresenta boa acurácia, a aplicabilidade em sistemas reais costuma exigir inferência em tempo real, o que impõe limitações severas de processamento e memória [Shao and Qian 2019]. Entre as arquiteturas de estágio único (*one-stage*), a família *You Only Look Once* (YOLO) [Redmon et al. 2016] consolidou-se como referência em tarefas que exigem rapidez, oferecendo uma alternativa atraente para integrar detecção facial e classificação de expressões com menor custo computacional, favorecendo aplicações em tempo real.

Nesse contexto, este trabalho investiga o uso da arquitetura YOLO para o reconhecimento de expressões faciais em cenários não controlados. Para isso, utiliza-se o *dataset* público "9 Facial Expressions for YOLO", disponível no Kaggle, composto por aproximadamente 68 mil imagens coletadas em ambientes reais (*in-the-wild*) e organizado em nove categorias emocionais. Diferentemente de bases obtidas em ambientes controlados, esse tipo de conjunto de dados apresenta grande variabilidade de iluminação, pose, oclusões parciais e expressões espontâneas, tornando a tarefa de reconhecimento significativamente mais desafiadora.

Os resultados experimentais indicam que os modelos avaliados apresentam desempenho competitivo, mas com *trade-offs* claros entre acurácia, velocidade de inferência e custo computacional. O YOLOv12n obteve o melhor desempenho em termos de mAP, enquanto o YOLOv10n apresentou a maior taxa de inferência. Considerando o equilíbrio entre precisão e eficiência, o YOLOv8n destacou-se como uma alternativa adequada para aplicações em tempo real sob restrições computacionais.

As principais contribuições deste trabalho podem ser resumidas em três aspectos: (i) a avaliação comparativa entre diferentes versões recentes da família YOLO aplicadas ao reconhecimento de expressões faciais; (ii) a análise do desempenho desses modelos em um conjunto de dados *in-the-wild*, que reflete condições mais próximas de aplicações reais; e (iii) a investigação conjunta de métricas de acurácia e eficiência computacional, visando identificar arquiteturas capazes de equilibrar desempenho e velocidade para aplicações interativas.

2. Trabalhos Relacionados

A arquitetura YOLO, apresentada inicialmente por [Redmon et al. 2016], consolidou-se como uma das principais abordagens para detecção de objetos em visão computacional. Diferentemente de métodos tradicionais de múltiplos estágios, o YOLO realiza detecção e classificação simultaneamente em uma única passada pela rede, permitindo inferência em tempo real com elevado desempenho. Ao longo dos anos, diversas versões da arquitetura foram propostas com o objetivo de melhorar precisão, velocidade e eficiência computacional. Segundo [Jegham et al. 2025], modelos recentes, como YOLOv8, YO-

LOv10, YOLO11 e YOLO12, incorporam melhorias arquiteturais e estratégias avançadas de detecção, mantendo desempenho adequado para aplicações em tempo real.

No reconhecimento de expressões faciais, abordagens de aprendizado profundo mostram alta capacidade de extrair automaticamente características discriminativas [Li and Deng 2018]. Porém, o reconhecimento em ambientes não controlados ainda é desafiador devido a variações de iluminação, pose, oclusões e diferenças individuais na expressão emocional [Mollahosseini et al. 2017]. Trabalhos recentes exploram a família YOLO nessa tarefa, reforçando sua viabilidade em computação afetiva e inferência em tempo real [Alshammari and Alshammari 2024]. O dataset público "9 Facial Expressions for YOLO" também vem sendo usado em estudos com CNN e YOLOv12 [Adimas et al. 2025]. Apesar dos resultados promissores, faltam comparações sistemáticas entre versões recentes da família YOLO sob um mesmo protocolo experimental. Assim, este trabalho compara YOLOv8n, YOLOv10n, YOLOv11n e YOLOv12n em cenários *in-the-wild*, considerando simultaneamente precisão, *recall*, mAP e desempenho em tempo real.

3. Materiais e Métodos

3.1. Conjunto de Dados

O conjunto de dados utilizado neste trabalho é o *"9 Facial Expressions for YOLO"*, disponibilizado publicamente na plataforma Kaggle[1]. Esse *dataset* foi preparado especificamente para modelos da família YOLO, contendo imagens anotadas no formato nativo da arquitetura, com rótulos e coordenadas normalizadas das caixas delimitadoras. As imagens estão nos formatos JPG/PNG e foram coletadas em condições *in-the-wild*, apresentando variações de iluminação, pose, oclusões parciais e diversidade demográfica.

O conjunto contempla nove classes de expressão facial: raiva (*angry*), desprezo (*contempt*), nojo (*disgust*), medo (*fear*), felicidade (*happy*), neutro (*natural*), tristeza (*sad*), sonolência (*sleepy*) e surpresa (*surprised*). A divisão em treino, validação e teste foi mantida conforme disponibilizada pelo autor do conjunto de dados, de modo a preservar o protocolo original e garantir a comparabilidade interna entre todas as arquiteturas avaliadas. No formato disponibilizado, o conjunto é composto por 64.866 imagens de treino (94,99%), 1.720 imagens de validação (2,52%) e 1.700 imagens de teste (2,49%).

Embora essa proporção seja menos convencional do que particionamentos como 70/15/15 ou 80/10/10, optou-se por não redefinir a divisão do conjunto, uma vez que o objetivo central deste trabalho é realizar uma análise comparativa entre diferentes versões da arquitetura YOLO sob um mesmo protocolo experimental. Além disso, analisou-se a distribuição de anotações por classe, observando-se desbalanceamento relevante: classes como *happy*, *angry* e *sad* possuem mais de 10 mil instâncias, enquanto *contempt* e *sleepy* apresentam número significativamente menor, o que pode aumentar confusões e reduzir o desempenho em categorias minoritárias.

3.2. Métricas de avaliação

A avaliação considera simultaneamente métricas de acurácia e eficiência computacional. Em tarefas de detecção de objetos, métricas como IoU, Precision, Recall, AP e mAP são

[1] `https://www.kaggle.com/datasets/aklimarimi/8-facial-expressions-for-yolo`.

amplamente empregadas para quantificar a qualidade da localização e da classificação das detecções, enquanto medidas como tempo de inferência, FPS e custo computacional auxiliam na análise da viabilidade de uso em tempo real [Redmon et al. 2016, Alshammari and Alshammari 2024].

Neste trabalho, reportam-se Precision, Recall, mAP@50 e mAP@50–95. O mAP@50 considera o limiar de IoU igual a 0,50, enquanto o mAP@50–95 calcula a média das APs para limiares de IoU entre 0,50 e 0,95, em passos de 0,05. Para caracterizar a eficiência, foram considerados o tempo de inferência (ms/imagem), FPS (*frames per second*) e número de parâmetros. O FPS foi estimado executando 100 inferências com *batch*=1, utilizando uma imagem fixa do conjunto de validação e desconsiderando o tempo de carregamento do modelo.

3.3. Modelos Utilizados

A família YOLO é amplamente utilizada em tarefas de detecção de objetos em tempo real devido à sua alta eficiência e por realizar localização e classificação em uma única passagem pela rede [Redmon et al. 2016]. Neste trabalho, adotou-se uma comparação sistemática entre quatro versões recentes da arquitetura, YOLOv8n, YOLOv10n, YOLOv11n e YOLOv12n, todas na variante *nano* (n), selecionada por apresentar baixo custo computacional e maior viabilidade para inferência em tempo real.

Ao restringir a análise às variantes *nano* e manter um protocolo experimental unificado, busca-se avaliar de forma justa como as mudanças introduzidas ao longo das versões impactam o trade-off entre acurácia e eficiência na tarefa de reconhecimento de expressões faciais em cenários *in-the-wild*, utilizando o dataset *"9 Facial Expressions for YOLO"*.

4. Resultados

Os experimentos foram conduzidos utilizando a divisão original do dataset *"9 Facial Expressions for YOLO"*, composta por 64.866 imagens para treinamento, 1.720 imagens para validação e 1.700 imagens para teste. Essa divisão foi mantida para preservar a consistência com o protocolo do conjunto de dados e garantir reprodutibilidade. Todos os modelos foram treinados sob o mesmo protocolo experimental, com 45 épocas, resolução de entrada de 640 *pixels* e *batch size* igual a 16.

O treinamento e a inferência foram realizados em ambiente controlado, utilizando a biblioteca Ultralytics YOLO (versão 8.3.x), executando em Python 3.10.14 e PyTorch 2.4.0. A GPU empregada foi uma NVIDIA Tesla T4, com 15 GB de memória. A adoção de um mesmo protocolo experimental buscou favorecer comparações mais justas entre as arquiteturas avaliadas.

Os resultados quantitativos obtidos para as quatro versões avaliadas, YOLOv8n, YOLOv10n, YOLOv11n e YOLOv12n, são apresentados na Tabela 1. Observa-se que todas as arquiteturas alcançaram valores elevados de mAP@50, indicando boa capacidade de reconhecimento das expressões faciais mesmo em cenários *in-the-wild*. O YOLOv12n apresentou os maiores valores de mAP@50 e mAP@50–95, refletindo ganhos de precisão. Em contrapartida, o YOLOv10n destacou-se pela maior velocidade de inferência, atingindo 73,85 FPS. O YOLOv8n apresentou um bom equilíbrio entre acurácia e eficiência computacional, com 56,17 FPS e mAP@50–95 de 68,1%, configurando-se

como uma alternativa atrativa para aplicações em tempo real sob restrições de processamento.

Modelo	Parâmetros	Treino (h)	P(%)	R(%)	mAP@50(%)	mAP@50–95(%)	FPS	Inferência (ms)
YOLOv8n	3.01M	8.99	77.5	79.2	83.7	68.1	56.17	17.8
YOLOv10n	2.78M	10.70	79.1	77.6	82.7	66.8	73.85	13.5
YOLOv11n	2.64M	8.79	75.1	79.0	82.5	67.3	62.42	16.0
YOLOv12n	2.60M	14.63	78.5	81.4	83.9	68.6	47.11	21.2

Tabela 1. Resultados gerais obtidos para as diferentes versões da família YOLO, incluindo custo de treinamento e desempenho de inferência.

O tempo de treinamento variou conforme a versão do modelo: o YOLOv11n teve o menor tempo, enquanto o YOLOv12n demandou o maior, com 14,63 horas. Isso indica que evoluções arquiteturais recentes, embora gerem pequenos ganhos de desempenho, podem aumentar o custo de treinamento e a latência de inferência.

A análise por classe do YOLOv8n, modelo que apresentou o melhor equilíbrio entre acurácia e eficiência computacional, revelou desempenho consistente nas expressões mais frequentes do conjunto de dados, como *happy*, *angry* e *sad*. Em contrapartida, categorias como *contempt* e *natural* apresentaram maior dificuldade de reconhecimento, reflexo tanto do desbalanceamento observado na distribuição das anotações quanto da semelhança visual entre algumas expressões.

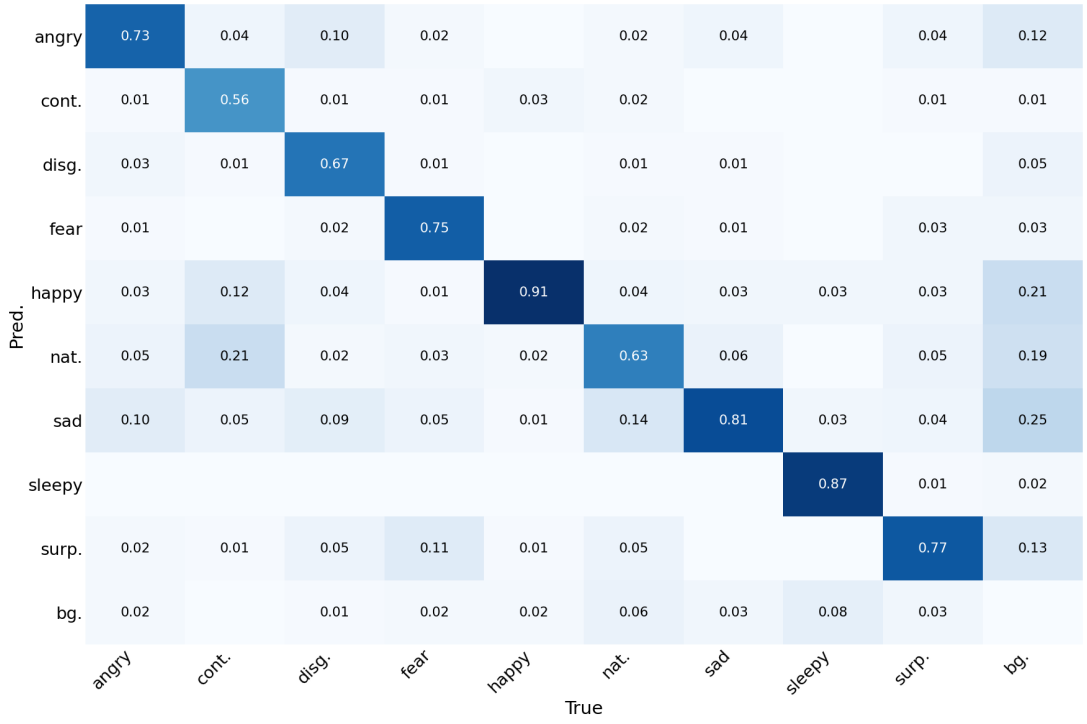

Figura 1. Matriz de confusão normalizada do modelo YOLOv8n no teste.

A Figura 1 apresenta a matriz de confusão normalizada obtida pelo modelo YOLOv8n no conjunto de teste. Observa-se que a maior parte das predições corretas concentra-se na diagonal principal da matriz, indicando boa capacidade de distinção entre as diferentes expressões faciais. Por outro lado, algumas confusões são observadas entre

classes com características visuais semelhantes, como *contempt* e *natural*, além de ambiguidades entre *angry* e *disgust*. De forma geral, os resultados indicam que a escolha do modelo mais adequado depende do cenário de aplicação: enquanto o YOLOv12n prioriza precisão, o YOLOv10n favorece velocidade e o YOLOv8n oferece melhor equilíbrio para aplicações em tempo real.

5. Conclusão

Este trabalho apresentou uma análise comparativa entre as versões YOLOv8n, YOLOv10n, YOLOv11n e YOLOv12n aplicadas ao reconhecimento de expressões faciais em cenários não controlados, sob um protocolo experimental unificado. Os resultados indicaram *trade-offs* claros entre acurácia, velocidade de inferência e custo computacional de treinamento: o YOLOv12n obteve o melhor desempenho em termos de mAP (mAP@50=0,839; mAP@50–95=0,686), enquanto o YOLOv10n apresentou a maior taxa de inferência (73,85 FPS). Considerando o equilíbrio entre precisão e eficiência, o YOLOv8n destacou-se como uma alternativa adequada para aplicações em tempo real sob restrições computacionais (mAP@50–95=0,681; 56,17 FPS).

A análise por classe evidenciou o impacto do desbalanceamento do dataset: expressões majoritárias, como *happy*, *angry* e *sad*, apresentaram desempenho superior, enquanto classes minoritárias e mais sutis, como *contempt*, exibiram maior taxa de confusão. Esses resultados sugerem que parte das limitações observadas decorre não apenas das arquiteturas, mas também das características intrínsecas do conjunto de dados. Como trabalhos futuros, pretende-se integrar o modelo selecionado em uma aplicação de inferência em tempo real e investigar estratégias para mitigar o desbalanceamento entre classes.

Referências

Adimas, R., Firmansyah, G., and Ardiansyah, A. (2025). Real-time facial expression detection using yolov12. *Journal of Knowledge and Technology Innovation*, 9(2):1–9.

Alshammari, A. and Alshammari, M. E. (2024). Emotional facial expression detection using YOLOv8. *Engineering, Technology & Applied Science Research*, 14(5):16619–16623.

Jegham, N., Koh, C. Y., Abdelatti, M. F., and Hendawi, A. (2025). Yolo evolution: A comprehensive benchmark and architectural review of yolov12, yolo11, and their previous versions. Preprint.

Kumari, J., Rajesh, R., and Pooja, K. M. (2015). Facial expression recognition: A survey. *Procedia Computer Science*, 58:486–491.

Li, S. and Deng, W. (2018). Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*.

Shao, J. and Qian, Y. (2019). Three convolutional neural network models for facial expression recognition in the wild. *Neurocomputing*, 355:82–92.