Object Recognition of School Supplies via Convolutional Neural Networks for Accessibility in the Educational Environment

Giulia C. Bezerra, João Victor M. Vantil e Letícia T. M. Zoby

Instituto de Educação Superior de Brasília (IESB)
– Brasília – DF – Brasil

{joao.vantil, giulia.bezerra,leticia.zoby}@iesb.edu.br

Abstract. *This paper presents a proof of concept that employs Convolutional Neural Networks (CNNs) to promote accessibility for visually impaired students in educational environments. By using YOLOv5 for real-time object detection and providing auditory feedback, the system enables users to identify school supplies present in a backpack. The implementation uses libraries such as PyTorch, OpenCV, and pyttsx3, and includes a graphical interface to enhance usability.*

Resumo. *Este trabalho apresenta uma prova de conceito baseada em Redes Neurais Convolucionais (RNCs), com o objetivo de promover acessibilidade no ambiente educacional para pessoas com deficiência visual. O sistema realiza identificação em tempo real de objetos escolares por meio da arquitetura YOLOv5 e fornece feedback sonoro ao usuário. A solução utiliza uma rede pré-treinada para detectar itens relevantes, integrando bibliotecas como PyTorch, OpenCV e pyttsx3 para conversão de texto em fala. Além disso, o modelo conta com uma interface gráfica que facilita a usabilidade, tornando-se uma ferramenta inclusiva no contexto educacional.*

1. Introdução

A inclusão de pessoas com deficiência visual no ambiente escolar requer soluções que promovam autonomia, acessibilidade e, principalmente, equidade no processo de aprendizagem. Nesse contexto, o uso de técnicas de Visão Computacional integradas à Inteligência Artificial tem se mostrado uma abordagem promissora para ampliar a independência desses alunos em atividades cotidianas. Segundo Russell e Norvig (2010), a Inteligência Artificial busca construir sistemas que simulem a inteligência humana para resolver problemas do mundo real, o que justifica sua aplicação nesse tipo de solução.

A educação é um dos pilares fundamentais para o desenvolvimento individual e social, e deve ser acessível a todos, independentemente de suas limitações físicas ou sensoriais. No caso de estudantes com deficiência visual, barreiras como a dependência constante de apoio humano ou a dificuldade em gerenciar materiais escolares de forma autônoma ainda são desafios presentes no cotidiano educacional. Prover tecnologias assistivas eficazes é um caminho para garantir não apenas o acesso ao conhecimento, mas também a permanência e o protagonismo desses estudantes no ambiente escolar.

Este trabalho propõe uma solução baseada em Redes Neurais Convolucionais (RNCs), utilizando a arquitetura YOLOv5 (You Only Look Once) (ULTRALYTICS, 2023), voltada à detecção em tempo real de objetos escolares no interior de mochilas. A proposta inclui o fornecimento de feedback auditivo por meio de síntese de voz, permitindo que o usuário seja informado sobre os itens presentes ou ausentes em sua mochila. O objetivo é apoiar estudantes na organização de seus materiais e oferecer maior autonomia no cotidiano escolar. Este artigo descreve a proposta, os fundamentos teóricos que a sustentam, a arquitetura do sistema implementado e os resultados esperados com a sua aplicação.

2. Referencial Teórico

Esta seção apresenta os fundamentos teóricos que sustentam o desenvolvimento do sistema proposto, com ênfase nas tecnologias utilizadas e nos trabalhos correlatos.

2.1 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (CNNs, do inglês Convolutional Neural Networks) são modelos amplamente utilizados em tarefas de visão computacional, especialmente por sua capacidade de identificar padrões espaciais em imagens com alto desempenho.

Para o desenvolvimento de aplicações baseadas em CNNs, a linguagem Python é amplamente adotada devido à sua simplicidade, flexibilidade e rica base de bibliotecas. Uma das arquiteturas mais reconhecidas na área é a YOLOv5, que equilibra precisão e velocidade na detecção de objetos em tempo real. Essa versão evolui a partir do YOLOv3, proposto por Redmon e Farhadi (2018), sendo adequada para aplicações com múltiplos objetos em uma única imagem.

A integração do YOLOv5 com ferramentas como OpenCV (para captura e pré-processamento de imagens), PyTorch (para modelagem e treinamento da rede) e pyttsx3 (para síntese de fala) permite a construção de sistemas acessíveis e interativos. O conjunto de dados COCO (Common Objects in Context) é utilizado no processo de inferência para detecção de objetos relevantes, como “book”, “backpack” e “scissors” (LIN et al., 2014).

Segundo a documentação oficial da Ultralytics (2024), o YOLOv5 mantém-se competitivo nos benchmarks de detecção por sua leveza, flexibilidade e facilidade de uso, sendo ideal para aplicações em tempo real.

2.2 Trabalhos Correlatos

Guimarães et al. (2025) avaliaram uma CNN enxuta no conjunto CIFAR-10 (60 000 imagens, 10 classes). A rede, composta por duas camadas convolucionais seguidas de pooling, dropout e uma camada totalmente conectada, obteve acurácia de validação em torno de 87 % após seis épocas, com perda estável e baixa diferença treino-validação. A matriz de confusão evidenciou alto desempenho geral, mas apontou confusões sistemáticas entre classes visualmente semelhantes, como “pássaro → avião” (81 casos) e “caminhão → automóvel” (96 casos). Os autores concluem que arquiteturas compactas podem alcançar boa generalização em tarefas de classificação estática, mas requerem estratégias adicionais (por exemplo, data augmentation) para separar classes ambíguas e não contemplam cenários em tempo real (GUIMARÃES et al., 2025).

Moura et al. (2021) propuseram um sistema de detecção de objetos em vídeos de câmeras de vigilância, voltado para identificar situações críticas como porte de armas, incêndios e pichações em tempo real. A solução utilizou a arquitetura YOLOv5, treinada com diferentes datasets especializados, como o Granada (armas), FiSmo (fogo) e FormasGraffitiDataset (pichação). Os resultados mostraram boa precisão para objetos pequenos (≈ 0,88 para armas), embora a presença de falsos positivos e a dificuldade em detectar padrões menos destacados em cenas reais indiquem a necessidade de mais dados realistas e ajustes de modelo. O trabalho destaca-se por integrar múltiplos detectores com potencial de aplicação direta em segurança pública (MOURA et al., 2021).

Já Bisi (2024) concentrou-se no reconhecimento de sinais manuais para inclusão de pessoas surdas. Três modelos de CNN foram treinados em duas bases: o ASL Alphabet (imagens estáticas) e a WLASL (vídeos de palavras). O melhor desempenho adveio de uma abordagem que reduz a entrada a 21 key-points 3-D capturados pelo MediaPipe, atingindo acurácia ≈ 0,998 e 30 fps de predição em hardware modesto. Em contrapartida, o modelo falhou em generalizar para sequências complexas do WLASL, com acurácia < 0,5, revelando desafios quando há elevada variabilidade temporal e ruído visual (BISI, 2024).

Esses estudos demonstram a robustez das CNNs em cenários controlados e a eficácia de representações compactas para mitigar variação de fundo. Contudo, ambos se limitam a classificação: não realizam detecção simultânea de múltiplos objetos nem integram feedback sensorial para usuários com deficiência visual. Moura et al. (2021), por sua vez, exploraram a detecção de múltiplas classes críticas — armas, incêndios e pichações — em vídeos de câmeras de vigilância, obtendo bons resultados para objetos pequenos, mas com desafios de adaptação a cenas reais. O presente trabalho expande o estado da arte ao empregar a arquitetura YOLOv5 para detectar diversos itens escolares em tempo real e fornecer retorno auditivo imediato, promovendo autonomia de estudantes cegos no manejo de materiais. Assim, transpomos as lições de precisão, redução de ruído e operação em tempo real observadas em Guimarães et al., Bisi e Moura et al. para um contexto assistivo multimodal, onde velocidade de inferência e acessibilidade auditiva são requisitos centrais.

3. Arquitetura do Sistema

A solução segue um fluxo Entrada → Processamento → Saída (Figura 1) projetado para operar em tempo real ou sobre imagens já armazenadas:

- **(a) Entrada – Aquisição:** Uma webcam USB capta quadros RGB a 30 fps. Para testes off-line, o sistema também aceita arquivos de imagem. Os quadros são duplicados: um vai para a janela de visualização, outro segue para a fila de inferência, evitando atrasos visíveis.
- **(b) Pré-processamento:** Cada quadro é redimensionado para 640 × 640 px e normalizado, atendendo ao pipeline padrão do YOLOv5. Esses passos são executados em < 5 ms em CPU moderna.
- **(c) Inferência e filtro semântico:** O modelo devolve "bounding-boxes" com classe e confiança. Um filtro retém apenas as nove classes escolares definidas no dicionário interno (backpack, book, scissors, etc.). Opcionalmente, a lista de classes pode ser editada pelo usuário na interface.
- **(d) Pós-processamento e exibição:** As caixas são desenhadas sobre o quadro usando OpenCV. Em paralelo, o sistema calcula o conjunto de itens detectados e faltando em relação a uma lista de materiais esperados, exibindo o resultado no painel da interface.
- **(e) Geração de feedback auditivo:** A biblioteca pyttsx3 transforma os nomes dos objetos em voz. Há suporte a Português ou Inglês (botão "Alternar Idioma" na GUI). Cada item recém-detectado é anunciado; quando algo da lista esperada não é encontrado, o sistema avisa "Falta X".

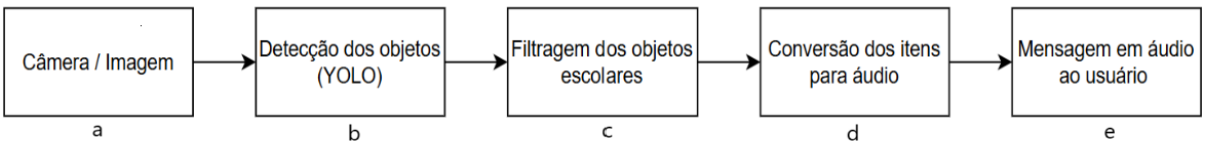

Figura 1 – Arquitetura Funcional

3.1 Implementação

A solução foi desenvolvida em Python 3.12 sobre três bibliotecas-chave — PyTorch, OpenCV-Python e pyttsx3 — mais os recursos nativos do Tkinter para interface gráfica. A Figura 1 resume o fluxo; os principais pontos de implementação são:

- **Carregamento do modelo:** Todos os scripts utilizam o YOLOv5. Os pesos são baixados diretamente do repositório oficial, evitando pré-configuração manual.
- **Pré-processamento de quadros:** Cada frame recebido da webcam é redimensionado para 640 × 640 px e normalizado em < 5 ms (CPU moderna), atendendo ao pipeline do YOLOv5.

- **Detecção e filtro semântico:** A inferência retorna bounding boxes com rótulo e confiança; um dicionário interno mantém nove classes consideradas “materiais escolares”. Essa lista pode ser editada no código ou pela GUI.
- **Síntese de voz offline:** pyttsx3 gera áudio sem depender de internet. A versão PT usa a voz brasileira quando disponível e a versão EN alterna idioma em tempo real via botão “**Alternar Idioma**”.
- **Interface e concorrência:** A GUI Tkinter roda na main thread; a captura de vídeo e a inferência rodam em uma thread separada, garantindo que a janela permaneça responsiva. Botões disponíveis são “**Iniciar Detecção**” que cria a thread e começa o loop, “**Alternar Idioma**” – troca língua/voz e “**Sair**” – encerra a thread e fecha a aplicação.
- **Comparação com lista esperada:** O usuário pode pré-definir um conjunto de materiais (arquivo JSON opcional). O painel da GUI mostra, a cada quadro, **Detectados**, ítens encontrados na mochila e **Faltando**, itens esperados que não apareceram.

4. Resultados

Para avaliar a eficácia do sistema de identificação de objetos com feedback auditivo, foram realizados testes offline utilizando imagens previamente selecionadas da internet. As imagens representavam os nove itens presentes no dataset, a saber: garrafa, mochila, tesoura, relógio, livro, notebook, celular, teclado e caneta. Cada item foi apresentado ao sistema 20 vezes, simulando diferentes condições visuais, como variações de foco, resolução e presença de ruído na cena. Os resultados obtidos são apresentados no Quadro 1.

Item	Acertos	Erros	Não identificado	Total
Garrafa	18	2	0	20
Mochila	10	6	2	20
Tesoura	19	0	1	20
Relógio	14	3	2	20
Livro	16	2	2	20
Notebook	15	3	2	20
Celular	17	2	1	20
Teclado	13	4	3	20
Caneta	12	5	3	20

Quadro 1 – Resultados dos testes com identificação de objetos

A partir dos dados do quadro, é possível observar que:

- A tesoura apresentou o melhor desempenho geral, sendo corretamente identificada em 19 das 20 tentativas, com apenas uma falha de identificação e nenhum erro de classificação. O feedback de áudio foi emitido corretamente em todas as ocasiões em que o item foi reconhecido com sucesso, demonstrando a eficiência do sistema nesse caso.

- A garrafa também obteve desempenho elevado, com 18 acertos e apenas 2 erros, sem falhas de não identificação. Isso evidencia a estabilidade e precisão do modelo para esse tipo de objeto, mesmo com pequenas variações de posicionamento e iluminação.
- Por outro lado, a mochila apresentou a menor taxa de acertos: foi corretamente reconhecida em apenas 10 das 20 tentativas, com 6 erros de classificação e 4 casos de não identificação. Esse desempenho inferior indica que o sistema encontra maior dificuldade em detectar esse item, especialmente quando aparece desfocado, com baixa resolução ou inserido em cenários visualmente poluídos.
- Itens como livro e notebook apresentaram desempenho intermediário (acertos de 16 e 15 respectivamente), sendo mais afetados por variações de iluminação e sobreposição parcial com outros objetos. O teclado também se mostrou mais suscetível a erros de classificação, com apenas 13 acertos e 4 erros, principalmente quando posicionado em ângulos não convencionais.
- O celular teve desempenho consistente, com 17 acertos, sendo impactado negativamente apenas em condições de forte reflexo na tela. A caneta, por ser um objeto pequeno e fino, apresentou maior dificuldade de detecção, com 12 acertos, 5 erros e 3 casos de não identificação, especialmente quando sobreposta a fundos visuais complexos.

Além da taxa de acerto, também foi verificado se o feedback de áudio era acionado corretamente após cada identificação. Em todos os casos em que o objeto foi reconhecido com sucesso, o sistema emitiu o áudio correspondente, confirmando o funcionamento adequado do módulo de síntese de voz. A Figura 1 apresenta um exemplo de imagem com baixo ruído visual, na qual a detecção foi bem-sucedida. Em contraste, a Figura 2 mostra um exemplo em que o objeto (mochila) aparece desfocado e cercado por outros elementos, resultando em ruído visual elevado e falha na detecção.

Figura 1 – Exemplo de imagem sem ruído visual que obteve êxito na detecção

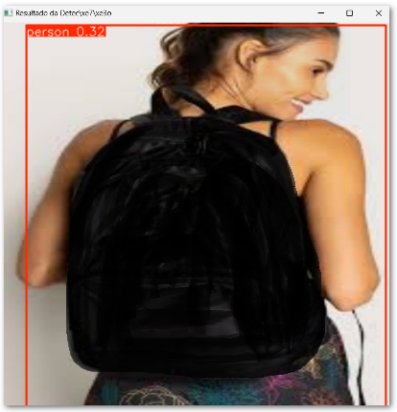

Figura 2 - Exemplo de imagem com ruído visual no qual o objeto não é detectado

Além dos testes offline com imagens previamente selecionadas, foram conduzidos testes em tempo real utilizando a webcam do sistema, com o objetivo de simular condições reais de uso por estudantes com deficiência visual. Nessa fase, os nove objetos físicos foram posicionados frente à câmera em diferentes ângulos, distâncias e sob diferentes condições de iluminação. O modelo YOLOv5 foi executado localmente, realizando a detecção e classificação dos objetos de forma contínua. Quando um objeto era identificado com sucesso, o sistema acionava imediatamente o módulo de síntese de voz, emitindo um feedback auditivo com a identificação do item detectado.

Os dados coletados durante esses testes práticos foram os seguintes:

- **Garrafa:** Utilizadas 3 garrafas diferentes, mostradas 20 vezes. O sistema obteve 18 acertos, 2 erros e nenhum não identificado
- **Mochila:** Testadas 4 mochilas distintas, totalizando 20 apresentações. Foram 11 acertos, 5 erros e 4 não identificados, reforçando a dificuldade observada nos testes offline.
- **Tesoura:** Utilizadas 3 tesouras diferentes, apresentadas 20 vezes. O modelo acertou 20 e não apresentou erros nem casos de não identificação.
- **Relógio:** Testados 2 relógios, totalizando 20 apresentações. Foram 15 acertos, 4 erros e 1 não identificado, com falhas mais comuns em baixa iluminação.
- **Livro:** Utilizados 6 livros diferentes, 20 apresentações. Foram 17 acertos, 3 erros e nenhum não identificado.
- **Notebook:** Testados 2 notebooks, 20 apresentações. O sistema acertou 16, com 2 erros e 2 não identificados, principalmente quando o equipamento estava parcialmente fechado.
- **Celular:** Utilizados 3 modelos diferentes, 20 apresentações. Obtidos 17 acertos, 1 erro e 2 não identificado, com desempenho estável mesmo sob reflexos moderados.
- **Teclado:** Testados 2 teclados diferentes, 20 apresentações. O modelo acertou 14, com 4 erros e 2 não identificados, apresentando maior dificuldade com teclados compactos.
- **Caneta:** Utilizadas 4 canetas distintas, 20 apresentações. Foram 13 acertos, 6 erros e 1 não identificado, sendo os erros mais comuns em fundos visualmente complexos.

Os resultados dos testes com webcam reforçam as observações feitas nos testes offline, indicando que itens com formatos mais padronizados e características visuais consistentes — como tesouras e garrafas — tendem a ser reconhecidos com maior precisão pelo modelo. Isso é ilustrado na Figura 3, em que, apesar da presença de ruído visual, o objeto é corretamente detectado devido ao seu formato mais regular.

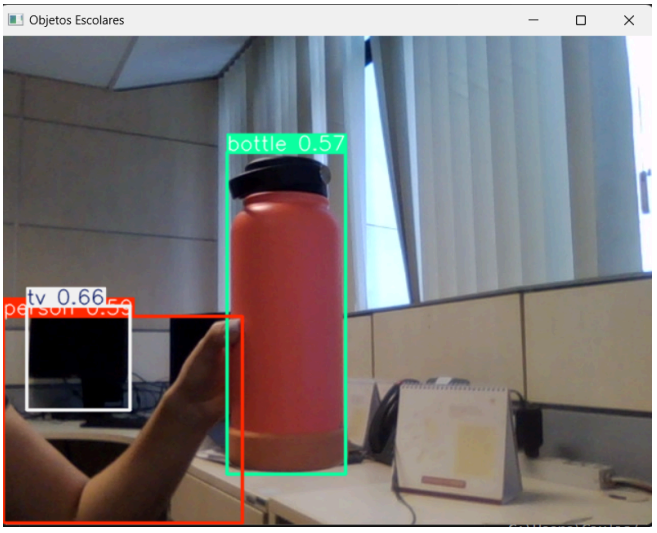

Figura 3 - Detecção bem-sucedida de uma garrafa em tempo real, com ruído visual

Em contrapartida, objetos com ampla variação de forma, cor, textura e tamanho, como mochilas, ainda representam um desafio para o sistema. A Figura 4 exemplifica esse cenário ao mostrar uma mochila que foi incorretamente classificada como "suitcase" em vez de "backpack", evidenciando a dificuldade do modelo em lidar com a diversidade visual desse tipo de item, agravada pela presença de ruído no ambiente.

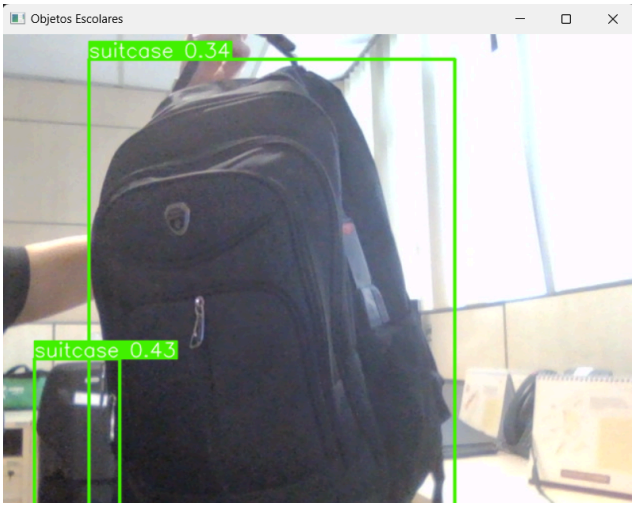

Figura 4 - Falha na detecção de uma mochila em ambiente com ruído visual

Esses resultados sugerem que a acurácia do modelo pode ser significativamente aprimorada com o aumento da variedade de exemplos no conjunto de dados de treinamento, especialmente para classes mais heterogêneas.

5. Conclusão

Este projeto oferece uma solução que combina Visão Computacional e Inteligência Artificial para promover a inclusão educacional de estudantes com deficiência visual. Ao fornecer feedback auditivo sobre o conteúdo da mochila escolar, o sistema contribui para a autonomia desses estudantes na organização de seus materiais.

Entre as melhorias futuras previstas, destacam-se:

- Treinamento com *datasets* personalizados, incluindo objetos escolares não presentes no COCO.

- Expansão da lista de materiais esperados, com possibilidade de personalização conforme o ano letivo.
- Aprimoramento da interface gráfica, com foco em acessibilidade (integração com leitores de tela, contraste ampliado, fontes maiores).

Tais melhorias ampliam a aplicabilidade do sistema, tornando-o mais preciso, personalizável e eficaz para o público-alvo.

Referências

LIN, T. Y. et al. Microsoft COCO: Common Objects in Context. arXiv preprint arXiv:1405.0312, 2014.

PYTTSX3. pyttsx3 Documentation. Disponível em: https://pyttsx3.readthedocs.io. Acesso em: 15 maio 2025.

REDMON, Joseph; FARHADI, Ali. YOLOv3: An Incremental Improvement. arXiv, 2018. Disponível em: https://arxiv.org/abs/1804.02767.

RUSSELL, Stuart; NORVIG, Peter. Inteligência Artificial. 3. ed. São Paulo: Pearson, 2010.

ULTRALYTICS. YOLOv5 Documentation. GitHub, 2023. Disponível em: https://github.com/ultralytics/yolov5.

TKINTER DOCS. Tkinter GUI Application Development. Disponível em: https://docs.python.org/3/library/tkinter.html. Acesso em: 15 maio 2025.

ZHANG, Wei et al. Object Detection for the Visually Impaired: A Review. Sensors, 2021.

SILVA, Maria C. et al. Sistema de Visão Computacional para Acessibilidade. Revista Brasileira de Informática Aplicada, 2022.

GUIMARÃES, A. J. S.; BARBOSA, D. S.; SOUSA, G. da S.; SANTOS, A. J. de S. Ação de redes neurais convolucionais na classificação de imagens: abordagens e resultados. Revista Contemporânea, v. 5, n. 1, p. e7190, 2025. DOI: 10.56083/RCV5N1-012. Disponível em: https://ojs.revistacontemporanea.com/ojs/index.php/home/article/view/7190. Acesso em: 21 mai. 2025.

BISI, J. V. dos S. Reconhecimento de padrões em linguagem de sinais utilizando redes neurais convolucionais. 2024. 50 f. Trabalho de Conclusão de Curso (Engenharia de Controle e Automação) – Instituto Federal do Espírito Santo, Linhares, 2024. Disponível em: https://repositorio.ifes.edu.br/handle/123456789/5491. Acesso em: 21 mai. 2025.

MOURA, Natan; CLARO, Daniela Barreiro; GONDIM, João Medrado. Análise experimental para a detecção de objetos em vídeos de câmeras de vigilância: uma abordagem para porte de arma, incêndio e pichação. In: I CONCURSO DE TRABALHOS DE INICIAÇÃO CIENTÍFICA (CTIC 2021), Minas Gerais. *Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, 2021. Disponível em: https://sol.sbc.org.br/index.php/webmedia_estendido/article/view/17608. Acesso em: 8 ago. 2025.