SoundEyes: Audiodescrição de Obstáculos para Pessoas com Deficiência Visual

Jerson V.P. Gomes, Wallace F. Oliveira, Fellipe G. Oliveira
Rafael H.N. Diniz, Matheus A. Souza, Felipe D. Cunha

Depto. de Ciência da Computação – Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – MG – Brasil

{jerson.gomes, wallace.oliveira.1413725, 1313536}@sga.pucminas.br,
rafahdiniz@yahoo.com.br, {felipe, matheusalcantara}@pucminas.br

Abstract. *Visually impaired individuals often face challenges when navigating unfamiliar or dynamic environments, where access to real-time spatial information is limited. This paper presents the development of SoundEyes, an assistive technology for visually impaired individuals that employs computer vision techniques to recognize objects and generate real-time audio descriptions. Designed for mobile devices with limited resources, the system uses an edge computing architecture and Bluetooth communication to ensure low latency and high autonomy. In practical tests with mobile devices, SoundEyes achieved a total response time of less than one second in HVGA mode on mid-range devices and demonstrated greater detection accuracy in XGA mode, showing promise for both dynamic and static environments.*

Resumo. *Pessoas com deficiência visual frequentemente enfrentam desafios ao se locomoverem em ambientes desconhecidos ou dinâmicos, onde o acesso a informações espaciais em tempo real é limitado. Este artigo apresenta o desenvolvimento do SoundEyes, uma tecnologia assistiva voltada para pessoas com deficiência visual. O sistema utiliza técnicas de visão computacional para reconhecer objetos e gerar descrições auditivas em tempo real, com foco em dispositivos móveis de baixo custo. Por meio de uma arquitetura baseada em Edge Computing e comunicação via Bluetooth, o SoundEyes permite maior autonomia e segurança durante a navegação. Em testes práticos com dispositivos mobile, o sistema obteve um tempo total de resposta inferior a 1 segundo no modo HVGA em dispositivos intermediários, e apresentou maior precisão no modo XGA, com potencial para ambientes dinâmicos e estáticos.*

1. Introdução

A visão é um sentido fundamental para a interação humana com o ambiente, permitindo a captação e interpretação imediata de informações visuais, o que amplia significativamente a compreensão espacial ao redor. Além disso, ela é essencial na comunicação não verbal, na identificação de objetos e símbolos, e na realização de atividades rotineiras, como o deslocamento seguro por diferentes ambientes. No entanto, segundo o *Relatório Mundial sobre a Visão da Organização Mundial da Saúde* [World Health Organization 2019], estima-se que mais de 2,2 bilhões de pessoas no mundo apresentem algum grau de deficiência visual, variando de baixa visão à cegueira completa.

Essa limitação sensorial impõe sérios desafios à autonomia, segurança e qualidade de vida dessas pessoas, dificultando interações sociais, profissionais e pessoais. Nesse cenário, tecnologias baseadas em computação ubíqua e pervasiva têm se mostrado fundamentais para promover acessibilidade em tempo real, integrando sensores, redes e dispositivos móveis ao cotidiano. Soluções como bengalas inteligentes, sistemas de orientação *indoor* e reconhecimento de objetos vêm explorando dispositivos conectados, com suporte à Internet das Coisas (IoT) e computação de borda (Edge Computing), para oferecer maior independência e segurança às pessoas com deficiência visual.

Neste contexto, este trabalho propõe o ***SoundEyes***, uma tecnologia assistiva baseada em computação móvel, visão computacional e comunicação Bluetooth, desenvolvida para fornecer descrições auditivas em tempo real a partir da detecção de objetos em ambientes reais. O sistema utiliza o *smartphone* como núcleo da solução, atuando como centro de processamento e resposta. Por estar constantemente presente com o usuário e integrar-se ao seu cotidiano, o celular configura-se como uma plataforma pervasiva, capaz de reagir dinamicamente ao ambiente e fornecer suporte assistivo de forma contínua e contextual. Dessa forma, as principais contribuições deste trabalho são:

- A concepção de uma arquitetura baseada em dispositivos móveis e comunicação via Bluetooth, que opera de forma totalmente local, sem necessidade de internet.
- A implementação de um *pipeline* de captura, detecção e audiodescrição com uso de modelos otimizados (YOLOv8) para funcionamento em tempo real.
- A avaliação empírica do sistema em diferentes dispositivos Android, comparando os modos velocidade (HVGA) e qualidade (XGA) quanto ao tempo de resposta e precisão da classificação.
- A proposição de um método simples de localização espacial de objetos na imagem, enriquecendo a audiodescrição com informações contextuais de direção.

O restante do artigo está estruturado da seguinte forma: a Seção 2 apresenta trabalhos relacionados. Na Seção 3 detalha-se o planejamento do projeto, etapas, métodos de avaliação, a arquitetura projetada e desafios de implementação. A Seção 4 traz a avaliação da solução e, por fim, na Seção 5 são feitas considerações finais e sugestões de trabalhos futuros.

2. Trabalhos Relacionados

Os recentes avanços nos campos da computação, inteligência artificial e Internet das Coisas (IoT) impulsionaram o desenvolvimento de tecnologias assistivas para pessoas com limitações sensoriais. Cada vez mais, são explorados o uso de sensores vestíveis, dispositivos móveis e Inteligência Artificial para implementar mecanismos que promovam maior acessibilidade e auxílio por meio da tecnologia.

A navegação assistida para indivíduos com deficiência visual tem sido uma área de intenso estudo, com esforços voltados para o aprimoramento da qualidade de vida. O sistema descrito em [Supekar and Patil 2022], por exemplo, emprega processamento de imagens em tempo real para converter informações visuais em texto, áudio ou sinais táteis, utilizando um microcomputador Raspberry Pi, câmeras e motores de vibração. Outra alternativa, apelidada de "*Guide-Me*"[Dissanayake et al. 2021], propõe o uso de *beacons* Bluetooth e reconhecimento de voz para criar um sistema de navegação em ambientes fechados, como prédios públicos, salões de eventos e aeroportos, garantindo maior segurança e precisão na locomoção.

Além disso, dispositivos conectados à Internet das Coisas (IoT), como a bengala inteligente SCBIoT [AbdElminaam et al. 2022], combinam sensores ultrassônicos e GPS(Global Positioning System) para detecção de obstáculos e monitoramento remoto de localização. Complementando essas soluções, [Osama et al. 2021] propõe um assistente móvel para identificação de objetos cotidianos, como cédulas e roupas, empregando redes neurais, como as MobileNets [Howard et al. 2017], para classificação em dispositivos de baixo custo.

A identificação eficiente de objetos em imagens desempenha um papel fundamental para garantir informações confiáveis e em tempo real. Nesse sentido, mecanismos de visão computacional como o YOLO (You Only Look Once) [Redmon et al. 2016] ganham destaque. Esse modelo utiliza um sistema de regressão unificado para determinar as coordenadas e probabilidades de objetos, reduzindo o número de interações necessárias para classificá-los.

Essa abordagem viabiliza alta eficiência computacional e processamento em tempo real, características essenciais para aplicações em dispositivos móveis ou vestíveis. Com a versão de [Jocher et al. 2023], houve um aprimoramento significativo na precisão e latência, tornando-a ideal para integração em sistemas assistivos. A necessidade de algoritmos rápidos e precisos também é ressaltada por [Devi and Subalalitha 2021], que revisa projetos de bengalas inteligentes equipadas com IA, destacando a importância da detecção contextual de obstáculos, como escadas e buracos.

3. Metodologia

No âmbito de desenvolvimento da solução proposta, foi aplicada uma abordagem que emprega diversos conceitos de computação móvel, redes e arquiteturas de computadores e processamento de imagens, para implementar um sistema de auxílio à mobilidade e audiodescrição para pessoas com deficiência visual.

Nesse sentido, escolhas como o processamento em bordas de rede (ou *Edge Computing*), o uso de redes bluetooth e computação mobile, se justificam pelo fato de ser uma aplicação voltada ao fácil acesso e à baixa latência em termos de tempo de processamento e resposta, e podem ser melhor compreendidas nas subseções que se seguem, por meio das seguintes etapas: 3.1 desenvolvimento da arquitetura do sistema, incluindo a escolha de sensores, algoritmos de visão computacional e estratégias de processamento de dados; 3.2 método de avaliação da solução; e 3.3 desafios enfrentados.

3.1. Arquitetura Proposta

A solução SoundEyes está dividida em dois ambientes funcionais: o Ambiente de Captura (ESP32-CAM + OV2640) e o Ambiente de Processamento e Resposta (*Smartphone* Android). A escolha de utilizar um ambiente móvel baseia-se em sua acessibilidade, portabilidade e conveniência. No entanto, essa escolha requer cuidados, pois os dispositivos móveis possuem limitações de memória e processamento. Já o ambiente de captura enfrenta desafios relacionados à qualidade das imagens capturadas e à latência na transmissão dos dados pela rede bluetooth, exigindo uma conexão estável entre os ambientes para garantir uma comunicação eficaz. A Figura 1 exibe o fluxograma da arquitetura do sistema, e a descrição dos blocos se encontra nos tópicos a seguir.

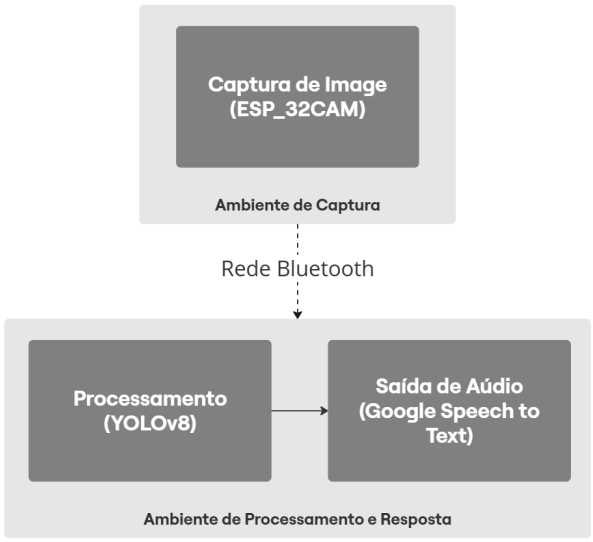

Figura 1. Fluxograma - Representação da Arquitetura do Sistema

- **Bloco de Captura**: O ambiente de captura de imagem é composto por um módulo ESP32-CAM e uma câmera OV2640, escolhidos por sua capacidade de capturar imagens de até 2 megapixels, ter um módulo Bluetooth embutido e a facilidade de programação, permitindo a captura e transmissão de imagens em tempo real. Na melhor configuração para o projeto, a câmera foi ajustada para capturar uma sequência contínua de frames em duas resoluções, priorizando velocidade e qualidade. Para maior velocidade, a imagem é limitada a 0,5 megapixels em resolução HVGA(480×320); para melhor qualidade, adotou-se a resolução XGA(1024×768) com 1 megapixel.Para facilitar o uso por pessoas com deficiência visual, o módulo ESP32-CAM é acoplado a um suporte no abdómen, alinhando o campo de visão da câmera com o eixo do corpo do usuário. Dessa forma, basta segurar o dispositivo com uma mão para obter imagens estáveis e sempre voltadas para a direção de deslocamento.

Figura 2. Módulo ESP32CAM com Câmera OV2640.

- **Processamento de Imagens**: Para a detecção de objetos, o sistema utiliza o YOLOv8n (nano) [Jocher et al. 2023], em que é pré-treinado no conjunto COCO (80 classes). A inferência ocorre diretamente no TensorFlow para Android, exigindo apenas ajustes de pré-processamento de imagem. Conhecido por sua alta precisão e capacidade de identificar múltiplos objetos em tempo real. Além disso, sua abordagem pré-treinada elimina a necessidade de ajustes complexos

ou treinamento personalizado, e possui ampla generalização para objetos do cotidiano.

O processamento da imagem recebida pelo bloco de captura inicia-se com um pré-processamento, no qual as imagens são convertidas para um formato compatível com o YOLO. Em seguida, o modelo identifica as coordenadas e probabilidades dos objetos detectados, informações que serão utilizadas pelo bloco de audiodescrição para gerar a saída em áudio.

- **Audiodescrição**: Após a detecção dos objetos, o sistema gera uma descrição auditiva dos itens identificados e suas posições. Para isso, é utilizada a tecnologia de síntese de fala Google Text-to-Speech [Google LLC 2024], que converte o texto em áudio. As descrições são criadas levando em conta a relevância dos objetos no contexto do usuário, de forma que informações mais importantes sejam priorizadas.
- **Integração dos Componentes**:Para a integração entre os diferentes componentes, foi utilizada a tecnologia Bluetooth Low Energy (BLE), que permite a transmissão de dados de forma eficiente sem a necessidade de uma conexão com a internet. Seu baixo consumo de energia e à capacidade de estabelecer uma rede de área pessoal (PAN) se mostra ideal para o projeto, pois possibilita a troca de informações de forma contínua e confiável entre os dispositivos.

O sistema foi configurado para funcionar em três etapas, que rodam em ciclos. Na primeira etapa, o Bloco de Captura obtém uma imagem, e realiza a compressão JPEG da mesma, afim de facilitar o envio. A imagem comprimida é então dividida em pacotes de 20.480 bytes e enviadas a uma taxa de 1Mbps via bluetooth para a aplicação.

Na segunda etapa, a imagem recebida é direcionada ao processamento YOLO, para definir a localização dos objetos dentro da cena, entre três possíveis posições: Esquerda, Frente e Direita. Para isso, a imagem de entrada é dividida em uma grade (grid) 3x1, e o posicionamento de cada objeto localizado é determinado pela comparação de sua área com as coordenadas da imagem em que foi detectado. Ou seja, a coluna na qual a maior parte da área do objeto está localizada define sua classificação de posicionamento. A Figura 3, exibe a detecção de múltiplos objetos dentro de uma cena, enquanto as Figuras 4 e 5, exibem respectivamente a detecção e a marcação de posicionamento para um veículo e pedestres caminhando.

Na terceira e última etapa, com a definição do objeto e seu posicionamento, o sistema gera uma descrição auditiva dos itens identificados com base na relevância e proximidade do usuário.

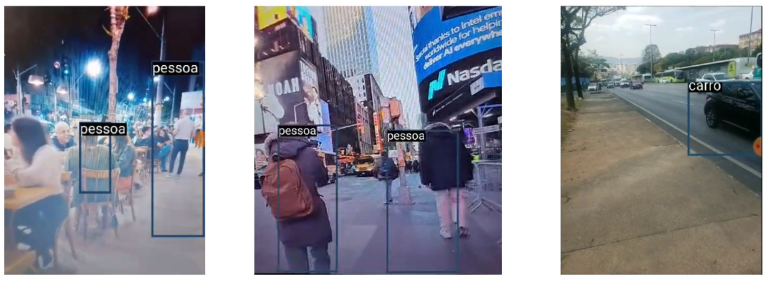

Figura 3. Saída da detecção de objetos utilizando YOLOv8.

Figura 4. Carro detectado à direita

Figura 5. Detecção de pessoas (esquerda e direita)

3.2. Método de Avaliação

A avaliação da solução proposta se dá por meio de métricas de desempenho, combinadas com testes e simulações. Como o projeto é baseado em modelos previamente treinados,as métricas de avaliação focam principalmente em métricas de desempenho do sistema como tempo de resposta e confiabilidade/qualidade da resposta.

- **Tamanho da Imagem:** A resolução da imagem impacta diretamente no tempo de envio e classificação dos frames. Foram consideradas duas resoluções:
  - **XGA (1024 x 768):** Exige maior volume de dados, aumentando a quantidade de pacotes e o tempo de envio.
  - **HVGA (480 x 320):** Exige menor volume de dados, resultando em menor latência.
- **Tempo de resposta:**
  - **Quantidade dos Pacotes:** Avalia o número de pacotes na transmissão via Bluetooth. A solução deve considerar que os frames são divididos em pacotes de 20.480 bytes e transmitidos a uma frequência de 1 Mbps.

É importante minimizar a quantidade de pacotes e as quebras durante a transmissão para reduzir a latência.
  - **Tempo de Envio dos Pacotes:** Mede o tempo total necessário para enviar todos os pacotes de uma imagem. A solução deve buscar reduzir esse tempo, considerando a taxa de transmissão (1 Mbps) e possíveis interferências no canal Bluetooth, como perdas ou retransmissões.
- **Tempo de classificação:** Avalia o tempo que o modelo leva para processar as informações e gerar uma resposta no hardware projetado. Espera-se que a solução apresente tempos curtos de resposta para cada ciclo de processamento. O tempo de resposta será medido a partir do recebimento de imagem até a emissão da resposta, usando ferramentas de monitoramento de tempo de execução.

O desempenho da solução será avaliado com base nos valores obtidos em cada uma dessas métricas, além da facilidade e acessibilidade oferecidas pela solução proposta.

3.3. Implementação

Durante a implementação do sistema, foram enfrentados desafios relacionados à captura, processamento e integração dos componentes. Um dos principais problemas observados foi o condicionamento da iluminação: o sistema apresentou redução na precisão em ambientes com pouca luz ou iluminação variável, o que exigiu ajustes nos parâmetros de detecção para preservar a acurácia do modelo. Além disso, a identificação de objetos pequenos ou parcialmente ocultos mostrou-se limitada, já que o modelo YOLO utilizado prioriza a generalização para objetos maiores, resultando em dificuldades na detecção de itens com menor área visível. Outro desafio relevante foi garantir a integração eficiente do áudio em tempo real, de forma que a descrição gerada estivesse sincronizada com a detecção visual; isso demandou otimizações no processamento e adoção de estratégias para redução da latência do sistema. O código-fonte do sistema e os artefatos desenvolvidos estão disponíveis para consulta no repositório oficial do projeto[1].

4. Testes e Resultados

Para o teste do sistema proposto, três diferentes dispositivos android, classificados com dispositivos de entrada ou intermediários, foram utilizados. As imagens capturadas pelo ESP32-CAM foram enviadas aos *smartphones* Samsung A35 5G, A14 e A10, onde o aplicativo Android executa o *pipeline* de inferência e síntese de fala. As especificações de cada dispositivo (Tabelas 1–3) impactam diretamente no tempo de classificação e recebimento das imagens disponibilizadas pela ESP32-CAM (Tabela 4). Suas especificações arquiteturais podem ser consultadas nas Tabelas 1, 2 e 3.

Tabela 1. Especificações do dispositivo Samsung A35 5G

Dispositivo Android	**Samsung A35 5G**
Processador	4x 2.4 GHz Cortex-A78 + 4x 2.0 GHz Cortex-A55
GPU	Mali-G68 MP5
RAM	6 GB
Bluetooth	5.3 com A2DP/LE/aptX

[1] `https://github.com/ICEI-PUC-Minas-CC-TI/plmg-cc-ti5-2024-2-g03-soundeyes`

Tabela 2. Especificações do dispositivo Samsung A10

Dispositivo Android	**Samsung A10**
Processador	2x 1.6 GHz Cortex-A73 + 6x 1.35 GHz Cortex-A53
GPU	Mali-G71 MP2
RAM	2 GB
Bluetooth	5.0 com A2DP/LE

Tabela 3. Especificações do dispositivo Samsung A14

Dispositivo Android	**Samsung A14**
Processador	2x 2.4 GHz Cortex-A78 + 6x 2.0 GHz Cortex-A55
GPU	Mali-G68 MP2
RAM	4 GB
Bluetooth	5.2 com A2DP/LE

Tabela 4. Especificações do ESP32-Cam

ESP32-Cam	**Características**
Câmera	OV2640 2MP
Velocidade do Clock	240 MHz
Conectividade	Bluetooth BLE 4.2
SRAM	520 Kbytes
Memória Flash	4 MB

A Tabela 5, compara os tamanhos da imagem geradas e o número total de pacotes necessários para envio via canal Bluetooth, para os dois diferentes modos de captura HVGA (velocidade — 430x320 pixels) e XGA (qualidade — 1024x768 pixels), conforme esperado, o modo HVGA requer menos bytes e consequentemente menor número de pacotes devido a redução no número de informações visuais capturadas.

Tabela 5. Tamanho Imagem HVGA x XGA

Descrição	**HVGA**	**XGA**
Tamanho da Imagem	0.3 MB	1.7 MB
Número de Pacotes	16	88

As Tabelas 6 e 7 realizam o comparativo de tempos obtidos para os modos velocidade e qualidade respectivamente, para cada dispositivo utilizado nos teste.

Tabela 6. Tempo de classificação/resposta modo HVGA(480×320) (Velocidade)

Descrição	**Samsung A35**	**Samsung A10**	**Samsung A14**
Classificação da Imagem	194.77 ms	337,72 ms	522.23 ms
Recebimento de Imagem	789.04 ms	552.41 ms	282.35 ms
Tempo Total	983,81 ms	877.75 ms	804.58 ms

Tabela 7. Tempo de classificação/resposta modo XGA(1024x768) (Qualidade)

Descrição	Samsung A35	Samsung A10	Samsung A14
Classificação da Imagem	190.79 ms	323.53 ms	490.00 ms
Recebimento de Imagem	1292.58 ms	901.43 ms	563.66 ms
Tempo Total	1483.37 ms	1224.95 ms	1053.66 ms

Com base nos tempos e na qualidade de classificação observados, foi possível notar que, embora o modo velocidade (HVGA) apresente tempos menores tanto para o envio quanto para o tempo total, suas respostas foram mais inconsistentes. Isso se justifica pela redução na resolução e pela maior sensibilidade a fatores externos, como iluminação e ângulo de captura, conforme discutido na Seção 3.3, que impactam diretamente o modelo de classificação.

A análise dos tempos de resposta revelou um importante compromisso entre desempenho e qualidade. O modo HVGA atingiu o menor tempo total de resposta no Samsung A14 (804 ms), representando uma redução de aproximadamente 24% em relação ao tempo obtido no modo XGA (1053 ms). Essa diferença chega a 46% quando comparados os extremos: 804 ms (A14/HVGA) contra 1483 ms (A35/XGA). No entanto, a agilidade compromete a acurácia, pois a menor resolução reduz a capacidade de detecção do modelo YOLO para objetos pequenos ou parcialmente ocultos.

Por outro lado, no modo XGA, a classificação foi mais precisa devido à maior riqueza de informações visuais, com tempos de processamento ainda aceitáveis — variando de 1053 ms a 1483 ms entre os dispositivos. A classificação da imagem no A35, por exemplo, foi 2,7% mais rápida no modo XGA (190,79 ms) em relação ao HVGA (194,77 ms), demonstrando que o tempo de envio — e não de processamento — é o principal fator de impacto em resoluções maiores.

Esses resultados reforçam a viabilidade de alternar entre os modos conforme o contexto de uso: priorizando velocidade em ambientes dinâmicos, como vias públicas, e precisão em ambientes estáticos, como corredores internos ou áreas de espera.

5. Conclusão

As pessoas com deficiência visual enfrentam diversas dificuldades, como mencionado anteriormente. Nesse contexto, os avanços tecnológicos são ferramentas valiosas para promover maior acessibilidade e inclusão. É essencial, portanto, que a comunidade científica e da computação continue explorando essas tecnologias como aliadas na promoção da autonomia e da dignidade para pessoas com deficiência.

Diante desse cenário, este trabalho apresentou o SoundEyes, um sistema assistivo baseado em computação móvel, visão computacional e comunicação local, projetado para fornecer descrições auditivas a partir da detecção de objetos em ambientes reais.

Os resultados experimentais demonstraram que a solução é viável para dispositivos móveis de baixo custo, alcançando tempos de resposta inferiores a 1 segundo no modo HVGA e maior precisão de reconhecimento no modo XGA. Tais resultados confirmam que o sistema atende aos objetivos propostos, oferecendo uma ferramenta acessível, eficiente e com potencial de aplicação em diferentes contextos de uso, tanto

dinâmicos quanto estáticos. Além disso, o uso de Edge Computing e modelos otimizados como o YOLOv8 reforça a capacidade do sistema de operar em ambientes com recursos limitados, sendo aplicável no contexto de computação ubíqua e pervasiva.

Como contribuições futuras, destaca-se a possibilidade de integrar sensores contextuais, como GPS (Global Positioning System) e acelerômetros, e explorar abordagens adaptativas para a priorização de objetos em diferentes situações de mobilidade. O uso de dispositivos móveis, inteligência artificial e redes pode ser expandido para outras aplicações assistivas. Tecnologias para identificação e tradução de sinais em Libras para múltiplos idiomas, por exemplo, ou sistemas de detecção e notificação de quedas para pessoas com mobilidade reduzida, podem ser fundamentais para garantir segurança e suporte imediato.

O avanço de tecnologias assistivas como o SoundEyes depende da continuidade de pesquisas voltadas à melhoria dos modelos de classificação, à redução de latência nas comunicações e ao aprimoramento da usabilidade em cenários reais.

Referências

[AbdElminaam et al. 2022] AbdElminaam, D. S., Ahmed, I. A.-E., and Sakr, F. (2022). SCBIoT: Smart cane for blinds using IoT. In *International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 371–377.

[Devi and Subalalitha 2021] Devi, S. K. and Subalalitha, C. N. (2021). Deep learning based audio assistive system for visually impaired people. *Computers, Materials and Continua*, 71(1):1205–1219.

[Dissanayake et al. 2021] Dissanayake, D. M. L. V., Rajapaksha, R. G. M. D. R. P., Prabhashawara, U. P., Solanga, S. A. D. S., and Anuradha Jayakody, J. A. D. C. (2021). Guide-me: Voice authenticated indoor user guidance system. In *IEEE Ubiquitous Computing, Electronics & Mobile Comm. Conf. (UEMCON)*, pages 0509–0514.

[Google LLC 2024] Google LLC (2024). Google Text-to-Speech (gTTS) API.

[Howard et al. 2017] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

[Jocher et al. 2023] Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics yolov8.

[Osama et al. 2021] Osama, M., Yehia, A., Mohamed, S., Sherief, R., Elmasry, N., Adel, V., and Hamdy, A. (2021). Design and implementation of visually impaired assistant system. In *Int. Mobile, Intelligent, and Ubiquitous Comp. Conf.*, pages 303–310.

[Redmon et al. 2016] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection.

[Supekar and Patil 2022] Supekar, A. and Patil, S. (2022). Design and development of portable navigation system for disabled person using image, text and audio. In *IEEE Delhi Section Conference (DELCON)*, pages 1–4.

[World Health Organization 2019] World Health Organization (2019). *World Report on Vision*. World Health Organization, Geneva, Switzerland. Acesso em: 19 jan. 2025.