SoundEyes: Audiodescrição de Obstáculos para Pessoas com Deficiência Visual

Fellipe Gabriel de Oliveira, Jerson Vitor de Paula Gomes, Wallace Freitas Oliveira

Departamento de Ciência da Computação (DCC)
Instituto de Ciências Exatas e Informática (ICEI)
Pontifícia Universidade Católica de Minas Gerais (PUC-MG)

{1313536,1416363,1413725}@sga.pucminas.br

Abstract. *This paper presents the development of SoundEyes, an assistive technology for visually impaired individuals, which employs object recognition algorithms and audio descriptions to map and describe environments. The goal is to provide users with greater autonomy and safety, facilitating navigation in spaces through mobile devices. The implementation includes the use of techniques such as YOLO for real-time object detection for optimization on resource-limited devices.*

Abstract. *Este artigo apresenta o desenvolvimento do SoundEyes, uma tecnologia assistiva para pessoas com deficiência visual, que utiliza algoritmos de reconhecimento de objetos e audiodescrição para mapear e descrever ambientes. O objetivo é proporcionar maior autonomia e segurança para os usuários, facilitando a navegação em espaços por meio de dispositivos móveis. A implementação inclui o uso de técnicas como YOLO para detecção de objetos em tempo real para otimização em dispositivos com recursos limitados.*

1. Introdução

A visão é um dos sentidos mais complexos e fundamentais para a interação humana com o ambiente. Ela nos permite captar e interpretar informações visuais de maneira imediata, facilitando a compreensão do espaço, a leitura de sinais e a comunicação não verbal. Esse sentido é essencial para a execução de diversas atividades cotidianas, desde tarefas simples, como a identificação de objetos, até atividades mais complexas, como a navegação em espaços desconhecidos ou o uso de ferramentas tecnológicas.

Estima-se que 2,2 bilhões de pessoas em todo o mundo apresentam algum grau de deficiência visual, variando de baixa visão à cegueira completa, segundo o *Relatório Mundial sobre a Visão* [World Health Organization 2019]. Essa limitação sensorial impõe desafios significativos à autonomia e à segurança dessas pessoas, afetando suas interações sociais, profissionais e pessoais. Sem o suporte e a acessibilidade adequados, elas tornam-se mais suscetíveis a acidentes e erros na execução de suas atividades diárias.

Com o objetivo de reduzir essas limitações e promover maior autonomia para pessoas com deficiência visual, diversos estudos têm buscado soluções inovadoras em diferentes áreas tecnológicas. Um exemplo é o estudo *Design and Implementation of Visually Impaired Assistant System* [Osama et al. 2021], que propõe uma solução para a identificação de objetos, como dinheiro e roupas, através da câmera de um smartphone. Outro exemplo é *A Design Review of Smart Stick for the Blind Equipped with Obstacle*

Detection and Identification using Artificial Intelligence [Niranjan Balu 2019], que revisa o design de uma bengala inteligente equipada com tecnologia de detecção e identificação de obstáculos utilizando inteligência artificial.

Diante desse cenário e das soluções em desenvolvimento, o presente artigo apresenta uma tecnologia assistiva voltada para a acessibilidade, explorando o uso de sistemas de reconhecimento de objetos e audiodescrição de imagens para auxiliar na mobilidade por meio do mapeamento de obstáculos. O objetivo é proporcionar maior independência e qualidade de vida para esse público. A estrutura do restante do artigo é organizada da seguinte forma: na seção Estudos Relacionado é discutido os algoritmos e tecnologias existentes que podem contribuir para o desenvolvimento da solução proposta. A seção Metodologia apresenta como se dará o desenvolvimento, bem como o cronograma e os método de avaliação. Em Desenvolvimento da Proposta discutimos sobre a arquitetura projetada, bem como os desafios de implementação. Em seguida abordamos os Testes e Resultados, na Seção 5, e por fim nas Seções 6 e 7, apresentamos a Conclusão do projeto e finalizamos propondo uma série de Trabalhos Futuros.

2. Estudos Relacionados

Durante a trajetória para a concepção do projeto, diversos artigos acadêmicos e publicações foram essenciais para o desenvolvimento de suas ideias centrais, especialmente no contexto de audiodescrição de ambientes para pessoas com deficiência visual.

O desenvolvimento de sistemas assistivos baseados em Deep Learning para pessoas com deficiência visual tem avançado significativamente, com pesquisas como a [Usman Masud 2022] focando na detecção de objetos com áudio descritivo para proporcionar maior autonomia e segurança. Enquanto isso, [Osama et al. 2021] propõe um sistema móvel que utiliza a arquitetura MobileNet [Howard et al. 2017] para a detecção de cédulas e também inclui detecção de roupas em suas funcionalidades, destacando-se pela praticidade em dispositivos portáteis. A relevância dessas pesquisas está na integração de soluções eficientes, como a aplicação do YOLO para detecção de objetos em tempo real [Redmon et al. 2016], e o uso do MobileNet [Howard et al. 2017], que permite o funcionamento adequado em dispositivos móveis com recursos limitados. Essas abordagens são fundamentais para o desenvolvimento de sistemas acessíveis e responsivos, garantindo uma experiência fluida e eficiente aos usuários.

Essas abordagens tecnológicas serviram de ponto de partida para o projeto, visto que este visa proporcionar maior autonomia a pessoas com deficiência visual por meio de um sistema assistivo que transforma informações visuais em descrições auditivas. A integração do YOLO para detecção eficiente de objetos e do MobileNet para otimização em dispositivos móveis, conforme discutido nas pesquisas mencionadas, reforça a proposta de um sistema ágil e acessível, capaz de funcionar em tempo real em smartphones, garantindo uma experiência de navegação segura e intuitiva. O uso dessas tecnologias otimiza a detecção de obstáculos e objetos no ambiente e permite a execução do sistema em dispositivos com recursos limitados, alinhando-se aos objetivos de portabilidade e acessibilidade do projeto.

3. Metodologia

3.1. Planejamento

A implementação de uma metodologia eficiente para o desenvolvimento de uma solução de assistência à mobilidade para deficientes visuais é fundamental para garantir uma abordagem organizada e bem estruturada.

Neste artigo, a metodologia de desenvolvimento consistiu, inicialmente, na elaboração de um cronograma detalhado do projeto, distribuindo as tarefas dentro do prazo estipulado. Em seguida, o foco foi na implementação das tecnologias e algoritmos propostos. Por fim, foram realizados testes seguidos de ajustes, utilizando métricas como velocidade de resposta, qualidade e desempenho na identificação de obstáculos em percursos simulados.

Cada uma dessas etapas será detalhada nas subseções a seguir, começando com o cronograma do projeto, seguido pela descrição das atividades de desenvolvimento e, por fim, a explicação do método de avaliação adotado.

3.2. Cronograma

A tabela 2 apresenta a estruturação do cronograma do projeto SoundEyes, que foi organizado de forma a garantir a entrega do projeto dentro do período de quatro *Sprints*, que contemplam em média blocos de 4 semanas, no intervalo dos meses de Setembro a Dezembro de 2024 e delimitam as respectivas datas de entregas de partes da solução.

Table 1. Cronograma de Trabalho

Sprints	Tópico	Data de Início	Data de Término
1	Estudar Trabalhos Relacionados	10/09/2024	29/09/2024
	Estudar Algoritmos Selecionados	15/09/2024	29/09/2024
	Desenvolver Processamento de Imagens	29/09/2024	18/10/2024
2	Desenvolver Aplicação Mobile	29/09/2024	18/10/2024
	Implementar Protótipos	06/10/2024	18/10/2024
	Realizar Testes Inicias	18/10/2024	22/10/2024
3	Realizar Ajustes e Rever Estratégias	21/10/2024	02/11/2024
	Realizar Novos Testes	02/11/2024	06/11/2024
	Concluir Ajustes e Desenvolvimento	05/11/2024	14/11/2024
4	Validar Proposta	16/11/2024	01/12/2024

De acordo com o cronograma, a primeira etapa do projeto envolve um levantamento bibliográfico detalhado, com o objetivo de identificar algoritmos e estudos relacionados à identificação de imagens. Em seguida, há uma fase de estudo aprofundado, focando na compreensão dos algoritmos mais relevantes para o tema, o que permite a escolha daqueles que melhor se adequam à solução de reconhecimento e à arquitetura mobile proposta. As etapas seguintes concentram-se na implementação dos processadores de imagem e no desenvolvimento da aplicação mobile, uma fase crucial para a elaboração da solução.

Na sequência, ocorre a criação de protótipos que validam a eficiência da solução desenvolvida, por meio de uma bateria de testes, cujo objetivo é identificar possíveis

falhas e oportunidades de melhoria. Com base nesses testes, ajustes são realizados, resultando em uma solução mais robusta e eficaz para os objetivos do projeto. Por fim, a validação da proposta é essencial para mensurar o nível de valor gerado pelo audiodescritor de ambientes SoundEyes.

3.3. 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 uso de memória, tempo de resposta e confiabilidade/qualidade da resposta.

- **Tamanho da Imagem:** A resolução da imagem impacta diretamente no tempo de resposta de envio de imagens e no tempo de classificação. Para estes casos, são 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 512 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.

4. Desenvolvimento da Proposta

4.1. Arquitetura Proposta

A solução SoundEyes está dividida em dois ambientes (Ambiente de Captura e Ambiente de Processamento e Resposta) e três blocos principais (Captura, Processamento e Audiodescrição). 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.

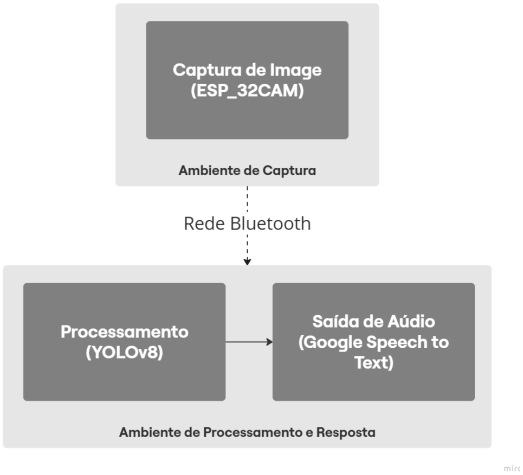

Figure 1. Fluxograma - Representação da Arquitetura do Sistema

- **Bloco de Captura**: O bloco de captura de imagem é implementado no ambiente de captura, que é constituido por um módulo ESP32CAM e uma câmera OV2640, com capacidade de captura de até 2 megapixels. Para facilitar captura e envio em tempo real via bluetooth, a câmera é configurada para capturar uma sequência contínua de frames em duas diferentes resoluções priorizando diferentes atributos (Velocidade e Qualidade). Para velocidade a captura da imagem é limitada a 0,5 megapixels em resolução HVGA, enquanto para qualidade foi definida a resolução XGA com 1 megapixel.

Figure 2. Módulo ESP32CAM com Câmera OV2640.

- **Processamento de Imagens**: Para a detecção de objetos, o projeto utiliza o modelo YOLO (You Only Look Once), que se destaca pela capacidade de detectar múltiplos objetos em tempo real com alta precisão. Foi utilizada a versão YOLOv8, devido às suas melhorias em relação às versões anteriores, como maior precisão e menor latência no processamento. O processamento é realizado em etapas: primeiramente, as imagens capturadas são convertidas em formato compatível com o modelo de detecção. Em seguida, o YOLOv8 processa os dados,

gerando uma lista de objetos detectados com suas respectivas localizações e probabilidades de detecção.

- **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**: A integração dos diferentes componentes é realizada de forma a garantir a comunicação contínua entre as etapas de captura, processamento e resposta. O sistema é configurado para funcionar em ciclos, onde a captura de imagens, o processamento e a geração de áudio ocorrem em sequência, permitindo a atualização constante das informações fornecidas ao usuário. Para a integração dos diferentes ambientes, é utilizada um rede Bluetooth BLE.

4.2. Desafios de Implementação

Durante a implementação do sistema, alguns desafios de captura, processamento e integração foram encontrados, elencamos nos tópicos que se segue os principais:

- **Condicionamento da Iluminação**: O sistema apresentou redução na precisão em ambientes com pouca luz ou iluminação variável, exigindo ajustes nos parâmetros de detecção.
- **Objetos Pequenos ou Ocultos**: Objetos menores ou parcialmente escondidos apresentaram dificuldades na detecção, sendo frequentemente ignorados pelo sistema.
- **Integração de Áudio em Tempo Real**: Garantir a sincronia entre a detecção e a geração de áudio exigiu otimização do processamento e técnicas para reduzir a latência.
- **Consumo de Recursos**: A execução em dispositivos móveis trouxe desafios relacionados ao consumo de CPU e memória, necessitando adaptações para manter o desempenho sem comprometer a autonomia do dispositivo.
- **Precisão na Identificação**: Diferentes versões do YOLO robustez e treinamento distintos, dessa forma, apresenta-se uma situação de *tradeoff* entre a capacidade de identificação e desempenho entre os modelos.

5. Testes e Resultados

Para validar a eficácia do sistema, foram realizadas várias experimentações em diferentes ambientes, com o objetivo de avaliar a precisão da detecção e a clareza das descrições fornecidas.

5.1. Descrição do Ambiente de Teste

Para o funcionamento da arquitetura proposta, foram utilizados os seguintes dispositivos:

Table 2. Especificações do dispositivo Android

Dispositivo Android	Características
Nome	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

Table 3. 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

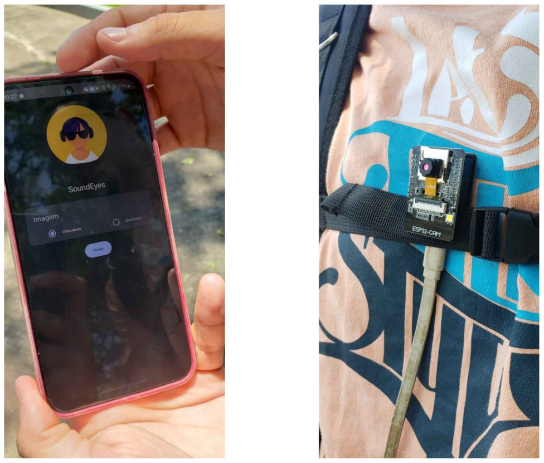

Figure 3. Dispositivos Utilizados no Teste

5.2. Detecção de Objetos

Os testes foram realizados em ambientes externos e em videos da internet, com diferentes condições de distância dos objetos. A Figura 4 mostra exemplos da saída gerada pelo sistema YOLOv8, onde múltiplos objetos foram detectados e identificados.

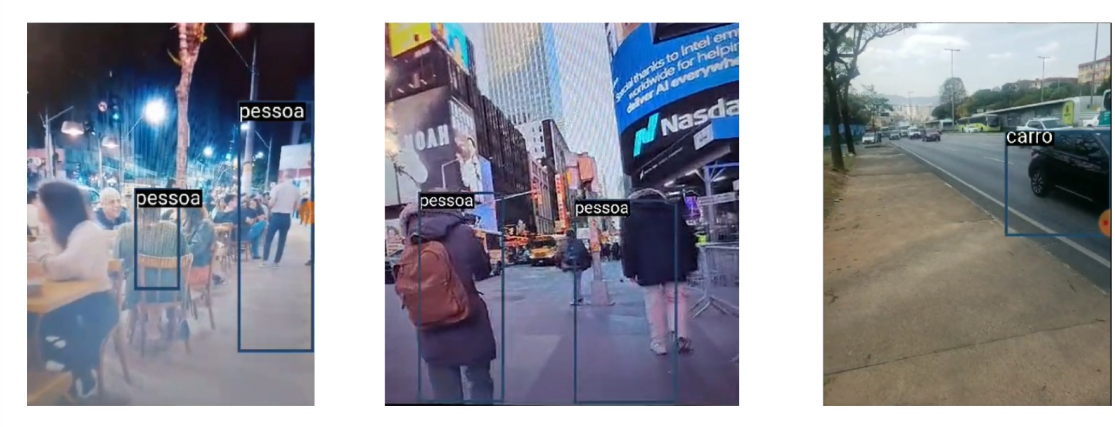

Figure 4. Saída da detecção de objetos utilizando YOLOv8.

5.3. Verificação de Posicionamento

A segunda etapa do processamento consiste em definir a localização do objeto 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.

Por exemplo, na Figura 5, o veículo detectado estaria classificado como à Direita, enquanto na Figura 6, as pessoas seriam classificadas como localizadas à Esquerda e à Direita, respectivamente.

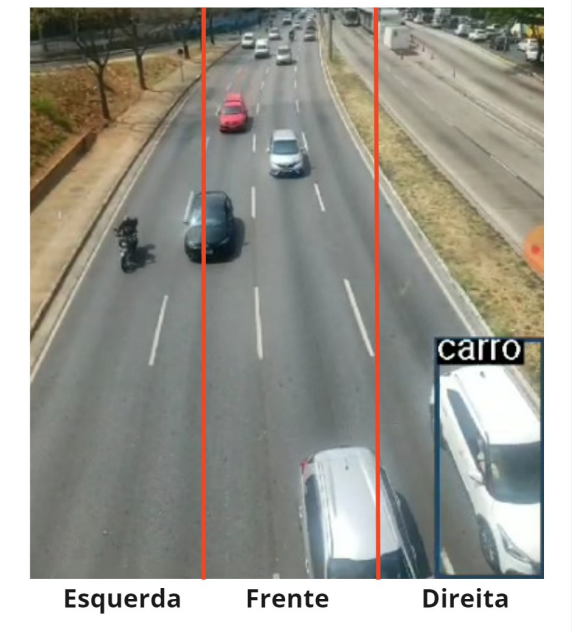

Figure 5. Carro detectado à direita

Figure 6. Pessoa detectada à esquerda e pessoa detectada à direita

5.4. Funcionamento e Retorno do Áudio

Após a definição do objeto e seu posicionamento, o sistema gera uma descrição auditiva dos itens identificados. As descrições são ordenadas de acordo com a relevância e a proximidade dos objetos ao usuário. Durante os testes, os usuários reportaram que o sistema foi útil para identificar obstáculos e objetos de interesse no ambiente. A latência entre a captura da imagem e a geração do áudio foi minimizada para menos de um segundo, proporcionando uma experiência quase em tempo real.

5.5. Resultados Obtidos

Nesta seção, é observado os dados da tabela 4, em que apresenta os impactos dos modos de captura e envio de imagens nas métricas de tempo de resposta e classificação do YOLOv8, conforme descrito na seção 3.3, Método de Avaliação. No modo Desempenho (HVGA), com imagens de 430x320 pixels, o menor tamanho da imagem resulta em processamento mais rápido, com um tempo total de aproximadamente 0,9 segundos. No entanto, a ampliação necessária para o frame de 640x640 do YOLOv8 compromete a precisão da identificação, introduzindo ruídos que dificultam a classificação adequada.

Já no modo Qualidade (XGA), com frames de 1024x768 pixels, a maior resolução proporciona uma classificação mais precisa, reduzindo a necessidade de ajustes no tamanho da imagem para o modelo. Contudo, o tempo de processamento ultrapassa 2 segundos devido ao maior volume de dados transmitidos e a compressão necessária para envio e processamento. Esses resultados destacam a relação de compromisso entre velocidade e precisão, exigindo um equilíbrio adequado conforme as necessidades da aplicação.

Table 4. Desempenho e qualidade da imagem nas resoluções HVGA e XGA

Descrição	Desempenho (HVGA)	Qualidade (XGA)
Tamanho da Imagem	8.350 bytes	43.400 bytes
Número de Pacotes	17	85
Classificação da Imagem	236,6 ms	341,4 ms
Recebimento de Imagem	597,8 ms	2.352,2 ms

6. Conclusão

As pessoas com deficiência visual enfrentam diversas dificuldades, como mencionado anteriormente. Nesse contexto, os avanços tecnológicos têm se mostrado ferramentas valiosas para promover maior acessibilidade e inclusão. Este artigo apresentou uma solução de *Edge Computing* para ambientes móveis, integrando conceitos de arquitetura de computadores, sistemas operacionais e redes, com o objetivo de desenvolver um assistente de locomoção para pessoas com deficiência visual.

Os resultados demonstram que a capacidade do SoundEyes de detectar objetos e oferecer orientação por comandos de voz torna-o uma ferramenta inovadora e eficaz, capaz de proporcionar maior segurança e confiança na mobilidade. Concluímos, portanto, que o SoundEyes se destaca como uma tecnologia promissora no auxílio à locomoção de pessoas com deficiência visual, com grande potencial para fomentar autonomia e inclusão.

7. Trabalhos Futuros

É fundamental continuar explorando as tecnologias como aliadas na promoção da inclusão e acessibilidade para pessoas com deficiência. Nesse sentido, o trabalho apresentado pode ser ampliado para diversos contextos, como a detecção de quedas em pessoas com deficiência locomotora e a tradução de linguagem de sinais para áudio. No caso de deficiência visual, há oportunidades para aplicações como a escolha de vestimentas, descrição de expressões faciais e paisagens.

Para alcançar esses avanços, é necessário aprofundar os processos de classificação, transmissão e precisão na identificação de objetos em imagens. Isso inclui a expansão das classes de treinamento, além de melhorias na eficiência do envio de imagens por meio de redes Bluetooth ou internet.

8. Código Desenvolvido

O código fonte do processador de imagem e da aplicação móvel desenvolvidos neste trabalho está disponível no repositório*GitHub* da equipe.

References

[Google LLC 2024] Google LLC (2024). Google Text-to-Speech (gTTS) API. Acesso em: 19 out. 2024.

[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.

[Niranjan Balu 2019] Niranjan Balu, Y. A. (2019). A design review of smart stick for the blind equipped with obstacle detection and identification using artificial intelligence.

[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 *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 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.

[Usman Masud 2022] Usman Masud, Fathe Jeribi, M. A. F. A. A. T. M. Y. N. (2022). Deep learning based audio assistive system for visually impaired people.

[World Health Organization 2019] World Health Organization (2019). *World Report on Vision*. World Health Organization, Geneva, Switzerland. Acesso em: 19 set. 2024.