Aplicação Assistiva para Audiodescrição de Imagens

**Iury Batista de Andrade Santos, Brendel Francisco Lima Santos,
Antonio Lucas de Almeida, Hiury Machado da Silva,
Paulo David Almeida da Silva, Mikaele Costa Mendonça,
Alcides Xavier Benicasa.**

[1] Departamento de Sistemas de Informação – Universidade Federal de Sergipe (UFS)
Itabaiana – SE – Brasil

iurybas@gmail.com, brendelsantos@gmail.com, lucas.d1995@gmail.com

hiury97@gmail.com, pdavidalmeida@gmail.com

mikaelemendonca@gmail.com, alcides@ufs.br

Abstract. *The inclusion of persons with disabilities in society is very important for the formation a broad and plural society. Due to several technological advances such as the miniaturization of computational devices and powerful artificial intelligence techniques, this work proposes the development of application to audiodescription images in real time, which allows that the persons with visually impaired to obtain information about the environment around them in the form of natural language sentences. Experiments were carried out demonstrating comparisons and analyzes between human descriptions and those generated by the application. The results were satisfactory, because informative sentences were obtained about the scene.*

Resumo. *A inclusão das pessoas com deficiência na sociedade é de grande importância para a formação de uma sociedade ampla e plural. Devido aos diversos avanços tecnológicos como a miniaturização de dispositivos computacionais e poderosas técnicas de inteligência artificial, o presente trabalho propõe o desenvolvimento de uma aplicação para audiodescrição de imagens em tempo real, permitindo que as pessoas com deficiência visual obtenham informações a respeito do ambiente ao seu redor no formato de sentenças em linguagem natural. Experimentos foram realizados demonstrando comparações e análises entre descrições humanas e as geradas pela aplicação. Os resultados foram satisfatórios, uma vez que, foram obtidas sentenças informativas a respeito da cena.*

1. Introdução

A deficiência é uma condição que atinge, segundo dados do Relatório Mundial sobre Deficiência, publicado no ano de 2011, cerca de 15% da população mundial. No Brasil, segundo censo 2010 realizado pela Instituto Brasileiro de Geografia e Estatística (IBGE), 23,9% das pessoas consultadas declaram ter algum tipo de deficiência dentre as investigadas, totalizando mais de 45 milhões de pessoas (DEMOGRÁFICO, 2010).

Ao longo das últimas décadas uma série de medidas foram tomadas visando incluir e destacando as questões sobre pessoas com deficiência, a exemplo de documentos internacionais como o Programa Mundial de Ações Relativo a Pessoa Com Deficiência (1982)

e as Regras Gerais Sobre Igualdade de Oportunidade para as Pessoas com Deficiência (1993), tratando este tema como um aspecto importante à causa dos direitos humanos (ORGANIZATION et al., 2011).

Recentemente, graças a avanços em *hardware* – com a miniaturização de dispositivos computacionais móveis, ao mesmo tempo que o poder computacional é incrementado, e avanços em técnicas de processamento de imagens e Inteligência Artificial, com destaque ao amadurecimento de abordagens que lidam com grande conjunto de dados e requerem alto poder computacional, chamadas comumente de Aprendizado Profundo, dispositivos de tecnologia assistiva cada vez mais dinâmicos e realizando atividades complexas vêm sendo propostos e desenvolvidos.

A utilização de técnicas de aprendizado profundo, a exemplo das Redes Neurais Convolucionais (CNNs) e Redes Neurais Recorrentes (RNN) que lidam, respectivamente, com dados topológicos espaciais (imagens, vídeos etc) e dados sequenciais (sentenças linguísticas, dados relacionados) vêm sendo adotadas em diversos trabalhos para descrição de imagens em linguagem natural (VINYALS et al., 2015; TRAN et al., 2016; KARPATHY, 2016).

Com tais conceitos em mente, este trabalho se propõe ao desenvolvimento de uma aplicação assistiva voltada para pessoas com deficiência visual, que realize a audiodescrição de ambientes e seus objetos componentes em tempo real, por meio da captura de imagens a partir de um *smartphone*, utilizando técnicas de inteligência artificial para análise de imagens e geração automática de descrições em linguagem natural, objetivando oferecer uma concepção que permita a interpretação e entendimento do ambiente ao redor do usuário, de maneira próxima a de um ser humano. Adicionalmente, por se valer de um dispositivo móvel do tipo *smartphone*, torna-se possível a proposta de uma tecnologia assistiva acessível à grande maioria das pessoas.

2. Trabalhos Relacionados

Trabalhos relacionados a proposição e desenvolvimento de TAs incluem a existência de dispositivos físicos específicos ou apenas proposição de métodos computacionais. Com vistas sobre tal conceito, são apresentados nesta seção trabalhos que envolvem métodos unicamente algorítmicos, assim como abordagens que também valem-se da proposição de dispositivos específicos. Adicionalmente, foram incluídos modelos não diretamente desenvolvidos com objetivos de TA, mas que enquadram-se a solução do problema e contemplam em suas propostas gerais o uso para fins de acessibilidade.

O recurso *Automatic Alternative Text*[1], disponibilizado pelo Facebook, realiza a análise automática de imagens postadas pelos usuários da rede, extraindo objetos presentes na mesma e disponibiliza-os de maneira legível por meio de leitores de tela.

De maneira semelhante, o trabalho de Johnson, Karpathy e Fei-Fei (2016) utiliza CNN e RNN para gerar descrições atômicas de uma cena, isto é, a imagem é particionada em diversas descrições sucintas de seus principais componentes individualmente. Tais descrições possuem um nível de informação ligeiramente superior em relação ao *Automatic Alternative Text*, a medida que estabelece características aos objetos e relaciona-

[1]Texto referente ao início da disponibilização aos usuários disponível em: https://pt-br.facebook.com/accessibility/videos/1082033931840331/

mentos.

Objetivando uma abordagem mais rica, avaliando-se a partir do referencial de proximidade se comparado a descrições como realizadas por humanos, alguns trabalhos buscaram a geração de descrições por meio de linguagem natural (LN), a exemplo de Vinyals et al. (2015) e Tran et al. (2016). Em ambos a identificação de elementos nas imagens é realizado por meio de uma CNN, diferenciando pelo modelo arquitetural utilizado. Adicionalmente, a rede responsável pelas construção das sentenças linguísticas para o trabalho de Vinyals et al. (2015) é uma rede do tipo RNN, diferentemente do trabalho de Tran et al. (2016). A Tabela 1 exibe um comparativo entre os trabalhos citados anteriormente.

Tabela 1. Trabalhos que realizam análise e descrição de cenas em imagens

Título do Trabalho/Aplicação	Autor(es)	Ano	Tecnologia Assistiva	Descrições em LN
Automatic Alternative Text	Facebook	2016	Sim	Não
DenseCap	Johnson, Karpathy e Fei-Fei	2016	Não	Não
Show and Tell	Vinyals et al.	2015	Não	Sim
Caption Bot	Tran et al.	2016	Não	Sim

O desenvolvimento de dispositivos de TA voltados as pessoas com deficiência visual, os quais incluem a presença de hardware específico, também se valem de métodos de processamento de imagem e IA.

Em Hicks et al. (2013) (*VA-ST Smart Specs*) e (LEWIS et al., 2013) (*eSight*), foram propostos o desenvolvimento de dispositivos que realizam a captura em tempo real de imagens, aplicando métodos de processamento de imagens de acordo com as necessidades requisitadas pelo usuário e apresentando-as em visores LED. Os dispositivos tem por objetivo auxiliar pessoas com baixa visão, daltonismo e demais deficiências que envolvam dificuldades em enxergar contrastes, enxergar à distância, realizar leitura de textos, diferenciar cores, entre outras, devendo, por tanto, estarem em níveis de acuidade visual superior aos níveis de cegueira.

Outra abordagem é ofertada pelo dispositivo *OrCam MyEye*, desenvolvido por N´aman, Shashua e Wexler (2012). Diferentemente dos demais dispositivos citados acima, o *OrCam MyEye* é composto por uma câmera de vídeo anexável a haste de qualquer óculos convencional, acompanhada por um microcomputador portátil e um dispositivo auditivo. Por meio de técnicas de inteligência artificial, o *OrCam MyEye* realiza atividades avançadas de reconhecimento de comandos por meio de gestos (um dedo apontando para algo) realizados pelo usuário e capturados pela câmera.

O *OrCam MyEye* pode ser utilizado por usuários que se enquadram nas características alvo dos demais trabalhos relacionados acima, bem como aqueles que não possuem visão residual, à medida que o mesmo não exibe nenhum tipo de informação em telas, mas sim em formato de áudio.

Os dispositivos *VA-ST Smart Specs*, *eSight* e *OrCam MyEye*, por se tratarem de produtos específicos, possuem valores que podem ser considerados proibitivos para considerável parcela de usuários com deficiência visual. Na Tabela 2, é realizado um comparativo entre estes, incluindo-se dentre outros fatores o valor de aquisição.

Tabela 2. Dispositivos de TA que usam técnicas a partir da captura de imagens

Título do Trabalho/Dispositivo	Autor(es)	Ano	Aplica IA	Requer Visão Residual	Valor
VA-ST Smart Specs	Hicks et al.	2013	Não	Sim	$1000 [2]
eSight	Lewis et al.	2013	Não	Sim	$10000 [3]
OrCam MyEye	Na'aman, Shashua e Wexler	2012	Sim	Não	$3500 [4]

Em vista dos trabalhos observados e suas funcionalidades, métodos para descrição de imagens e dispositivos assistivos não possuem uma solução integrada. Assim sendo, busca-se realizar a junção das capacidades apresentadas nos trabalhos anteriormente expostos para descrição de imagens e dispositivos assistivos portáteis, em uma única aplicação.

Desta forma, este trabalho visa oferecer descrições em tempo real, providas em linguagem natural, de acordo com o ambiente ao redor do usuário, oferecendo uma nova abordagem de aplicação assistiva, permitindo a interpretação de cenas por pessoas com deficiência visual, de maneira natural. Para tanto, é proposto uma aplicação integralizando diversas técnicas, voltada ao usuário final, especificamente a um aparelho do tipo *smartphone*, que demonstra a acessibilidade proposta. A descrição detalhada do modelo proposto é apresentada a seguir.

3. Biblioteca *Neuraltalk*

O relacionamento entre imagens e linguagem natural é um campo de estudo que desperta interesse dentro da área de Inteligência Artificial, apresentando-se como um desafio de grande complexidade. Devido aos avanços possibilitados pelas técnicas de CNNs e RNNs, foram propostos modelos valendo-se de tais métodos em busca de resultados mais naturais e inteligentes no tocante a obtenção de descrições linguísticas para cenas complexas, a exemplo do trabalho realizado por Karpathy (2016), o qual é utilizado pelo presente trabalho como biblioteca para descrição de imagens.

A biblioteca *NeuralTalk* [5], fruto do referido trabalho citado anteriormente, foi implementada utilizando a linguagem *Lua* e disponibilizado para uso livremente [6]. O modelo arquitetural é composto por uma CNN e RNN. O modelo utiliza uma CNN para extrair informações visuais presentes em uma imagem, como objetos, características e relacionamentos. Os resultados desta são posteriormente utilizados por uma RNN, responsável por gerar uma sentença linguística descritiva da imagem.

O trabalho de Karpathy (2016) permite a geração de descrições de imagens por meio de sentenças linguísticas de maneira bastante satisfatória. Devido a interligação de técnicas como CNNs e RNNs, o modelo permite obter descrições em linguagem natural com alguns resultados próximos de descrições dadas por humanos. Considerando a

[2]Valor obtido em: https://www.technologyreview.com/s/538491/augmented-reality-glasses-could-help-legally-blind-navigate/ , acessado em 03 de Abril de 2017

[3]Valor obtido em: http://www.theverge.com/circuitbreaker/2017/2/16/14637804/esight-3-augmented-reality-headset-legally-blind-see , acessado em 03 de Abril de 2017

[4]Valor obtido em: https://www.washingtonpost.com/news/worldviews/wp/2016/05/21/could-a-new-smart-cam-designed-for-the-blind-help-my-dyslexic-daughter/ , acessado em 03 de Abril de 2017.

[5]O código utilizado neste trabalho é o presente em *NeuralTalk2*, possuindo modificações em relação ao *NeuralTalk* como descrito em (KARPATHY, 2016). No entanto, o modelo conceitual e arquitetural permanecem consistentes.

[6]Disponível em https://github.com/karpathy/neuraltalk2

disponibilização do código-fonte da biblioteca, materiais permissíveis para sua utilização, acesso a artigos e trabalho dissertativo descrevendo a mesma, bem como motivado pela sua abordagem mais inteligente e natural, a *NeuralTalk* foi adotada por este trabalho como biblioteca para a geração de descrições de imagens.

4. Modelo

Uma vasta gama de informações do ambiente ao nosso redor são de difícil ou impossível acesso as pessoas com deficiência visual. O contexto do ambiente é importante para o entendimento pleno do cenário ao qual estamos inseridos, bem como para o desfrute deste e a participação social integral. As pessoas com deficiência visual encontram-se constantemente em situação desfavorável quando em ambientes onde informações são providas exclusivamente de maneira visual e em muitas situações de interações sociais das quais não participam plenamente.

Buscando prover um contexto mais amplo sobre o ambiente e as características de uma cena de maneira acessível a uma pessoa com deficiência visual, a aplicação aqui desenvolvida realiza a análise do ambiente por meio do uso de técnicas de Inteligência Artificial, análise de imagens e transcrição de texto-em-fala, fornecendo assim informações pertinentes daquele por meio da audiodescrição.

4.1. Arquitetura do Modelo

O modelo é composto por uma aplicação cliente, executada em um dispositivo móvel do tipo *smartphone*, e outro conjunto de programas executados em um computador que desempenha função de servidor, como visualizado na Figura 1. A aplicação presente no dispositivo móvel é responsável pela captura das imagens do ambiente e receptação da descrição desta. Ainda no dispositivo é realizado o processo de transcrição de texto-em-fala por meio das funcionalidades oferecidas pela própria plataforma operacional do aparelho.

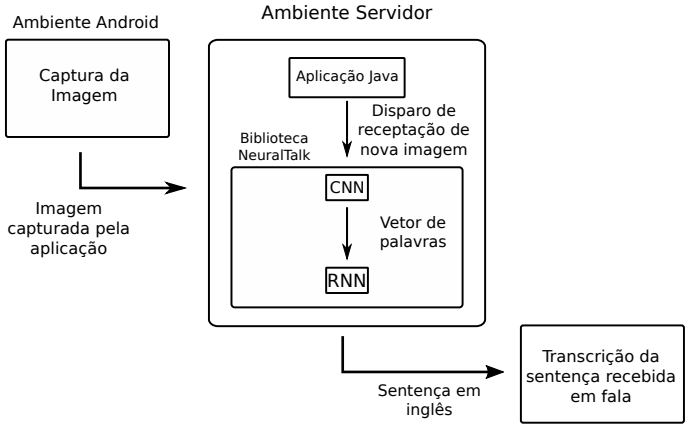

Figura 1. Diagrama do ciclo de vida da solução proposta. O ciclo se repete indefinidamente

No servidor são executados os procedimentos de processamento da imagem capturada enviada pelo aplicativo a partir do aparelho do usuário. A imagem é recebida pela aplicação Java, sendo posteriormente analisada por meio da biblioteca *NeuralTalk*, proposta por (KARPATHY, 2016), onde se é obtida uma descrição no formato de uma

sentença linguística e textual, contendo informações a respeito da cena de maneira geral ou, em alguns casos, de objetos específicos, seus posicionamentos e características.

O esquema apresentado na Figura 1 é repetido indefinidamente até o encerramento da aplicação por parte do usuário, sendo realizada a captura de imagens de maneira automática dado um intervalo de tempo (atualmente definido em 15 segundos) considerado suficiente para a execução de todos os processos realizados.

4.2. Desenvolvimento

A aplicação cliente foi desenvolvida para a plataforma *Android*. Foi desenvolvida uma interface simples e acessível, requisito essencial para a aplicação. Por meio do acionamento de um único botão, todo o processo é iniciado, desde a captura da imagem até a descrição em som para o usuário. O Algoritmo 1 descreve o processo de envio da imagem para o ambiente servidor.

No ambiente servidor (Figura 1(Ambiente Servidor)), o processo foi realizado por meio de duas aplicações distintas, trabalhando de maneira sincronizada. A primeira aplicação foi desenvolvida na linguagem *Java*, sendo responsável pela receptação da imagem enviada pela aplicação cliente, gravado-a em disco e, por fim, disparando um sinal para a aplicação executora dos métodos de Inteligência Artificial, que se encontra em espera. A segunda aplicação é a biblioteca *NeuralTalk*, a qual é escrita na linguagem *Lua* e executada independentemente. Esta última é responsável por aguardar o sinal de disparo enviado pela aplicação em *Java*, realizando o processamento da imagem recebida por esta e gerando a descrição da mesma.

Algoritmo 1: Rotina de descrição da aplicação cliente

```
Aguarda inicialização da aplicação cliente
while AppCliente_inicializada do
    Captura da imagem na aplicação cliente
    Envio da imagem para o servidor
    Recepção da descrição gerada pelo servidor
    Transcrição de texto para voz da descrição obtida
    Pausa câmera por tempo mínimo
end
```

Modificações foram realizadas na biblioteca *NeuralTalk* com o intuito de torná-la mais dinâmica e impedir a realização de processos custosos em tempo e recursos computacionais. Originalmente, a biblioteca se mostra apta a descrever um conjunto de imagens localizadas em um diretório previamente informado. Devido ao uso de bibliotecas adicionais e implementação de uma CNN e RNN, a inicialização da aplicação é de significativo custo computacional, tanto no tangente a processamento, como a memória, mais intensamente desta última, devido ao grande número de parâmetros requiridos pelas redes neurais. A alocação destes recursos exige um tempo considerável quando julgada sua execução repetida no contexto da aplicação deste trabalho, considerando-se que, originalmente, após realizar a descrição das imagens contidas no diretório definido, a aplicação *NeuralTalk* encerrava seu processo, desalocando a memória utilizada.

Para contornar tal problema, a biblioteca foi modificada para executar a classificação de imagens apenas quando disparado o sinal da aplicação *Java*, mantendo-se

Algoritmo 2: Biblioteca *NeuralTalk* modificada para execução mediante sinal da aplicação *Java*

```
while True do
    Aguarda sinal da aplicação java
    if Sinal_recebido then
        Carrega imagem recebida pela aplicação Java
        CNN realiza análise da imagem e identifica elementos
        RNN recebe elementos e gera descrição linguística
        Descrição é enviada para aplicação cliente
    end
end
```

ainda em execução após a conclusão do respectivo processo, aguardando novo disparo até que seja manualmente encerrada, dispensando assim a realocação desnecessária dos recursos e tornando-a mais adequada ao propósito desejado. Ao receber o sinal da aplicação em *Java*, a biblioteca *NeuralTalk* processa a imagem recebida pela aplicação cliente e, como resultado, retorna uma descrição em formato de sentença linguística textual em inglês. O processo de funcionamento da biblioteca *NeuralTalk* com as modificações realizadas está apresentado no Algoritmo 2.

5. Experimentos e Resultados

Os experimentos foram divididos em partes e apresentados nas seguintes subseções. A primeira parte verificou-se a aplicação por meio de uma prova de conceito em ambiente real. Já na segunda, está direcionada ao comportamento do modelo quando aplicado a um conjunto de imagens diversas condizentes ao contexto proposto neste trabalho, aplicação assistiva.

5.1. Experimentos × Ambiente Real

No intuito de validar a aplicação e testar em um cenário de ambiente real, aplicou-se uma PoC (Prova de Conceito). A prova de conceito consistiu em disponibilizar a aplicação móvel para o uso de usuários voluntários em suas próprias residências, sendo supervisionados pela equipe da pesquisa, por intermédio de visitas.

Figura 2. Análise qualitativa (1-4) do modelo proposto em cenas reais próprias, comparado com descrições humanas e descrição gerada.

A cada visita realizada pela equipe, instruía-se o voluntário a utilizar de forma livre à aplicação pelos cômodos da residência. A fim de coletar métricas para posterior avaliação, observou-se os logs fornecidos pela aplicação e uma pequena amostra aleatória

Tabela 3. Avaliação dos Dispositivos

Modelo	SDK	Tempo Inicial	Tempo Final	Diferença
ASUS X008DB	24	08:55:21.870	08:55:22.364	0.494
Moto G XT1069	23	09:55:25.299	09:55:25.958	0.659
Lenovo A6020136	23	11:15:30.307	11:15:30.938	0.631

de imagens, foram coletados. As visitas duravam em média 45 minutos, tempo gasto para preparar a aplicação e disponibilizar para o usuário.

A figura 2 traz uma amostra das imagens obtidas ao longo das visitas. A descrição dada pelo modelo acertou em um alto nível o ambiente que se fazia presente na imagem, isto se deu pelo fato do treinamento da rede neural ter acontecido com imagens contextualizadas no cenário americano. Contudo o modelo se mostrou eficaz em cenários semelhantes a sua base.

Para melhor compreensão da aplicação buscou-se por voluntários com diferentes dispositivos, a tabela 3 traz alguns dados coletados dos *logs*, o tempo inicial refere-se ao tempo que a imagem chega a rede neural, consequentemente o tempo final é referente ao término do processamento realizado para gerar a descrição. Percebesse que entre as versões (SDK) houve uma melhoria para a linha 1 em relação as outras, pelo fato do aperfeiçoamento e correções que cada nova versão traz. Contudo entre versões iguais e modelos distintos a diferença não foi tão relevante.

5.2. Experimentos $\times$ Descrições Humanas

O segundo conjunto de experimentos foi realizado para verificar a qualidade das descrições obtidas em um contexto próprio. Neste caso, descrições foram geradas pela aplicação para imagens coletadas localmente, em ambientes dentro do contexto da aplicação, totalizando 56 imagens, sendo comparadas com descrições fornecidas por humanos. A comparação foi realizada de maneira qualitativa, verificando-se o quanto a mesma se insere dentro do observado pelas descrições humanas. Cada imagem recebeu 5 descrições providas por seres humanos[7]. A quantidade de descrições coletadas é igual aquela presente nos conjuntos de dados utilizados para treinamento da rede. As imagens e suas respectivas 5 descrições providas por humanos e a descrição gerada pelos processos de inteligência artificial são exibidos na Figura 3.

As descrições variam entre graus de precisão próximos aquelas providas por humanos, em alguns momentos até mesmo mostrando-se mais detalhadas. Na Figura 3 (1-3) é possível observar que os principais elementos da imagem são citados e o contexto geral é corretamente descrito, apresentando uma boa representatividade perante à diversidade das imagens e ambientes. No entanto, exemplos com erros foram verificados, como na Figura 3 (4), onde uvas são confundidas com *"donuts"*, comida popular nos países de origem das imagens de treinamento, mas fora do contexto regional local relativo ao Brasil.

De maneira geral, a aplicação *NeuralTalk* apresentou bons resultados quando avaliada com o conjunto inicial, composto por imagens heterogêneas e homogêneas, realizando a percepção dos principais elementos na maioria dos casos observados e de-

[7]As descrições não foram corrigidas ou editadas gramaticamente, na maioria dos casos, assim como o que ocorre nas descrições presentes no conjunto de teste

	Imagem	Descrições Humanas	Processo$_{NeuralTalk}$
(1)		Fazenda com árvores e coqueiros, alguns bois brancos no meio do pasto, na margem da BR. Gado pastando em vegetação verde com coqueiros e árvores próximos e uma serra ao fundo. Uma serra e um campo com animais e algumas árvores. Animais no pasto. Paisagem de árvores, gado e serra ao fundo.	*A herd of animals grazing on a lush green hillside*
(2)		Balcão com micro-ondas branco em cima. Em cima do micro-ondas uma sanduicheira preta, ao lado direito uma garrafa térmica, ao lado esquerdo uma máquina de espremer laranja preta. Quase em frente um pote com açúcar. Recipiente com substância branca, espremedor de frutas, micro-ondas, torradeira e garrafa de café sobre uma bancada de madeira. Alguns eletrodomésticos da cozinha. Micro-ondas branco. Sanduicheira em cima do micro-ondas.	*A microwave oven sitting on top of a counter*
(3)		Geladeira com porta aberta. Dentro tem ovos vermelhos, refrigerantes, cervejas, suco, nescau. Geladeira aberta contendo refrigerantes, cervejas, suco, ovos, leite condensado e verduras. Geladeira aberta e com alimentos. Geladeira aberta com alimentos. Geladeira com alimentos.	*A refrigerator filled with lots of food and drinks*
(4)		Prateleira, pratinhos de uvas, uvas verdes e vermelhas. Uvas embaladas expostas em prateleira de supermercado. Uvas na prateleira do supermercado. Prateleira com uvas. Uvas verdes e vinho a venda.	*A display case filled with lots of donuts*

Figura 3. Análise qualitativa (1-4) do modelo proposto em cenas reais próprias, comparado com descrições humanas e descrição gerada.

screvendo corretamente a cena observada. De maneira especial, é válido notar os resultados do segundo grupamento, onde a mesma identificou placas nos casos onde estes elementos se apresentavam de maneira bastante presente, mesmo sem ter sido treinada de maneira direcionada a realizar tal reconhecimento.

6. Conclusão

A aplicação proposta oferece um produto de tecnologia assistiva diferente dos demais trabalhos apresentados, realizando descrições de uma cena em linguagem natural, em tempo real utilizando um dispositivo móvel como, por exemplo, *smartphones*. Para tanto, realizou-se uma combinação de diversas técnicas de inteligência artificial aplicadas a um problema real conjuntamente com desenvolvimento móvel, demostrando a possibilidade de aplicações assistivas de alto nível a serem exploradas.

Os processos integrados para a composição deste trabalho foram testados se valendo de diversos parâmetros e comparados com resultados humanos. Observou-se nos resultados bons níveis de proximidade aos equivalentes humanos, onde elementos de importância na cena, tanto quando avaliados por meio dos mapas de fixação, como quando se valendo de descrições humanas, foram identificados e citados nas descrições geradas. Ainda, o modelo mostrou-se flexível mesmo quando exposto a cenários diferentes e plu-

rais.

Devido a utilização de biblioteca de terceiro e o custo computacional da mesma, os métodos de inteligência artificial são executados em um ambiente servidor, sendo necessária a comunicação via rede entre os componentes. O desenvolvimento de uma arquitetura CNN/RNN própria, com base em modelos mais compactos, é um trabalho futuro de grande relevância a ser realizado. Tal modelo permitiria a execução de todos os processos localmente no aparelho, dispensando assim a necessidade de qualquer comunicação externa.

Conclui-se que o trabalho se mostrou exitoso em explorar e propor uma abordagem de tecnologia assistiva de alto nível, promovendo descrições a respeito de cenas de maneira semelhante à mesma realizada por humanos, sendo projetada para operar em dispositivos móveis acessíveis e de custos inferiores a outros observados, de carácter inclusivo e permitindo as pessoas com deficiência visual acesso a informações visuais ao seu redor, almejando abrir novas possibilidades de interações sociais, bem-estar e independência.

Referências

DEMOGRÁFICO, I. C. características gerais da população, religião e pessoas com deficiência. *Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística*, 2010.

HICKS, S. L. et al. A depth-based head-mounted visual display to aid navigation in partially sighted individuals. *PLOS ONE*, Public Library of Science, v. 8, n. 7, p. 1–8, 07 2013. Disponível em: ⟨http://dx.doi.org/10.1371\%2Fjournal.pone.0067695⟩.

JOHNSON, J.; KARPATHY, A.; FEI-FEI, L. Densecap: Fully convolutional localization networks for dense captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016.

KARPATHY, A. *Connecting Images and Natural Language*. Tese (Doutorado) — Stanford University, 2016.

LEWIS, C. et al. *Apparatus and method for augmenting sight*. Google Patents, 2013. US Patent 8,494,298. Disponível em: ⟨https://www.google.com/patents/US8494298⟩.

N´AMAN, E.; SHASHUA, A.; WEXLER, Y. *User wearable visual assistance system*. Google Patents, 2012. US Patent App. 13/397,919. Disponível em: ⟨https://www.google.com/patents/US20120212593⟩.

ORGANIZATION, W. H. et al. *World report on disability*. [S.l.]: World Health Organization, 2011.

TRAN, K. et al. Rich image captioning in the wild. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. [S.l.: s.n.], 2016.

VINYALS, O. et al. Show and tell: A neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2015. p. 3156–3164.