Image Descriptions' Limitations for People with Visual Impairments: Where Are We and Where Are We Going?

Alessandra Helena Jandrey
School of Technology - PUCRS
alessandra.jandrey@edu.pucrs.br

Duncan Dubugras Alcoba Ruiz
School of Technology - PUCRS
duncan.ruiz@pucrs.br

Milene Selbach Silveira
School of Technology - PUCRS
milene.silveira@pucrs.br

ABSTRACT

Image descriptions aim to transcribe the visual content and are essential for people who do not have eyesight. Such image descriptions are generated manually or by Artificial Intelligence (AI) models. Despite its relevance, the emergence of automatic image descriptions was not motivated by people with visual impairments; thus, they still cause dissatisfaction in this audience. This paper provides a snowballing review of the limitations of image descriptions for people with visual impairments. We encountered thirteen image description issues, including those analogous to the ethics, such as people's appearances, gender, race, and disability. We identified five reasons why sighted people do not write image descriptions for the content they share, exposing the necessity of accessibility campaigns to awareness people of the social importance of image descriptions. Moreover, we discussed recommendations found in the literature that may support automated tools and sighted people for writing high-quality image descriptions. We hope our results will highlight the social significance of image descriptions and encourage the community to pursue further interdisciplinary researches that could potentially minimize the issues pointed out in our study.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility theory, concepts and paradigms**.

KEYWORDS

image description, visually impaired people, accessibility, snowballing

ACM Reference Format:
Alessandra Helena Jandrey, Duncan Dubugras Alcoba Ruiz, and Milene Selbach Silveira. 2021. Image Descriptions' Limitations for People with Visual Impairments: Where Are We and Where Are We Going?. In *XX Brazilian Symposium on Human Factors in Computing Systems (IHC'21), October 18–22, 2021, Virtual Event, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3472301.3484356

IHC'21, October 18–22, 2021, Virtual Event, Brazil

ACM ISBN 978-1-4503-8617-3/21/10…$15.00
https://doi.org/10.1145/3472301.3484356

1 INTRODUÇÃO

Pessoas com deficiência visual consomem os conteúdos digitais através de Tecnologias Assistivas [13, 25], *e.g.,* leitores de tela e magnificadores de tela, para a leitura das informações textuais disponíveis no computador [13, 23]. Para a leitura de imagem, por exemplo, o leitor de tela depende de textos alternativos que descrevam a informação visual [29], sendo essa uma das recomendações mais básicas das diretrizes de acessibilidade [11, 23, 29, 35]. Apesar dessa recomendação, a maioria das imagens digitais não dispõe de textos descritivos [36] e é inacessível às pessoas com deficiência visual [13, 36]. A ausência dos respectivos textos ocorre, entre outros motivos, por ser uma tarefa laboriosa, já que são incluídos manualmente por aqueles que compartilham os conteúdos visuais [39]. Uma alternativa a isso é a aplicação de modelos de aprendizado de máquina [13] que geram sentenças descritivas automáticas a partir dos elementos identificados nas imagens [1].

A transformação do conteúdo visual para texto tem uma vasta aplicabilidade [3, 30], *e.g.,* produção de interações naturais entre robôs e humanos [3], entendimento de cenas [2], veículos autônomos [2], além de auxiliar pessoas com deficiência visual em suas tarefas diárias [3, 17]. Tal aplicabilidade motiva diferentes linhas de pesquisa a aperfeiçoar os geradores automáticos para se assemelharam à habilidade humana [3]. O maior desafio dessa tarefa é gerar sentenças diversificadas e significativas [6, 10], visto que requer não somente a identificação dos elementos visuais, mas, também, a compreensão de como esses elementos se relacionam e a geração de uma sentença descritiva em linguagem natural [5] como a Língua Portuguesa.

Apesar do impacto social, as pessoas com deficiência visual não foram a motivação inicial do surgimento dos modelos de descrições automáticas [22], cujas sentenças geradas ainda não satisfazem as necessidades desse público [22, 32, 38]. Isso nos motivou a investigar quais são os problemas em descrições de imagens relatados por pessoas com deficiência visual na literatura. Por meio da técnica de *Snowballing*, identificaram-se treze problemas de descrição de imagens, incluindo aqueles análogos à ética, como aparência das pessoas, gênero, raça e deficiência. Também identificaram-se cinco motivos pelos quais pessoas videntes, *i.e.,* sem deficiência visual, não descrevem os conteúdos visuais que compartilham, discutindo, ainda, recomendações encontradas que podem contribuir para a melhoria da qualidade das descrições. Esperam-se que os resultados enalteçam a relevância social das descrições de imagens para as pessoas com deficiência e encorajem a comunidade a prosseguir com pesquisas interdisciplinares que possam minimizar os problemas apontados neste estudo.

Esse trabalho está organizado da seguinte forma: a Seção 2 fundamenta as classificações de descrições de imagens e as abordagens utilizadas para avaliar a qualidade das descrições. A Seção 3

apresenta o planejamento do *Snowballing* e a Seção 4 detalha sua execução. Os resultados obtidos estão discutidos na Seção 5; por fim, a Seção 6 apresenta as considerações finais, as limitações e as perspectivas de trabalhos futuros.

2 DESCRIÇÕES DE IMAGENS

Os seres humanos possuem facilidade em descrever uma imagem [12] ou o ambiente a sua volta [3], podendo fornecer mais detalhes à medida que lhe são solicitados [6]. Para descrever uma imagem, fazemos o uso dos conhecimentos adquiridos, dos saberes do cotidiano e das experiências pessoais, destacando os objetos julgados importantes [17, 18]. Além disso, podemos utilizar da imaginação para descrever de maneira vívida e cativante [17]. Possibilitar que computadores interpretem o mundo visual dessa forma é um objetivo de longa data de pesquisadores da Inteligência Artificial e suas sub-áreas [3].

Ao contrário das descrições humanas, a geração de sentenças automáticas baseia-se nos elementos identificados pelo modelo do gerador [1] e não conforme o que as pessoas estão interessadas em saber sobre uma imagem [6]. Além disso, há muitos fatores influenciadores no que seria uma descrição considerada apropriada por todos [32], *e.g.,* necessidade, conhecimento prévio, tradições, crenças [18] e identidade racial e de gênero [4].

2.1 Tipos de Descrições

As descrições de imagens são classificadas conforme a quantidade de conhecimento requerido e no quão informativas elas são [15, 16, 19]. Segundo as definições de Hollink *et al.* [16], uma descrição **não-visual** refere-se aos metadados da imagem, *e.g.,* hora, fotógrafo e local, sendo essa uma descrição objetiva, *i.e.,* não afetada por qualquer interpretação, contendo informações que não são explicitamente derivadas do conteúdo ou da aparência da imagem. Uma descrição **perceptiva** captura as características dos elementos da imagem, *e.g.,* posição, orientação e distância relativa dos elementos, assim como as propriedades visuais da imagem, *e.g.,* se é uma fotografia ou uma pintura, textura, cor e tamanho, não requerendo conhecimento prévio para interpretar o conteúdo visual. Por fim, uma descrição **conceitual** provê informações sobre o conteúdo da imagem, concentrando-se na cena representada, nos elementos que a compõem, seus respectivos atributos e como esses elementos se relacionam.

Devido às inúmeras maneiras de descrever uma imagem [19], uma descrição conceitual pode ainda ser classificada como abstrata, *i.e.,* descreve a cena genericamente, ou específica, *i.e.,* apresenta mais detalhes sobre o conteúdo da imagem [15, 16, 19]. Conforme James e Chang [19] destacam, uma descrição conceitual é inferida pelo conhecimento prévio e pela interpretação do observador sobre a imagem alvo, portanto, tende à subjetividade. Produzir descrições conceituais é o objetivo dos geradores automáticos, já que estas são significativas para a compreensão do conteúdo visual [15].

2.2 Métodos de Avaliação de Descrições

A qualidade das descrições automáticas é inferida por métricas quantitativas, *e.g.,* BLEU [26], METEOR [8], ROUGE [21] e CIDEr [33], além de pessoas videntes voluntárias. De acordo com Hrga e Ivašić-Kos [17], as respectivas métricas devem satisfazer dois critérios: (a) descrições consideradas boas por humanos devem atingir escores altos e; (b) descrições que atingiram escores altos têm que ser consideradas boas por avaliadores humanos. As avaliações humanas ocorrem através de plataformas online pelas quais os avaliadores normalmente escolhem a opção que melhor descreve o conteúdo visual de uma imagem [28]. Todavia, tais avaliações são pouco viáveis devido ao tempo necessário e ao custo envolvido [17] e são subjetivas, já que cada pessoa descreve uma imagem de maneira única conforme seu conhecimento prévio e sua interpretação [19]. Além disso, as descrições automáticas podem ser avaliadas qualitativamente e buscam explorar as experiências e as opiniões dos avaliadores com as descrições geradas [22, 38].

3 PLANEJAMENTO DO *SNOWBALLING*

Os estudos sistemáticos da literatura, incluindo as revisões e mapeamentos, surgiram como uma forma de sintetizar as evidências, permitindo que pesquisadores chegassem a um entendimento sobre determinada área de pesquisa [37]. Para o presente trabalho, utilizou-se a técnica de *Snowballing* seguindo os procedimentos propostos por Wohlin [37]. A presente Seção engloba a Questão de Pesquisa que motivou a execução deste trabalho (Subseção 3.1), o protocolo definido para a seleção dos estudos (Subseção 3.2), o conjunto inicial de trabalhos (Subseção 3.3) e, por fim, os procedimentos de busca executados (Subseção 3.4).

3.1 Questão de Pesquisa

A definição da Questão de Pesquisa (QP) é a atividade mais importante da etapa de planejamento de uma revisão da literatura [9]. Esse estudo visou responder a seguinte QP: **"Quais são os problemas em descrição de imagens para pessoas com deficiência visual?"**.

3.2 Protocolo de Pesquisa

Essa Seção apresenta os critérios de inclusão, de exclusão e de qualidade adotados para a execução do *Snowballing*:

- **Critérios de inclusão:** consideraram-se os estudos publicados nos locais a seguir cuja escolha ocorreu por serem, na nossa visão, os principais veículos de publicação relacionados aos tópicos de Interação Humano-Computador e Acessibilidade:
  - ASSETS (Conference on Computers and Accessibility);
  - CHI (Conference on Human Factors in Computing Systems);
  - IHC (Brazilian Symposium on Human Factors in Computing Systems);
  - WWW (World Wide Web Conference);
  - CSCW (Conference on Computer Supported Cooperative Work and Social Computing);
  - IJHCI (International Journal of Human-Computer Interaction);
  - PACMHCI (Proceedings of the ACM on Human-Computer Interaction Journal Series).
- **Critérios de exclusão:** rejeitaram-se os estudos seguindo a ordem especificada:

(1) Estudos já selecionados ou duplicados, ou seja, analisados anteriormente na execução do protocolo;

(2) Títulos fora do escopo do *Snowballing*. Essa remoção ocorreu apenas nos casos explícitos, ou seja, quando os títulos indicavam outra temática que não descrições de imagens sob as perspectivas de pessoas com deficiência visual;

(3) Resumos fora do escopo do *Snowballing*;

(4) Introduções, resultados e considerações finais fora do escopo do *Snowballing*;

(5) Não responderam à Questão de Pesquisa.

- **Critérios de qualidade:** para garantir a qualidade do *Snowballing*, seguiram-se os seguintes critérios:

– Para cada estudo desconsiderado, foi atribuído um *motivo de rejeição* com base nos critérios de exclusão definidos. Dessa forma, foi possível identificar a quantidade de estudos rejeitados em cada critério de exclusão;

– Os autores do estudo em análise devem discorrer sobre os problemas apontados, e não apenas mencioná-los brevemente, possibilitando assim, uma discussão em profundidade dos dados extraídos.

3.3 Conjunto Inicial

Segundo Wohlin [37], a técnica de *Snowballing* requer a definição de um conjunto inicial de estudos que responda à Questão de Pesquisa (QP) [37]. O conjunto inicial deste *Snowballing* é composto por seis estudos identificados durante uma investigação preliminar em diferentes bases de dados, incluindo o Google Acadêmico, ACM e IEEE. A seleção dos seis trabalhos, além do escopo e por responderem a QP, ocorreu devido às contribuições e os pontos de vista relatados pelos autores. Ademais, estes estudos são alguns dos mais referenciados nas bases de dados citadas acima, exceto um [29], selecionado após apresentação do mesmo no evento IHC 2020. A Tabela 1 apresenta os seis estudos do conjunto inicial, dos quais ocorreram os procedimentos de busca.

Tabela 1: Estudos pertencentes ao conjunto inicial.

Estudo e Ano	Objetivo	Total de participantes
Morris *et al.* [24] (2016)	Coletar as motivações de uso e as opiniões de pessoas cegas sobre as descrições de imagens no Twitter	112 pessoas cegas
Wu *et al.* [38] (2017)	Avaliar o recurso de geração de descrição de imagem automática do Facebook com pessoas com deficiência visual	4 pessoas cegas e 375 usuários de leitores de tela
MacLeod *et al.* [22] (2017)	Explorar as experiências de pessoas com deficiência visual com descrições de imagens automáticas nas redes sociais	106 pessoas com deficiência visual
Gleason *et al.* [13] (2019)	Explorar o recurso de inserção manual de descrição de imagem no Twitter e entender as motivações das pessoas videntes que incluem descrições em suas postagens	20 pessoas videntes
Stangl, Morris e Gurari [32] (2020)	Investigar as experiências e as preferências de descrições de imagens de pessoas com deficiência visual em diferentes contextos digitais	28 pessoas com deficiência visual
Sacramento *et al.* [29] (2020)	Investigar as dificuldades de acesso aos conteúdos visuais de usuários de leitores de tela e os hábitos de descrições das pessoas videntes	333 pessoas videntes e 100 usuários de leitores de tela

3.4 Procedimentos de busca

Conforme Wohlin [37] expoe, a técnica de *Snowballing* não requer pesquisas em bases de dados com o uso de uma *string*. Dessa forma, a busca por potenciais trabalhos ocorreu por meio da análise das referências, conhecido também como *Backward Snowballing*) e das citações (*Forward Snowballing*) dos seis estudos do conjunto inicial. A técnica de *Snowballing* é realizada por iterações, *i.e.,* para cada novo estudo aceito (selecionado), analisam-se as referências e as citações destes, até que nenhum novo estudo seja selecionado [37].

4 EXECUÇÃO DO *SNOWBALLING*

De acordo com Snyder [31], uma revisão da literatura de qualidade precisa ser replicável, *i.e.,* a execução deve ser descrita de forma que um leitor externo possa replicar o estudo e chegar às conclusões semelhantes. Portanto, a presente Seção discorre detalhadamente sobre os procedimentos executados em **duas iterações**, sendo que a Primeira Iteração é relatada na Subseção 4.1 e a Subseção 4.2 relata a Segunda Iteração.

4.1 Primeira Iteração

Essa iteração refere-se aos seis estudos do conjunto inicial. A execução dos procedimentos de *Backward Snowballing* e *Forward Snowballing* apresentam-se na Seção 4.1.1 e Seção 4.1.2, respectivamente.

4.1.1 Backward Snowballing. Nesse procedimento analisaram-se as referências dos seis estudos do conjunto inicial, sendo que o procedimento de busca retornou 219 potenciais trabalhos. A análise e seleção dos estudos estão expostos a seguir:

O estudo de Morris *et al.* [24] possui 34 referências e rejeitaram-se 22 devido aos critérios de inclusão. Dos 12 estudos restantes, rejeitaram-se 6 deles após a leitura dos títulos e outros 4 estudos após a leitura dos resumos. Efetuou-se necessária a leitura das introduções, dos resultados e das considerações finais dos 2 estudos restantes, dos quais se rejeitou 1 deles. Aceitou-se o único trabalho restante após a sua leitura completa: **Voykinska** *et al.* **[34]**.

O estudo de Wu *et al.* [38] possui 34 referências e rejeitaram-se 21 devido aos critérios de inclusão. A partir dos 13 estudos restantes, removeram-se 5 duplicados, sendo que 2 eram estudos aceitos. Restaram 8 estudos e rejeitaram-se 5 destes após a leitura dos títulos. Realizou-se necessária a leitura das introduções, dos resultados e das considerações finais dos 3 estudos restantes, dos quais se rejeitaram 2 deles. Aceitou-se o único estudo restante após a sua leitura completa: **Lazar** *et al.* **[20]**.

O estudo de MacLeod *et al.* [22] possui 38 referências e rejeitaram-se 28 devido aos critérios de inclusão. A partir dos 10 estudos restantes, removeram-se 6 duplicados, dos quais 3 eram estudos selecionados. Dos 4 estudos restantes, rejeitaram-se todos devido aos títulos fora do escopo do *Snowballing*.

O estudo de Gleason *et al.* [13] possui 37 referências e rejeitaram-se 27 devido aos critérios de inclusão. A partir dos 10 estudos restantes, removeram-se 5 estudos duplicados, sendo que 2 já foram selecionados anteriormente. Dos 5 estudos restantes, rejeitaram-se 3 após a leitura dos títulos. Realizou-se a leitura das introduções, dos resultados e das considerações finais dos outros 2 estudos e, posteriormente, a leitura completa para a tomada de decisão. Após essa

etapa, aceitou-se um deles por responder à Questão de Pesquisa: **Zhao** *et al.* **[39]**.

O estudo de Stangl, Morris e Gurari [32] possui 45 referências e rejeitaram-se 27 devido aos critérios de inclusão. A partir dos 18 estudos restantes, removeram-se 11 estudos duplicados, sendo que 6 já haviam sido selecionados anteriormente. Dos 7 estudos restantes, rejeitaram-se 6 devido aos títulos fora do escopo. Rejeitou-se o único estudo restante após a leitura completa do mesmo por não responder à Questão de Pesquisa.

O estudo de Sacramento *et al.* [29] possui 31 referências das quais se rejeitaram 26 devido aos critérios de inclusão. A partir dos 5 estudos restantes, removeram-se 4 estudos duplicados, sendo todos já aceitos em alguma etapa anterior. Rejeitou-se o único estudo restante devido ao resumo fora do escopo do *Snowballing*.

4.1.2 Forward Snowballing. Nesse procedimento analisaram-se as citações dos seis estudos do conjunto inicial, sendo que o procedimento de busca retornou 267 potenciais trabalhos. A análise foi conduzida utilizando o Google Acadêmico e a coleta de dados ocorreu no dia 26 de janeiro de 2021.

O estudo de Morris *et al.* [24] possui 68 citações e rejeitaram-se 39 devido aos critérios de inclusão. A partir dos 29 estudos restantes, removeram-se 8 estudos duplicados, sendo que 5 foram aceitos em etapas anteriores. Dos 21 estudos restantes, rejeitaram-se 16 devido aos títulos fora do escopo do *Snowballing*. Para os 5 estudos restantes, efetuou-se a leitura das introduções, dos resultados e das considerações finais, dos quais se rejeitaram 3 estudos por não se enquadrarem no escopo do *Snowballing*. Após a leitura completa dos 2 trabalhos restantes, aceitaram-se ambos por responderem à Questão de Pesquisa: **Gleason** *et al.* **[14]** e **Bennett** *et al.* **[4]**.

O estudo de Wu *et al.* [38] possui 100 citações e rejeitaram-se 73 devido aos critérios de inclusão. A partir dos 27 estudos restantes, removeram-se 13 duplicados, dos quais 7 eram estudos já aceitos anteriormente. Dos 14 estudos restantes, rejeitaram-se 9 devido aos títulos e outros 2 devido aos resumos fora do escopo. Após a leitura das introduções, dos resultados e das considerações finais dos 3 estudos restantes, rejeitaram-se 2 por não se enquadrarem no escopo do *Snowballing*. Rejeitou-se o único trabalho restante após a leitura completa por não responder à Questão de Pesquisa.

O estudo de MacLeod *et al.* [22] possui 70 citações e rejeitaram-se 50 devido aos critérios de inclusão. A partir dos 20 estudos restantes, removeram-se 19 duplicados, dos quais 5 já foram selecionados em etapas anteriores. Rejeitou-se o único trabalho restante devido ao título fora do escopo do *Snowballing*.

O estudo de Gleason *et al.* [13] possui 18 citações e rejeitaram-se 6 devido aos critérios de inclusão. A partir dos 12 estudos restantes, removeram-se 12 duplicados, dos quais 4 deles eram estudos anteriormente selecionados. Com isso, não restaram trabalhos a serem considerados.

O estudo de Stangl, Morris e Gurari [32] possui 11 citações e rejeitaram-se 6 devido aos critérios de inclusão. A partir dos 5 estudos restantes, removeram-se 3 duplicados, sendo que 1 deles foi selecionado em alguma etapa anterior. Rejeitaram-se os 2 trabalhos restantes devido aos títulos fora do escopo do *Snowballing*.

O estudo de Sacramento *et al.* [29] não possuía citações no momento da coleta dos dados. O estudo foi apresentado na conferência IHC em outubro de 2020.

Como resultado da Primeira Iteração, selecionaram-se cinco trabalhos. Conforme mencionado anteriormente, é necessário refazer os procedimentos de *Backward* e *Forward Snowballing* para os novos estudos selecionados [37].

4.2 Segunda Iteração

Essa iteração refere-se aos cinco estudos selecionados ao término da Primeira Iteração, conforme apresenta a Tabela 2. A execução dos procedimentos de *Backward Snowballing* e *Forward Snowballing* apresenta-se na Seção 4.2.1 e Seção 4.2.2, respectivamente.

Tabela 2: Estudos aceitos a partir da Primeira Iteração.

Estudo e Ano	Objetivo	Total de participantes
Lazar *et al.* [20] (2007)	Investigar as frustrações das pessoas cegas ao acessar a web	100 pessoas com deficiência visual
Voykinska *et al.* [34] (2016)	Investigar as motivações, os desafios, e as interações de pessoas cegas com conteúdos visuais	71 pessoas com deficiência visual
Zhao *et al.* [39] (2017)	Avaliar o recurso da aplicação *mobile* do Facebook para a geração de descrição de imagem automática	12 pessoas com deficiência visual
Gleason *et al.* [14] (2020)	Avaliar uma extensão de navegador para a geração e recuperação de descrições de imagens do Twitter	13 usuários de leitores de tela
Bennett *et al.* [4] (2021)	Explorar os conceitos de raça, gênero e deficiência em descrições de imagens	25 usuários de leitores de tela

4.2.1 Backward Snowballing. Nesse procedimento analisaram-se as referências dos cinco estudos, cujo procedimento de busca retornou 301 potenciais trabalhos. A análise e a seleção dos potenciais trabalhos são expostas a seguir:

O estudo de Lazar *et al.* [20] possui 39 referências e rejeitaram-se 34 devido aos critérios de inclusão. A partir dos 5 estudos restantes, rejeitaram-se 4 devido aos títulos e mais 1 após a leitura do resumo por não se enquadrarem no escopo do *Snowballing*. Dessa forma, não se consideraram novos estudos.

O estudo de Voykinska *et al.* [34] possui 46 referências, das quais se rejeitaram 29 devido aos critérios de inclusão. A partir dos 17 estudos restantes, removeram-se 9 duplicados, sendo que 1 deles já pertencia ao conjunto de estudos selecionados. Em relação aos 8 trabalhos restantes, rejeitaram-se 5 deles devido ao título e mais 1 deles após a leitura do resumo devido ao escopo dos estudos. Rejeitaram-se os outros 2 trabalhos restantes após a leitura das introduções, dos resultados e das considerações finais por não se enquadrarem no escopo do *Snowballing*.

O estudo de Zhao *et al.* [39] possui 63 referências, rejeitaram-se 50 devido aos critérios de inclusão. A partir dos 13 estudos restantes, removeram-se 10 duplicados, dos quais 4 eram trabalhos já aceitos anteriormente. Rejeitaram-se os demais 3 estudos devido aos títulos fora do escopo.

O estudo de Gleason *et al.* [14] possui 27 referências e rejeitaram-se 14 devido aos critérios de inclusão. Os 13 estudos restantes eram estudos duplicados, dos quais 4 eram trabalhos já aceitos em alguma etapa anterior. Dessa forma, não restaram estudos a serem considerados.

O estudo de Bennett *et al.* [4] possui 126 referências e rejeitaram-se 89 devido aos critérios de inclusão. A partir dos 37 estudos restantes, removeram-se 18 duplicados, dos quais 7 eram estudos já selecionados anteriormente. Dos 19 estudos restantes, rejeitaram-se 18 deles após a leitura dos títulos e 1 após a leitura do resumo devido ao escopo. É relevante destacar que este estudo explorou especificamente descrições de raças e de gêneros com pessoas LGBTQIA+; dessa forma, desconsideram-se muitas das referências devido à alusão nessa temática e áreas afins como, por exemplo, diversidade.

4.2.2 Forward Snowballing. Nesse procedimento analisaram-se as citações dos 5 estudos, sendo que o procedimento de busca retornou 441 potenciais trabalhos. A análise foi conduzida utilizando o Google Acadêmico e a coleta dos dados ocorreu em 10 de fevereiro de 2021.

O estudo de Lazar *et al.* [20] possui 321 citações e rejeitaram-se 298 devido aos critérios de inclusão. A partir dos 23 estudos restantes, removeram-se 3 duplicados e já aceitos anteriomente. Dos demais 20 estudos, rejeitaram-se 17 devido aos títulos e outros 3 devido aos resumos fora do escopo do *Snowballing*.

O estudo de Voykinska *et al.* [34] possui 91 citações e rejeitaram-se 66 devido aos critérios de inclusão. A partir dos 25 estudos restantes, rejeitaram-se 17 estudos duplicados, dos quais 6 já foram selecionados em etapas anteriores. Referente aos 8 trabalhos restantes, rejeitaram-se 4 deles devido aos títulos e outros 2 deles devido aos resumos fora do escopo. Rejeitaram-se os demais 2 estudos após a leitura das introduções, dos resultados e das considerações finais por não se esquadrarem no escopo do *Snowballing*.

O estudo de Zhao *et al.* [39] possui 21 citações e rejeitaram-se 10 devido aos critérios de inclusão. A partir dos 11 estudos restantes, removeram-se 10 duplicados, sendo que 3 deles eram trabalhos anteriormente selecionados. Rejeitou-se o único estudo restante devido ao título fora do escopo.

O estudo de Gleason *et al.* [14] possui 8 citações e rejeitaram-se 3 devido aos critérios de inclusão. A partir dos 5 estudos restantes, removeu-se 1 estudo duplicado. Referente aos 4 trabalhos restantes, rejeitaram-se 2 deles devido aos títulos e mais 1 devido ao resumo fora do escopo. Rejeitou-se o único trabalho restante após a leitura da introdução, dos resultados e das considerações finais por não se esquadrar no escopo do *Snowballing*.

O estudo de Bennett *et al.* [4] não possuía citações até a data de coleta dos dados. O estudo foi apresentado na conferência CHI em maio de 2021 e um dos autores do trabalho disponibilizou-o em sua página pessoal da Carnegie Mellon University, cujo link direto foi apontado pelo Google Acadêmico.

4.3 Análise quantitativa das Iterações

Na Primeira Iteração, os procedimentos de busca retornaram 486 potenciais estudos. Destes, rejeitaram-se 325 após a aplicação dos critérios de inclusão e outros 86 após a remoção dos duplicados, restando 75 trabalhos. Em seguida, rejeitaram-se 52 estudos após a leitura dos títulos, outros 7 após a leitura dos resumos e demais 8 após a análise das introduções, dos resultados e das considerações finais por não pertencerem ao escopo do *Snowballing*. Por fim, rejeitaram-se 3 trabalhos por não responderem à Questão de Pesquisa, resultando em 5 novos estudos selecionados.

Já na segunda Iteração, os procedimentos de busca retornaram 742 estudos iniciais. Destes, rejeitaram-se 593 após a aplicação dos critérios de inclusão e outros 81 após a remoção dos duplicados, restando 68 estudos. Em seguida, rejeitaram-se 54 estudos após a leitura dos títulos, outros 9 após a leitura dos resumos e demais 5 após a análise das introduções, dos resultados e das considerações finais por não pertencerem ao escopo do *Snowballing*. Sendo assim, não se selecionaram novos estudos.

A Figura 1 apresenta uma síntese das etapas da seleção e a quantidade de estudos rejeitados em cada etapa. Como resultado, selecionaram-se onze trabalhos, dos quais 6 pertencem ao conjunto inicial e os outros 5 foram selecionados a partir da execução do *Snowballing*.

4.4 Extração de dados

Analisaram-se o contexto da pesquisa, o número de participantes e a técnica de coleta de dados utilizada em cada um dos onze trabalhos selecionados. Nove estudos pertencem ao contexto das mídias sociais e as mais proeminentes são o Facebook [29, 32, 34, 38, 39] e o Twitter [13, 14, 22, 24, 32, 34]. Quatro estudos englobaram demais plataformas como, por exemplo, YouTube [29], WhatsApp [29], Instagram [29, 34], LinkedIn [34], Snapchat[34] e Indeed [32]. Um estudo explorou também descrições de imagens em aplicações para produtividade [32], *e.g.,* Microsoft Word, Microsoft PowerPoint e Google Docs. Quatro estudos exploraram o contexto online e investigaram descrições de imagens em sites [32, 32], frustrações ao utilizar leitores de tela [20] e questões de diversidade em descrições de imagens [4].

No que se refere às técnicas de coleta de dados utilizadas, notou-se uma variância conforme o número de participantes. Cinco estudos coletaram seus dados através de questionário online e o número de respostas variou entre 1 até 100 [22, 29, 34], entre 101 até 200 [24] e de 300 ou mais respostas [29, 38]. Dois estudos utilizaram questionário em forma de diário pessoal, sendo que 6 respostas foram obtidas para um estudo de uma semana [39] e 100 respostas para um estudo de nove meses [20]. Além disso, um estudo relatou a preferência de alguns participantes em responder o questionário online por telefone [29]. Dois estudos coletaram dados também por meio da inspeção manual, um visando analisar as interfaces das aplicações selecionadas [29] e outro coletar dados públicos disponíveis no Twitter [24]. Oito estudos utilizaram entrevistas e o número de participantes variou entre 1 até 10 [22, 38, 39], de 11 até 20 [13, 14, 34, 39] e de 21 ou mais participantes [4, 32].

5 RESULTADOS

Essa Seção discorre sobre os resultados do *Snowballing*. A Subseção 5.1 apresenta e discute a resposta da Questão de Pesquisa, a Subseção 5.2 apresenta fatores que inibem a geração de descrições manuais elaboradas por pessoas videntes e, por fim, a Subseção 5.3 expõe recomendações identificadas que podem amenizar os problemas em descrições de imagens encontrados nesse levantamento.

5.1 Problemas em descrições de imagens

A execução do *Snowballing* resultou na identificação de 13 problemas em descrições de imagens discutidos a seguir.

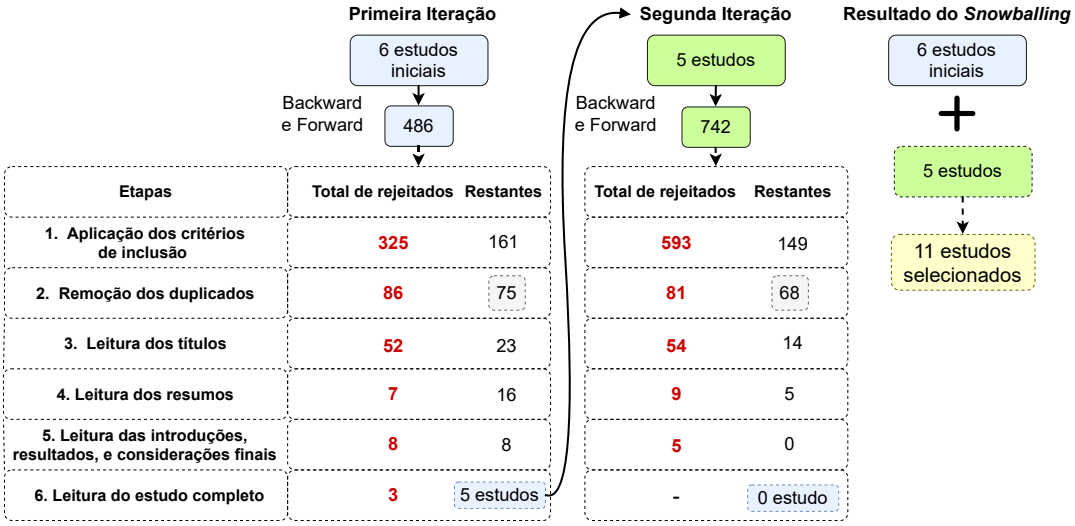

Figura 1: Síntese das etapas de seleção e suas respectivas quantidades de estudos rejeitados e restantes, em cada Iteração.

(1) **Textos insuficientes:** esse problema refere-se a descrições muito genéricas que, apesar de existentes, não fornecem informações ou detalhes suficientes para contextualizar a cena [14]. Descrições ricas em detalhes são necessárias para que as pessoas com deficiência visual consigam ter a mesma, ou semelhante, experiência das pessoas videntes ao consumirem os conteúdos visuais [32]. Conforme Morris *et al.* [24] investigaram, até junho de 2015, as imagens representavam 93,6% do conteúdo do Twitter e 61,8% dos textos associados a esses conteúdos seriam descrições insuficientes, significando que uma pessoa cega ouvindo o texto não teria nenhuma percepção do que a imagem representava. Ainda segundo os autores, 55,6% das imagens foram consideradas essenciais, *i.e.,* uma pessoa ouvindo o texto sem visualizar a imagem não entenderia o propósito do tuíte [24].

Segundo Voykinska *et al.* [34], a falta de descrições contextuais úteis nas mídias sociais prejudica o envolvimento das pessoas com deficiência visual com os conteúdos. Stangl, Morris e Gurari [32] complementaram que descrições insuficientes ou até mesmo ausentes inibem a interação das pessoas com deficiência visual nas mídias sociais, causando frustração e confusão. Ainda segundo os autores, somente a identificação dos objetos de uma imagem é insuficiente para o entendimento da cena, resultando em descrições incompletas e interpretações incoerentes do conteúdo visual [32].

Conforme Wu *et al.* [38] observaram, as descrições automáticas do Facebook são insuficientes para que as pessoas com deficiência visual se sintam encorajadas a interagir por comentários ou *likes*. A necessidade de melhoria dos sistemas automáticos também evidencia-se no contexto brasileiro, conforme Sacramento *et al.* [29] observaram com os 100 participantes com deficiência visual do estudo. Os autores identificaram que a segunda maior dificuldade enfrentada pelos participantes no acesso ao conteúdo visual refere-se a descrições muito genéricas e que elas "atendem apenas em alguns casos as expectativas" de descrições.

(2) **Textos inexistentes:** conteúdos visuais sem uma descrição associada são inacessíveis às pessoas com deficiência visual, prejudicando a inclusão digital e a participação ativa na Internet [34]. Conforme relatado no estudo de Wu *et al.* [38], a escassez de descrições de imagens causa sentimentos de isolamento e frustração nas pessoas com deficiência visual, visto que elas se sentem impedidas de interagir nas mídias sociais por não saberem o que os conteúdos visuais estão representando.

Desde 2016, o Twitter permite que os usuários incluam descrições de imagens manuais e, em 2018, Gleason *et al.* [13] investigaram a prevalência dessas descrições. Os autores coletaram mais de 9 milhões de tuítes públicos, dos quais mais de 1 milhão (11,84%) continham pelo menos uma imagem e apenas 0,1% destes continham textos alternativos. Sacramento *et al.* [29] identificaram que nenhuma das quatro mídias sociais mais utilizadas no Brasil (WhatsApp, YouTube, Facebook e Instagram) dispunham de descrições automáticas em suas versões mobile, sendo que apenas o Instagram habilitou a inserção manual de descrições através do aplicativo. Além disso, a ausência de descrições dos conteúdos visuais foi a maior dificuldade relatada pelos participantes nas quatro mídias exploradas no estudo [29].

Lazar *et al.* [20] discutiram que descrições de imagens inapropriadas podem impedir, por exemplo, o acesso das pessoas com deficiência visual em websites que possuem verificação de segurança via digitação de um texto presente na imagem. Segundo os autores, a ausência de descrições de imagens foi a quarta maior causa de frustração relatada pelos 100 participantes do estudo. Conforme Voykinska *et al.* [34], esse problema também resulta em um sentimento de isolamento, visto que as pessoas com deficiência visual se sentem excluídas ou incapazes de participar nas mídias sociais. Os autores relataram, ainda, que entre as estratégias adotadas pelas pessoas com deficiência visual estão a procura por informações adicionais nos comentários da postagem ou a solicitação de

ajuda para pessoas próximas [34]. Gleason *et al.* [14] também observaram que a falta de conteúdos acessíveis é a maior barreira para a participação das pessoas com alguma deficiência.

(3) **Textos errôneos:** geradores de descrições automáticas ainda são imperfeitos e podem gerar textos divergentes das imagens [22]. Segundo Gleason *et al.* [14], descrições imprecisas podem induzir as pessoas com deficiência visual a crerem que uma imagem contém algo que não está realmente presente. Macleod *et al.* [22] observaram que as pessoas com deficiência visual tendem a confiar em descrições automáticas mesmo quando elas apresentaram incongruências. Além disso, os autores observaram que a confiança dos participantes em descrições de imagens está associada às suas expectativas de descrições e, mesmo nos casos em que elas não eram atendidas, os participantes tentaram justificar ou criar cenários para conectar a descrição com a imagem inesperada ao invés de suspeitar que o texto descritivo estivesse incorreto [22].

Wu *et al.* [38] observaram que os participantes do estudo em um ambiente controlado manifestaram a preferência por mais objetos identificados mesmo que isso significasse descrições menos acuradas, entretanto, esse fato não foi observado no estudo em larga escala, onde os participantes demonstraram mais intolerância para os casos incorretos, gerando desconfiança no algoritmo e nas descrições geradas. Além disso, Sacramento *et al.* [29] revelaram que a presença de descrições incorretas é a terceira maior dificuldade enfrentada nas quatro mídias sociais mais utilizadas no Brasil.

(4) **Textos que não descrevem as informações textuais contidas na imagem:** esse problema ocorre quando há um texto na imagem não contido na descrição. A inclusão dos textos embutidos ajuda no entendimento do contexto e do propósito da imagem como, por exemplo, uma notícia contendo uma foto de uma pessoa segurando um cartaz em uma manifestação pública [24]. O estudo de Lazar *et al.* [20] aborda como o uso de imagens com textos para a autenticação de websites pode causar frustração nas pessoas com deficiência visual. Ainda, mesmo com uma versão alternativa de áudio, essa solução não é ideal devido às diversas pronúncias de uma mesma palavra.

Segundo Gleason *et al.* [14] observaram, as descrições automáticas do Facebook informam o usuário da presença de textos na imagem, *e.g.,* "Essa imagem pode conter texto." e, para contornar as limitações, as pessoas com deficiência visual utilizam estratégias como, por exemplo, o uso de aplicativos que realizam Reconhecimento Ótico de Caracteres (OCR, do inglês *Caracter Optical Recognition*). De acordo com Morris *et al.* [24], há diversas categorias de imagens com informações textuais, entre elas *screenshots*, frases motivacionais, gráficos, memes e propagandas, amplamente utilizadas para transmitir conteúdo.

Essa variância de categorias e as limitações atuais sugerem que os geradores automáticos necessitam empregar diferentes abordagens para a construção de sentenças como, por exemplo, visão computacional para reconhecimento dos elementos e de OCR para reconhecimento dos textos embutidos nas imagens [14, 24].

(5) **Textos confusos:** esse problema refere-se a descrições com textos não coerentes e claros. Segundo Wu *et al.* [38], na ocorrência de descrições automáticas confusas, as pessoas com deficiência visual tentam adivinhar por conta própria a intenção do algoritmo, buscando entender o que há na imagem para que a descrição fosse gerada daquela forma. Entretanto, de acordo com Sacramento *et al.* [29], esse problema também está sujeito a ocorrer nas descrições manuais e, até mesmo, por profissionais de audiodescrição, ao usarem palavras regionalistas, o que causa descontamento e confusão naqueles que desconhecem tais expressões.

(6) **Textos sem abrangência ao humor e às emoções:** Voykinska *et al.* [34] abordaram as limitações das tecnologias atuais para expressarem o humor contido em imagens como, por exemplo, sarcasmo e conteúdos de comédia, assim como as emoções das pessoas, *e.g.,* raiva e tristeza. Wu *et al.* [38] destacaram que "pessoas" presentes em uma imagem são consideradas o elemento mais interessante, sendo o humor a característica mais importante, seguida pela ação delas. Todavia, os sistemas de descrições automáticas ainda são limitados em descrever humor e emoções, dado que é extremamente difícil treinar algoritmos para interpretar esses conceitos nos contextos de uso [38].

(7) **Textos que não descrevem apropriadamente as características das pessoas:** as descrições de imagens ainda são limitadas em descrever raça, gênero, ou se a pessoa possui alguma deficiência, o que evidencia a necessidade de abranger tais questões contemporâneas nos geradores automáticos [32]. Segundo Stangl, Morris e Gurari [32], descrever uma pessoa fisicamente é uma tarefa subjetiva e nem sempre as características estão implícitas na imagem, *e.g.,* raça ou etnia, o que pode acarretar descrições incorretas. Além disso, os autores observaram que informações sobre raça e gênero são preferidas em diferentes contextos digitais, *e.g.,* redes sociais e portais de notícias, enquanto características mais específicas como estilo de cabelo, estatura corporal, peso e cor dos olhos, são preferidas apenas no contexto de sites de relacionamento.

Tais informações são relevantes por uma questão de equidade, já que pessoas videntes as conseguem perceber pelo senso de visão [32]. O estudo de Bennett *et al.* [4] aborda a emergência de adequar o treinamento dos geradores automáticos para considerar também pessoas representantes da diversidade de gênero, visto que as descrições automáticas utilizam apenas dois gêneros (masculino e feminino) para descrever uma pessoa.

(8) **Textos que não descrevem apropriadamente as expressões faciais e corporais das pessoas:** essa limitação é semelhante ao da emoção, porém descrever expressões faciais refere-se aos traços do rosto, *e.g.,* uma pessoa com uma expressão séria, ou mostrando a língua, ou com um olhar malicioso [32]. Sacramento *et al.* [29] relataram que alguns dos 100 participantes com deficiência visual demandaram mais detalhes das descrições automáticas como, por exemplo, expressões faciais e corporais.

Segundo Stangl, Morris e Gurari [32], as expressões faciais e linguagem corporal são muito relevantes nos contextos das mídias sociais e em sites de relacionamento, já que tais expressões ajudam a pessoa com deficiência visual a decidir como reagir à imagem. Além disso, descrever essas expressões auxilia as pessoas com deficiência visual a compreender o objetivo da imagem [32], *e.g.,* uma pessoa mostrando os músculos pode significar tanto um contexto de academia, quanto aludir às campanhas de empoderamento feminino como, por exemplo, #WeCanDoIt.

(9) **Textos não descritivos:** esse problema refere-se quando uma descrição contém apenas a palavra "imagem" ao invés de descrever o seu conteúdo visual. O estudo realizado por Lazar *et al.* [20]

expôs que a presença de "textos não descritivos" é a segunda maior causa de frustração nos usuários de leitores de tela na categoria de "textos alternativos".

(10) **Textos homogêneos para imagens semelhantes:** conforme Zhao *et al.* [39], os geradores de descrições automáticas são limitados em fornecer descrições mais heterogêneas para imagens semelhantes. Isso pode dificultar o processo de imaginação da pessoa com deficiência visual, sua compreensão sobre o está sendo representado na imagem, assim como obstar a identificação de imagens variadas de um mesmo evento, *e.g.,* um álbum no celular contendo fotos de um almoço em família. Os autores observaram também que para as pessoas com baixa visão, os detalhes ou particularidades das imagens, eram muito mais relevantes do que para as pessoas cegas, já que elas esperam que as descrições transmitam mais informações do que o resquício visual as permitem perceber.

(11) **Textos errôneos para imagens de baixa qualidade:** descrições automáticas são geradas independente da qualidade da imagem, conforme os objetos identificados. Entretanto, muitas das imagens digitais podem apresentar borrões e desfoques, principalmente as capturadas pelas pessoas com deficiência visual. Segundo Zhao *et al.* [39], uma limitação dos geradores automáticos é não informar sobre as condições das imagens mesmo nos casos insuficientes para uma identificação adequada dos objetos. Os autores debateram que essa informação é útil para que as pessoas possam submeter imagens com qualidade superior, evitando a geração de descrições incorretas ou insuficientes [39].

(12) **Textos que não transmitem a intenção da imagem:** segundo Stangl, Morris e Gurari [32], pessoas com deficiência visual desejam saber qual é o propósito da imagem para entender o que a torna importante. Os autores observaram que a intenção da imagem é muito relevante no contexto das redes sociais, principalmente quando os comentários da imagem não referenciam seu conteúdo, para que as pessoas com deficiência visual entendam a representação visual e o porquê ela foi escolhida para ser compartilhada. Os autores discutiram, ainda, que as preferências de descrições variam conforme o elemento central da imagem, *e.g.,* em uma imagem de alguém exibindo o anel de noivado, descrever apenas "um anel" não ajudaria uma pessoa com deficiência visual a entender a intenção da imagem.

(13) **Textos que não descrevem apropriadamente as ações das pessoas:** segundo Wu *et al.* [38], descrever o que as pessoas estão realizando na imagem, *i.e.,* a ação, é considerada a segunda característica mais interessante pelas pessoas com deficiência visual, sendo essa uma das sugestões de melhoria de 26% dos participantes do estudo em larga escala. Os autores abordaram, ainda, que descrever em detalhes as ações das pessoas na imagem possibilita uma contextualização do que está sendo representado.

A Tabela 3 sumariza a frequência dos problemas discutidos nessa Seção. Observa-se que os três problemas mais frequentes referem-se a: "Textos insuficientes" e "Textos inexistentes", citados por seis dos onze trabalhos selecionados e "Textos errôneos", citados por quatro dos onze trabalhos. É relevante destacar que pode haver intersecção entre os trabalhos, visto que eles relataram mais de um problema em descrição de imagens.

Tabela 3: Frequência dos problemas identificados em descrições de imagens.

Problemas identificados	Frequência
Textos insuficientes	6
Textos inexistentes	6
Textos errôneos	4
Textos que não descrevem as informações textuais contidas na imagem	3
Textos confusos	2
Textos sem abrangência ao humor e às emoções	2
Textos que não descrevem apropriadamente as características das pessoas	2
Textos que não descrevem apropriadamente as expressões faciais e corporais das pessoas	2
Textos não descritivos	1
Textos homogêneos para imagens semelhantes	1
Textos errôneos para imagens de baixa qualidade	1
Textos que não transmitem a intenção da imagem	1
Textos que não descrevem apropriadamente as ações das pessoas	1

5.2 Fatores que inibem a geração de descrições manuais

Além dos problemas em descrições de imagens, o presente estudo identificou, também, fatores contribuintes para a inibição de descrições manuais elaboradas por pessoas videntes. São eles:

(1) **Desconhecer como descrever uma imagem:** os autores de descrições manuais são recomendados a seguir as orientações fornecidas pelas Diretrizes de Acessibilidade para o Conteúdo da Web (WCAG, do inglês *Web Content Accessibility Guidelines*) [13]. Todavia, as pessoas videntes possuem dificuldade para descrever os conteúdos visuais em palavras, por desconhecerem como elaborar as alternativas e por julgarem complexa a tarefa de descrever uma imagem [13, 29].

(2) **Esquecimento:** pessoas videntes alegam esquecer de incluir descrições de imagens [13, 29]. Segundo Gleason *et al.* [13], os motivos incluem pressa, múltiplas postagens em um curto tempo, até mesmo esquecimento da existência e do objetivo dessa funcionalidade. No estudo de Sacramento *et al.* [29], "esquecimento" foi o terceiro motivo mais frequente informado pelos participantes videntes, mesmo por aqueles que relataram criar descrições para os conteúdos nas mídias sociais.

(3) **Falta de tempo:** apesar da importância, as pessoas videntes alegam não ter tempo de descrever uma imagem em palavras, visto que essa tarefa requer tempo e energia [13, 29].

(4) **Problemas na interface:** os recursos para inserção de descrições de imagens nas mídias sociais requerem configuração manual e não são amplamente divulgados para seus usuários [13, 29]. Dessa forma, são necessários diversos passos para habilitar a funcionalidade, sendo que a referida não é fácil de ser encontrada nas configurações das contas [13, 29].

No estudo de Sacramento *et al.* [29], quase 50% dos participantes videntes da pesquisa afirmaram não adicionar descrições de imagens

por desconhecerem como adicioná-las nas postagens. Além disso, as interfaces não permitem adicionar descrições em imagens em postagens não originais [13, 29]. Segundo Sacramento *et al.* [29], as interfaces carecem de instruções, dado que apenas as referidas do Facebook e Instagram contêm orientações sobre como adicionar uma descrição de imagem;

(5) **Desconhecer a importância de descrições de imagens:** pessoas videntes desconhecem a necessidade de descrições de imagens e o quanto elas são essenciais para garantir o acesso ao conteúdo e a inclusão digital das pessoas com deficiência visual, mesmo os amigos ou familiares destes [29]. O estudo de Sacramento [29] identificou, ainda, a falta de interesse das pessoas videntes em elaborar conteúdos para pessoas com deficiência visual, evidenciando a urgência de ações de conscientização.

A Tabela 4 sumariza os motivos pelos quais as pessoas videntes não fornecem suas próprias descrições para os conteúdos visuais.

Tabela 4: Fatores que inibem a geração de descrições manuais por pessoas videntes.

Fatores identificados	Frequência
Desconhecer como descrever uma imagem	2
Esquecimento	2
Falta de tempo	2
Problemas na interface	2
Desconhecer a importância das descrições de imagens	1

5.3 Sugestões de melhorias

Outro resultado do *Snowballing* foi a identificação de recomendações para alguns dos problemas encontrados, cuja sumarização encontra-se na Tabela 5.

Visando melhorar a qualidade de descrições manuais (humanas) e diminuir a defasagem destas, as mídias sociais necessitam **fornecer instruções sobre como descrever uma imagem**, explicando o propósito dessa ação e o porquê as pessoas deveriam adotar esse hábito [13, 29]. Outras estratégias incluem **utilizar perguntas estruturadas** para a geração da descrição [13], **criar recursos para lembrar** as pessoas de descrever os conteúdos visuais [29] e **incentivar as pessoas com deficiência visual** a compartilharem instruções sobre como escrever descrições que satisfaçam suas necessidades [14].

Foi exposto que as interfaces das mídias sociais deveriam **facilitar o acesso às configurações específicas** do recurso de descrição de imagem [13, 29]. Gleason *et al.* [13] sugeriram a **habilitação automática** desta funcionalidade para os usuários. Entretanto, essa ação poderia resultar no uso incorreto como, por exemplo, a inserção de links externos ou spam ao invés de usar o recurso para descrever os conteúdos visuais [13]. Outra recomendação refere-se a **aumentar o limite do número de caracteres** no campo da descrição para a inclusão de textos mais significativos, especialmente para as imagens com textos imbutidos [13]. Além disso, **solicitar a confirmação do usuário** antes de postar um conteúdo visual sem descrição também é uma recomendação para a defasagem de textos alternativos [29].

Gleason *et al.* [13] e Sacramento *et al.* [29] sugeriram que as interfaces permitam a **inserção de descrições de imagens após o ato de postagem** do conteúdo e em conteúdos originalmente criados por terceiros. Todavia, Sacramento *et al.* [29] expõem a preocupação com direito autoral e aconselham que o autor do conteúdo seja notificado sobre as descrições inseridas para este poder aprovar ou complementar as descrições, ou, ainda, que elas sejam geradas colaborativamente. Os autores destacam que essa ação poderia contribuir também para a conscientização das pessoas videntes sobre os conteúdos visuais acessíveis [29].

Referente à disseminação de descrições de imagens, identificaram-se recomendações como **realizar ações para sensibilizar as pessoas videntes** [29], **tornar o campo de descrição mais saliente** para que mais pessoas conheçam a funcionalidade [13, 34] e a **ampliação da funcionalidade de marcação nas imagens** para permitir a inclusão de comentários manuais, ao invés de apenas identificar os elementos presentes [34]. De acordo com Lazar *et al.* [20], alguns dos problemas mais frequentes em descrições de imagens como, por exemplo, a ausência e textos não descritivos, poderiam ser mitigados caso fossem seguidas as orientações das diretrizes de acessibilidade. Os autores sugeriram ações como **capacitação** para as equipes de desenvolvimento, **políticas governamentais** mais robustas e o **uso de ferramentas** que incorporam mais fortemente Acessibilidade.

Recomendações para descrições automáticas nas mídias sociais referem-se à **garantia do funcionamento do recurso em diferentes versões**, *e.g.,* desktop e aplicativo *mobile* [29] e à **ampliação do escopo das descrições** para englobarem também outros conteúdos visuais, *e.g.,* emojis, figurinhas e GIFs [13, 29, 34]. No caso das imagens com informações textuais, *e.g., screenshots*, a **adoção de múltiplas abordagens** pode tanto identificar os objetos, quanto reconhecer caracteres de texto, resultando em descrições mais completas para as pessoas com deficiência visual [14, 24].

6 CONSIDERAÇÕES FINAIS

Este trabalho propôs-se a responder, através de uma revisão da literatura utilizando a técnica de *Snowballing*, a seguinte pergunta: **"Quais são os problemas em descrição de imagens para pessoas com deficiência visual?"**. Ao todo, selecionaram-se onze estudos para responder à pergunta. Após a extração e análise dos estudos selecionados, identificaram-se treze problemas em descrições de imagens nas perspectivas das pessoas com deficiência visual. Observou-se que, apesar de ser uma recomendação do nível mais básico de acessibilidade, as pessoas com deficiência visual ainda se deparam com imagem sem textos descritivos, cujo problema foi reportado em seis estudos.

Identificou-se uma demanda por melhorias nos geradores automáticos para geração de descrições mais ricas, detalhando especialmente o gênero, as expressões faciais e as ações das pessoas na imagem. Um outra demanda refere-se ao aperfeiçoamento para estes atenderem às expectativas de descrição e que descrevam conteúdos visuais diversos, como GIFs e imagens com textos embutidos. Observou-se, também, que a ausência de descrições inibe a participação das pessoas com deficiência visual nas plataformas de mídias sociais e causam sentimentos de frustração e isolamento, expondo a

Tabela 5: Frequência das recomendações identificadas para problemas em descrições de imagens.

Recomendações identificadas	Frequência
Ampliar o escopo dos modelos de descrições automáticas	3
Fornecer instruções sobre como descrever uma imagem	2
Facilitar o acesso as configurações específicas do recurso de descrição automática	2
Permitir a inserção de descrições de imagens após o ato de postagem	2
Permitir a inserção de descrições de imagens em conteúdos compartilhados por terceiros	2
Tornar o campo de descrição mais saliente	2
Adotar múltiplas abordagens nos modelos de descrições automáticas	2
Habilitar automaticamente as configurações específicas do recurso de descrição automática	1
Utilizar perguntas estruturadas para auxiliar a geração de descrição manual	1
Incluir recursos para lembrar as pessoas de descreverem os conteúdos visuais	1
Incentivar as pessoas com deficiência visual a compartilharem suas necessidades de descrições	1
Aumentar o limite do número de caracteres	1
Solicitar a confirmação ao postar um conteúdo visual sem descrição	1
Realizar ações para conscientizar as pessoas videntes	1
Ampliar a funcionalidade de marcação nas imagens	1
Capacitar as equipes de desenvolvimento	1
Criar políticas governamentais mais fortes	1
Utilizar ferramentas que incorporam Acessibilidade	1
Garantir o funcionamento do recurso de descrição em múltiplas plataformas	1

necessidade de políticas públicas que reforçam a pertinência de conteúdos visuais acessíveis. Além disso, é fundamental conscientizar pessoas videntes, visto que elas alegam desconhecer a importância das descrições de imagens, além de qualificá-las através de treinamentos e fornecer instruções sobre como descrever conteúdos visuais para pessoas com deficiência.

Identificou-se que a descrição de características pessoais como, por exemplo, raça, gênero, etnia e deficiência, é uma questão de equidade para as pessoas que não podem utilizar o sentido da visão [32]. No entanto, os modelos de descrições automáticas ainda falham em descrever as referidas características, visto que é extremamente difícil treiná-los para interpretar os conceitos de identidade, aparência e classe social nos contextos [38]. Uma das razões para essa dificuldade é que os conjuntos de dados utilizados nos treinamentos dos modelos automáticos, *e.g.,* MSCOCO [7] e Flickr30k [27], omitem essas informações na tentativa de permanecerem objetivos ou evitarem preconceitos [4]. Se por um lado a omissão dessas características pode excluir pessoas já demonstradas em desvantagem pela IA e pela sociedade em geral [4], por outro, descrever esses conceitos inadequadamente ou incorretamente pode causar danos e constrangimentos [32].

Observou-se nos estudos selecionados, o entusiasmo das pessoas com deficiência visual para com o potencial dos modelos automáticos de contribuir para a acessibilidade dos conteúdos visuais, porém há, ainda, uma expressiva hostilidade no que se refere à precisão, à homogeneidade e à ética das descrições automáticas. A partir dessas lições, anseia-se contribuir para a comunidade científica da IHC e das demais áreas de pesquisas, ao identificar algumas das limitações atuais de descrições de imagens produzidas manualmente e por modelos automáticos, nas perspectivas das pessoas com deficiência visual. Espera-se que a discussão apresentada chame a atenção da comunidade e inspire futuras pesquisas que visam minimizar as limitações apontadas neste levantamento.

Em relação às limitações deste trabalho, apesar de não se ter buscado especificamente pelo contexto da mídia social [1], nove dos onze estudos selecionados referiram-se a esse contexto e podem não englobar todos os problemas em descrições de imagens. Além disso, nosso levantamento não identificou se os problemas em descrições estão atrelados ao grau de deficiência visual, visto que muitos dos estudos selecionados utilizaram o termo "pessoas com deficiência visual" ou "usuários de leitores de tela". Outra limitação refere-se à dificuldade em recrutar voluntários com deficiência visual, uma mesma pessoa pode ter participado de mais de um estudo dos autores recorrentes, por conhecer o pesquisador e estar habituado a participar de suas pesquisas, portanto, pode afetar a variedade das respostas reportadas nos estudos selecionados. Ademais, estudos relevantes para o *Snowballing* podem não ter sido inclusos na análise devido à limitação do escopo dos veículos de publicações considerados.

Perspectivas de trabalhos futuros incluem: investigar se os problemas em descrições estão relacionados com o grau de deficiência visual, *i.e.,* se ocorrem exclusivamente para pessoas cegas ou para pessoas com baixa visão; analisar as características das descrições buscando identificar quais problemas estão presentes apenas em sentenças automáticas e apenas em sentenças humanas e; propor melhorias para os modelos automáticos, visando sanar, por exemplo, a limitação de descreverem apenas gêneros binários [4], para que os modelos sejam mais inclusivos e abrangentes.

AGRADECIMENTOS

Pesquisa financiada pela HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda. com recursos provenientes da contrapartida da isenção ou redução de IPI conforme a Lei nº 8.248, de 1991.

REFERÊNCIAS

[1] Maria Lúcia Toledo Moraes Amiralian. 1997. *Compreendendo O Cego - Uma Visão Psicanalítica Da Cegueira Por Meio De Desenhos-Estórias* (1ª ed.). Casa do Psicólogo, São Paulo.

[2] Soheyla Amirian, Khaled Rasheed, Thiab R Taha, and Hamid R Arabnia. 2019. A Short Review on Image Caption Generation with Deep Learning. In *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition. World Congress in Computer Science, Computer Engineering, and Applied Computing (IPCV'19).* CSREA Press, 10–18.

[3] Shuang Bai and Shan An. 2018. A survey on automatic image caption generation. *Neurocomputing* 311 (May 2018), 291–304. https://doi.org/10.1016/j.neucom.2018.05.080

[4] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI*

[1] O termo mídia social foi utilizado por autores dos estudos selecionados para se referirem às aplicações de relacionamentos, empregos e demais redes sociais, portanto, decidiu-se por manter tal nomenclatura.

'21). Association for Computing Machinery, New York, NY, USA, Article 375, 19 pages. https://doi.org/10.1145/3411764.3445498

[5] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research* 55, 1 (Jan 2016), 409–442. https://doi.org/10.1613/jair.4900

[6] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 9959–9968. https://doi.org/10.1109/CVPR42600.2020.00998

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO Captions: Data Collection and Evaluation Server. 8693 (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[8] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Workshop on Statistical Machine Translation (WMT '14)*. Association for Computational Linguistics, Baltimore, Maryland, USA, 376–380. https://doi.org/10.3115/v1/W14-3348

[9] Diego Dermeval, Jorge A. P. de M. Coelho, and Ig I. Bittencourt. 2020. Mapeamento Sistemático e Revisão Sistemática da Literatura em Informática na Educação. In *Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa*, Patricia Jaques, Mariano Pimentel, Sean Siqueira, and Ig Bitencourt (Eds.). SBC. Série Metodologia de Pesquisa em Informática na Educação, Porto Alegre, Chapter 3. https://metodologia.ceie-br.org/livro-2

[10] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, and Tom Sercu. 2019. Adversarial Semantic Alignment for Improved Image Captions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 10455–10463. https://doi.org/10.1109/CVPR.2019.01071

[11] e MAG. 2014. Modelo de Acessibilidade em Governo Eletrônico (eMAG 3.1). http://emag.governoeletronico.gov.br/

[12] Desmond Elliott and Frank Keller. 2013. Image Description using Visual Dependency Representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1292–1302. https://www.aclweb.org/anthology/D13-1128/

[13] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's Almost like They're Trying to Hide It": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 549–559. https://doi.org/10.1145/3308558.3313605

[14] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376728

[15] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research* 47, 1 (May 2013), 853–899. https://doi.org/10.1613/jair.3994

[16] Laura Hollink, A. Th. Schreiber, Bob J. Wielinga, and Marcel Worring. 2004. Classification of user image descriptions. *International Journal of Human-Computer Studies* 61, 5 (Nov 2004), 601–626. https://doi.org/10.1016/j.ijhcs.2004.03.002

[17] Ingrid Hrga and Marina Ivasic-Kos. 2019. Deep Image Captioning: An Overview. In *Proceedings of the International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO '19)*. IEEE, 995–1000. https://doi.org/10.23919/MIPRO.2019.8756821

[18] Marina Ivasic-Kos, Ivo Ipsic, and Slobodan Ribaric. 2015. A knowledge-based multi-layered image annotation system. *Expert Systems with Applications* 42, 24 (2015), 9539–9553. https://doi.org/10.1016/j.eswa.2015.07.068

[19] Alejandro Jaimes and Shih-Fu Chang. 2000. A Conceptual Framework for Indexing Visual Information at Multiple Levels. *Electronic Imaging* 3964 (Jan 2000), 2–15. https://doi.org/10.1117/12.373443

[20] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. *International Journal of Human–Computer Interaction* 22, 3 (2007), 247–269. https://doi.org/10.1080/10447310709336964

[21] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Workshop on Text Summarization Branches Out (WAS 2004)*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1000

[22] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5988–5999. https://doi.org/10.1145/3025453.3025814

[23] João Marcelo Santos Marques, Simone Bacellar Leal Ferreira, and Claudia Cappelli. 2020. Identificando as principais dificuldades na compreensão de gráficos pelos cidadãos cegos. *Brazilian Journal of Development* 6, 11 (Nov 2020), 88683–88704. https://doi.org/ttps://doi.org/10.34117/bjdv6n11-332

[24] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "With Most of It Being Pictures Now, I Rarely Use It": Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5506–5516. https://doi.org/10.1145/2858036.2858116

[25] Emma Murphy, Ravi Kuber, Graham McAllister, Philip Strain, and Wai Yu. 2007. An empirical investigation into the difficulties experienced by visually impaired Internet users. *Universal Access in the Information Society* 7, 1 (Oct 2007), 79–91. https://doi.org/10.1007/s10209-007-0098-4

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL 2002)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[27] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision* 123, 1 (May 2017), 74–93. https://doi.org/10.1007/s11263-016-0965-7

[28] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Los Angeles, California) *(CSLDAMT '10)*. Association for Computational Linguistics, USA, 139–147. https://doi.org/10.5555/1866696.1866717

[29] Carolina Sacramento, Leonardo Nardi, Simone Bacellar Leal Ferreira, and João Marcelo dos Santos Marques. 2020. #PraCegoVer: Investigating the Description of Visual Content in Brazilian Online Social Media. In *Proceedings of the Brazilian Symposium on Human Factors in Computing Systems* (Diamantina, Brazil) *(IHC '20)*. Association for Computing Machinery, New York, NY, USA, Article 1, 10 pages. https://doi.org/10.1145/3424953.3426489

[30] Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. 2020. Image Captioning: A Comprehensive Survey. In *Proceedings of the International Conference on Power Electronics IoT Applications in Renewable Energy and its Control (PARC '20)*. IEEE, 325–328. https://doi.org/10.1109/PARC49193.2020.236619

[31] Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research* 104 (Aug 2019), 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

[32] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?"What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376404

[33] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 4566–4575.

[34] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How Blind People Interact with Visual Content on Social Networking Services. In *In Proceedings of the Conference on Computer-Supported Cooperative Work Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1584–1595. https://doi.org/10.1145/2818048.2820013

[35] WCAG. 2008. Web Content Accessibility Guidelines (WCAG 2.1). https://www.w3.org/TR/WCAG21/

[36] WebAIM. 2020. The WebAIM Million. https://webaim.org/projects/million/

[37] Claes Wohlin. 2014. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. In *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering* (London, England, United Kingdom) *(EASE '14)*. Association for Computing Machinery, New York, NY, USA, Article 38, 10 pages. https://doi.org/10.1145/2601248.2601268

[38] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-Text: Computer-Generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1180–1192. https://doi.org/10.1145/2998181.2998364

[39] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The Effect of Computer-Generated Descriptions on Photo-Sharing Experiences of People with Visual Impairments. *Human-Computer Interaction* 1, CSCW, Article 121 (Dec 2017), 22 pages. https://doi.org/10.1145/3134756