PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

ALESSANDRA HELENA JANDREY

IMAGERY CONTENTS DESCRIPTIONS FOR PEOPLE WITH VISUAL IMPAIRMENTS

Porto Alegre
2021

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

IMAGERY CONTENTS DESCRIPTIONS FOR PEOPLE WITH VISUAL IMPAIRMENTS

ALESSANDRA HELENA JANDREY

Master Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. Duncan Dubugras Alcoba Ruiz
Co-Advisor: Prof. Milene Selbach Silveira

Porto Alegre
2021

Ficha Catalográfica

J33i Jandrey, Alessandra Helena

Imagery contents descriptions for people with visual impairments / Alessandra Helena Jandrey. – 2021.
123 f.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Duncan Dubugras Alcoba Ruiz.
Co-orientadora: Profa. Dra. Milene Selbach Silveira.

1. Good practices. 2. Image descriptions. 3. Visually impaired People. 4. Qualitative study. 5. Snowballing. I. Ruiz, Duncan Dubugras Alcoba. II. Silveira, Milene Selbach. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

Alessandra Helena Jandrey

IMAGERY CONTENTS DESCRIPTIONS FOR PEOPLE WITH VISUAL IMPAIRMENTS

This Master Thesis/Doctoral Thesis has been submitted in partial fulfillment of the requirements for the degree of Doctor/Master of Computer Science, of the Graduate Program in Computer Science, School of Technology of the Pontifical Catholic University of Rio Grande do Sul.

Sanctioned on 25th August, 2021.

COMMITTEE MEMBERS:

Prof. Dr. Simone Bacellar Leal Ferreira (PPGI/UNIRIO)

Prof. Dr. Isabel Harb Manssour (PPGCC/PUCRS)

Prof. Dr. Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS - Advisor)

Prof. Dr. Milene Selbach Silveira (PPGCC/PUCRS – Co-Advisor)

“Ema daubuema
Daubuera daubuema
Dãnbabãndabãnabauema
Dauba daubaemaparências
Raruri rararerê enganar meu coração.”
(Alcione - Evidências)

“Mom says, “*Where did anxiety come from?*”
Anxiety is the cousin visiting from out of town that depression felt obligated to invite to the party
Mom, I am the party, only I am a party I don’t want to be at. (...)
Mom says, “*Try counting sheep*”
But my mind can only count reasons to stay awake
So I go for walks, but my stuttering kneecaps clank like silver spoons held in strong arms with loose wrists
They ring in my ears like clumsy church bells, reminding me I am sleepwalking on an ocean of happiness that I cannot baptize myself in (...)”
(Sabrina Benaim — *Explaining my depression to my mother: a conversation*)

ACKNOWLEDGMENTS

I thank my advisors, Prof. Duncan, for the opportunity and guidance, and Prof. Milene for considerations throughout this study. I thank HP[1] for the full scholarship.

There are some special people that I would like to thank:

First, to my old friend Henrique Jung. Thank you for your support and for the countless times you helped me.

Second, to Hildegard Jung. Thank you for your guidance and tireless help over the years. You will always be an inspiration to me.

Third, to friends and colleagues from LIS for their partnership over the past two years. A special thanks to Andrey Salvi for his compassion, emotional help, and encouragement in the final stretch.

Finally, I thank my father, Donário Jandrey, for his support and for encouraging my education.

[1] Research supported by HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda. using financial incentives of IPI refund reggarding the Law (Law nº 8.248 of 1991).

DESCRIÇÕES DE CONTEÚDOS DE IMAGENS PARA PESSOAS COM DEFICIÊNCIA VISUAL

RESUMO

Descrições de imagens visam expressar, em palavras, o conteúdo visual e são essenciais para pessoas que não têm visão. Tais sentenças descritivas são geradas manualmente ou por modelos de Inteligência Artificial (IA). Apesar da sua relevância, a emergência de geradores de descrições automáticas não foi motivada por pessoas com deficiência visual. Portanto, elas ainda causam insatisfação em sua audiência. Neste estudo, nós investigamos problemas em descrições de imagens na literatura por meio da técnica de *Snowballing*, onde encontramos treze problemas, incluindo aqueles relacionados à Ética, tais como a aparência física, gênero e identidade, raça e deficiência. Nós identificamos cinco razões do porquê pessoas videntes não escrevem descrições para os conteúdos visuais, demonstrando a necessidade de campanhas de acessibilidade para conscientizá-las da importância social das descrições de imagens. Além disso, nós realizamos um conjunto de entrevistas com oito participantes com baixa visão. Nós exploramos as características das sentenças descritivas de 25 imagens de ambientes internos e coletamos as expectativas de descrições de imagens dos participantes. Portanto, através dos resultados do *Snowballing* e das entrevistas, nós propomos um conjunto de Boas Práticas para auxiliar as ferramentas automáticas e as pessoas videntes na escrita de descrições de imagens de mais satisfatórias e de qualidade. Nós esperamos que os nossos resultados ressaltem a relevância social de sentenças descritivas e encorajam a comunidade a prosseguir com pesquisas interdisciplinares que possam potencialmente minimizar os problemas encontrados no nosso estudo.

Palavras-Chave: Boas práticas, Descrições de imagens, Pessoas com deficiência visual, Estudo qualitativo, *Snowballing*.

IMAGERY CONTENTS DESCRIPTIONS FOR PEOPLE WITH VISUAL IMPAIRMENTS

ABSTRACT

Image descriptions intend to express, in words, the visual content and are essential for people who do not have eyesight. Such descriptive sentences are generated manually or by Artificial Intelligence (AI) models. Despite its relevance, the emergence of automatic description generators was not motivated by people with visual impairments; thus, they still cause dissatisfaction in their audience. In this study, we investigate image descriptions issues reported in the literature through the Snowballing technique, where we found thirteen problems, including those related to Ethics, such as physical appearance, gender and identity, race, and disability. We have identified five reasons why sighted people do not write descriptions for visual content, raising the need for accessibility campaigns to make them aware of the social importance of image descriptions. In addition, we conducted a set of interviews with eight low vision participants. We explored the characteristics of the descriptive sentences of 25 indoor images and collected the participants' expectations of image descriptions. Therefore, through the results of the Snowballing and the interviews, we propose a set of Best Practices to help automatic tools and sighted people in writing more satisfactory and quality descriptive sentences. We hope our results will highlight the social relevance of image descriptions and encourage the community to pursue further interdisciplinary researches that could potentially minimize the issues encountered in our study.

Keywords: Good practices, Image descriptions, Visually impaired people, Qualitative study, Snowballing.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1. INTRODUCTION

There has been a noteworthy increase in online visual content production, such as photos, videos, and user-generated images [60, 31], having revolutionized our digital communication practices to a majority imagery-based type [60]. The increase is, among other reasons, since human beings process visual data faster than textual ones [60, 21]. Nevertheless, without proper descriptions, imagery contents are inaccessible and discriminate against those unable to use eyesight, such as visually impaired people [15, 31].

To consume digital content, people with visual impairments use assistive technologies as, for example, screen readers [31, 69]. Regarding imagery content, screen readers rely on descriptive sentences, also known as alternative texts, to convey the visual content, and it is an elementary recommendation from accessibility standards [69, 24, 96]. Nevertheless, most digital images remain inaccessible to screen reader users due to insufficient or lacking descriptions [31].

As WebAIM Million [97] reveals, an annual accessibility analysis conducted by WebAIM, in February 2021, 60.6% of the 1,000,000 most worldwide accessed sites lacked alternative text for images. One of the reasons is that manual descriptive sentences are laborious since they depend on the professionals engaged in the website's creation [102, 31], such as developers and designers. Moreover, accessibility guidelines only offer general recommendations for the alternative texts' generation [69, 82], and there is no common consensus on how to describe the visual content nor standards defining its process [82].

An option to the alternative texts' absence is automatic image descriptions generators [31], which employ Artificial Intelligence (AI) models and integrates several of AI's research fields, including image processing, computer vision, and natural language processing [55]. These AI models aim to recognize and understand the images' elements, fitly describe their relationship, and generate a descriptive sentence [5]. Therefore, it requires a thorough understanding of what is relevant and imperative in an image description [55]. Despite AI models' progress, their identification of assorted images' elements is scanty and restrict to the most eminent components, yielding less heterogeneous image descriptions [13, 22]. Thereby, it is a daunting task to build AI models that produce sundry, creative, and akin to those manually generated descriptive sentences [57, 22].

The emerge of automatic image descriptions generators did not occur because of people with visual impairments, despite their benefits for this audience [57]. Instead, its emerge was due to the evolution of summarization systems [67, 92]. Thereby, their generated sentences may be meager, as their images' inference does not consider the visually impaired people's preferences of image descriptions [57]. From this statement, and since lacking instructions on how to describe imagery-based content [69, 82], and because sighted

people are unaware of the descriptive sentences' social impact [76, 27], we identified an opportunity to explore image descriptions with people with visual impairments.

In this study we explored automatic and manual (human) image descriptions with low vision participants. We aimed to identify characteristics of satisfactory and unsatisfactory descriptive sentences, their expectations of image descriptions, and investigate how to describe an image that enhances visually impaired people's image understanding. Our main contribution is a set of good practices for image descriptions. From our results, we expect to support imagery descriptions' generation, hence benefiting visually impaired people.

1.1 Research Questions and Objectives

The following research questions guided the development of this study:

1. What are the current image descriptions' issues for people with visual impairments?
2. What are the characteristics of satisfactory and unsatisfactory image descriptions?
3. What are the image descriptions' expectations of visually impaired people?

Research Question 1 sought to identify, in the literature, the current image descriptions' issues for people with visual impairments through the Snowballing technique. Research Question 2 aimed to identify the characteristics of satisfactory and unsatisfactory image descriptions for visually impaired people. Finally, Research Question 3 intended to discover the imperative content of descriptive sentences that meet visually impaired people's expectations.

The main objective of this research is to **assist the generation of descriptive sentences through a set of good practices for writing them**. Therefore, we defined the specific objectives based on the Research Questions (RQ) formerly presented:

1. Identification of usual issues of image descriptions for those with visual impairments (RQ 1);
2. Delineation of satisfactory and unsatisfactory image descriptions (RQ 2);
3. Identification of imperative content of descriptive sentences (RQ 3);
4. Identification of **what** and **how** to describe imagery content for visually impaired people (RQ 3).

1.2 Methodology

Regarding the objectives, this research is exploratory since we aimed to acquire insights and familiarity with the participants [51]. We used a multi-method strategy for the data collection and analysis. The term multi-method refers to qualitative and quantitative procedures usage either concurrently or sequentially, as each research phase had a particular aim and addressed a distinct purpose [11]. The steps are delineated as follows:

1. **Theoretical Framework:** We investigated human and automatic image descriptions' generation, seeking to identify its practices and evaluation methods. In this phase, we decided the number of images, sources, and descriptive sentences. In total, we selected 25 images from the MS COCO dataset [14] that portray internal environments, such as houses' rooms, offices, libraries, restaurants, among others. We decided on the internal scope since outside environments did not contain many objects and may not contain rich imagery details since they are mainly landscapes.

   We used the IBM Image Caption Generator Tool to generate the respective automatic image descriptions. Automatic descriptions are from the IBM MAX Image Caption Generator Model [44], a tool based on the study of Vinyals *et al.* [93], which are the winners of the 2015 MSCOCO Image Captioning Challenge. The model is freely available and displays three generated captions describing an input image.

   The image descriptions, either human and automatic, are generated in the English language; thus, a Linguistic Professional living abroad for over 15 years translated the descriptive sentences into the Portuguese language. Besides, in this step, we identified studies that explored visually impaired people's experiences with automatic image descriptions;

2. **Research Protocol:** we searched for qualitative studies with visually impaired participants. We did not perform a deep investigation since our objective was to corroborate our research protocol, *i.e.,* to verify the adequacy of the data collection method and the number of participants we had defined. This step was crucial for the research project's submission to the Research Ethics Committee (REC). Therefore, the selected studies had specified the research protocol, the number of participants, and the data collection technique. Thereby, we set our research protocol as follows:

   (a) **Sample profile selection:** participants with low vision, screen reader users, and professionally active people. We decided to recruit low-vision people because of their visual acuity and functional vision. Besides, most low-vision people had acquired the visual impairment in some moment of life; thus, they have imagery memories. Further, to respect the isolation imposed by Health Authorities, the interviews were remotely

executed, which required screen reader users. Professionally active people have personal experiences in non-familiar contexts and may provide more meaningful insights;

(b) **Data collection method:** we decided to collect data through semi-structured interviews. We sought qualitative studies with visually impaired participants to support our choice to use this data collection method. We delineated the selected studies in Chapter 3, Section 3.1.

(c) **Required data and resources:** images and their human descriptions are from the MS COCO dataset [14], and the automatic image descriptions are from IBM CODAIT [44];

(d) **Submission to the Research Ethics Committee (REC):** we submitted the research project to the REC. After its approval, we performed the interview study;

(e) **Pilot test:** we carried out a pilot test with one low vision participant to check necessary adjustments in the interview guide;

(f) **Participants recruitment:** we recruited participants through personal contacts and organizations;

(g) **Study conduction:** an interview study with the volunteer participants, lasting about 2 hours. We performed the interviews through the Zoom meeting tool;

(h) **Result analysis:** we analyzed the characteristics of the image descriptions and the participants' expectations.

3. **Image Descriptions Rating:** we developed and distributed an online survey for sighted people to rated the human and automatic-generated image descriptions through a Likert Scale. This step was necessary because either image descriptions sources provided more than one descriptive sentence, more specifically, five human image descriptions and three automatic ones. Thus, it would be very time-consuming to evaluate all original image descriptions in an interview study. Thereby, sighted people rated the descriptive sentences we selected the best-rated (one of each type) for the interview study with visually impaired participants.

   We received 57 responses, whose ratings of the 25 images and their eight image descriptions are available in Appendix C. One image had a technical tie in their description ratings, so we opted for those with the highest grade of 5 on the Likert Scale. The 25 images previously mentioned and their respective automatic and human-generated descriptions are in Appendix D;

4. **Snowballing:** we investigated, through the Snowballing technique, image descriptions' issues. We aimed to understand what visually impaired people find unsatisfactory in descriptive sentences. Besides, we identified factors inhibiting manual image descriptions' generation, and recommendations to some of the identified issues. The snowballing results will be presented in the Brazilian Symposium on Human Factors

in Computing Systems (IHC-2021) conference [47], which is the main event in the Human-Computer Interaction area in Brazil;

5. **Interview Study:** we performed an interview study with low vision participants. We used semi-structured interviews for a deeper understanding of research subjects' perceptions and experiences [66]. We analyzed the data through the qualitative content analysis method, using the inductive coding approach. In this step, we explored satisfaction and unsatisfactory reasons in descriptive sentences and image descriptions' expectations of the participants.

6. **Results Analysis:** lastly, we analyzed the data collected in the interview study, extracted meaningful insights to answer the Research Questions, and achieved the defined objectives. In this step, we identified a set of good practices for writing image descriptions.

Figure 1.1 presents the previously delineated steps.

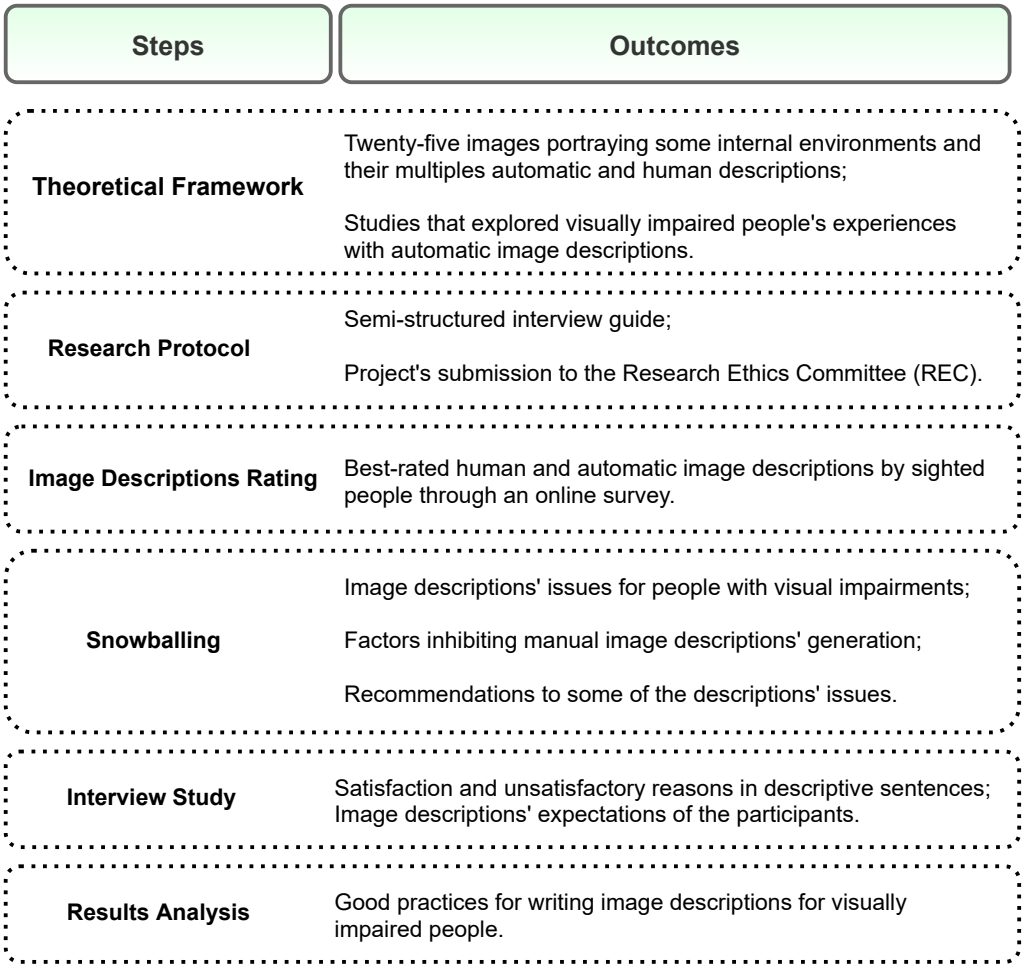


Figure 1.1: Research Methodology Steps.
Source: From the authors.

1.3 Contributions

This research aimed to assist the generation of image descriptions for people with visual impairments. Through a literature investigation, we identified thirteen image descriptions issues faced by visually impaired people, as well as factors inhibiting manual image descriptions' generation, and recommendations to some of the identified problems. Furthermore, we sought to explore, with low vision participants, human and automatic image descriptions, seeking relevant imagery elements and identify satisfactory and unsatisfactory image descriptions' characteristics. Moreover, we collected image descriptions' expectations from the research subjects. Their reported expectations were crucial for understanding how suitably describe imagery for visually impaired people.

Therefore, the main contribution of this research is a set of **nineteen good practices for writing image descriptions for people with visual impairments**. We expect to contribute to future AI models and manual image descriptions by sharing the needs and expectations we collected and clarifying how to describe imagery content.

1.4 Document Structure

The remaining of this study is: Chapter 2 presents the theoretical framework, Chapter 3 presents the related studies and the Snowballing results. Chapter 4 describes the results of the interview study, and Chapter 5 delineates the main contribution of this study, which is a set good practices for writing image descriptions. Lastly, Chapter 6 presents the study's final considerations.

2. THEORETICAL FRAMEWORK

This chapter delineates image descriptions generation. Thus, in Section 2.1 we present the classifications of image descriptions. Section 2.2 supports the generation of automatic image descriptions, and due to its complexity, we detail the approaches, datasets, and evaluation metrics used. Finally, Section 2.3 reports manually generated image descriptions.

2.1 Classification of Image Descriptions

Classification of image descriptions occurs according to the body of expertise required and their detail level output. As Hollink *et al.* [39] define, a **non-visual** image description refers to the image's metadata, *e.g.,* time, photographer, and the scene location, and is uninfluenced by any human interpretation. In contrast, a **perceptual** image description contains the characteristics of the image's elements, *e.g.,* their position, orientation, and relative distance. Besides, it delineates the visual properties of the image, for example, either it is a photograph or a painting, its texture, color, and size, and does not require prior knowledge about the image's components to interpret its visual content.

Conceptual image description delineates the imagery content, centering on the scene, its elements, and their attributes, as well as their relationship. Due to the umpteen possibilities of describing imagery content, a conceptual description is either **abstract**, *i.e.,* generically describes an image's scene, or **specific**, *i.e.,* meticulously describes the image's content [46, 39, 38]. As highlighted by James and Chang [46], this image description's type requires former knowledge and the observer's interpretation, thus tending to be subjective. Conceptual descriptive sentence production is the goal of AI image descriptions as it is the most relevant to the image understanding task [38]. From the definitions above, we decided to use **conceptual image descriptions** for the interview study.

Figure 2.1 exemplifies the three classifications of image descriptions.

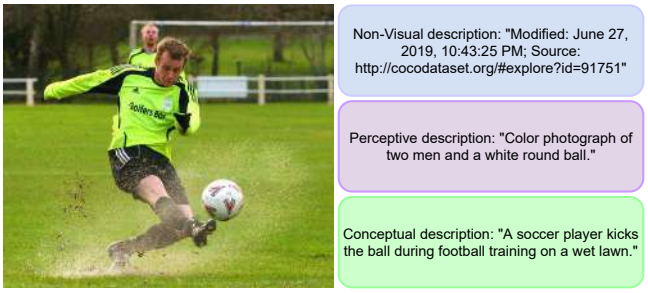


Figure 2.1: Classification of image descriptions: non-visual, perceptual, or conceptual. Source: MSCOCO dataset [14], id=91751. Adapted by the authors.

2.2 Generation of Automatic Image Descriptions

Generation of automatic image descriptions employs AI (Artificial Intelligence) models [83], a task known as image captioning. The generators' AI models recognize the salient elements, understand their proper relationship, and generate a sophisticated descriptive sentence of an image's content [5]. Therefore, this task requires a deep image understanding and is seemingly more complex than object detection or segmentation tasks [59, 9]. Several areas employ automatic image descriptions [55, 40], including biomedical [52], social media platforms [30, 57, 84], web searching [33, 18], and assistive technologies [84, 59].

Automatic image descriptions generators employ deep learning techniques as, for example, encoder-decoder framework and attention mechanism [40, 78]. Regarding the first mentioned approach, given an image as input, the encoder reads it, extracts its data, and places it in a feature vector, so the decoder generates descriptive words of the visual content [3], as Figure 2.2 presents. Most AI models employ a Convolutional Neural Network (CNN) as the encoder and a Recurrent Neural Network (RNN) as the decoder, especially Long Short-Term Memory (LSTM) [37].

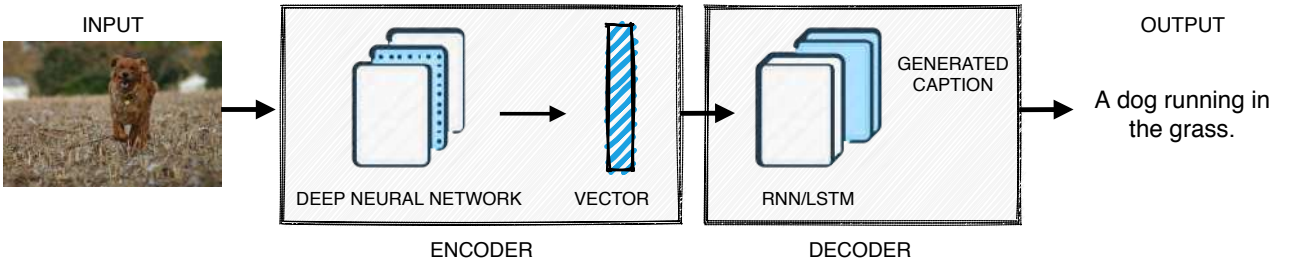


Figure 2.2: Encoder-decoder framework. A deep learning model encodes an input image into a feature vector. The language model decodes the input vector to generate a descriptive sentence. Source: adapted from [3].

However, rather than judging images as a set of various frames (or regions), this approach generates descriptive sentences by considering imagery content as a whole [40]. In contrast, the attention mechanism intends to replicate the human behavior to pay attention to specif images' regions before describing them [85], as Figure 2.3 shows. This approach is widespread since it dynamically focuses on various imagery regions during output production [40].

2.2.1 Datasets

Image descriptions generators based on deep learning methods require data for training, testing, and evaluating their image inference. There are many available datasets,

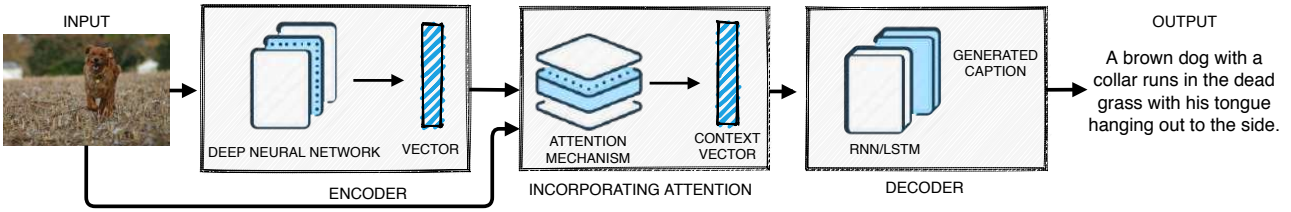


Figure 2.3: The attention-based model learns to focus on different regions of an image. Instead of encoding all content as a static vector, the attention mechanism adds the corresponding data of each region into the content vector, generating a more detailed descriptive sentence. Source: adapted from [6].

and they differ in the number of images, the number of captions per image, the elements categories, *e.g.,* people, animals, and objects, among other features. We delineated some of the public datasets widely in the AI-based image descriptions generation:

- Flickr 8K [38]: it has 8,000 images collected from Flickr. Each image contains five descriptions provided by human annotators;
- Flickr 30K [73]: It contains 30,000 images from Flickr, focusing mainly on people and animals, and 158,000 captions provided by human annotators;
- MS COCO [14]: Microsoft Common Objects in COntext (MS COCO) dataset contains 80 object categories, 330,000 images, and five captions per image;
- Conceptual Captions [79]: it has around 3.3M image-URL and caption pairs. It contains many imagery contents, including landscapes, products, professional photos, cartoons, drawings, and so on. The descriptions are from the Web and present a wider variety of styles;
- VizWiz-Captions [34]: it consists of over 39,000 images originating from people who are blind and has five descriptions per image.

2.2.2 Automatic Evaluation Measures

There are many metrics to evaluate automatic image descriptions. For example, BLEU [67], METEOR [20], and ROUGE [54] metrics emerged to estimate automated translation and summarization systems, whereas the CIDEr [92] metric specifically evaluates descriptive sentences, showing more accuracy than the previously mentioned metrics. These metrics outcomes a score regarding the similarity between the candidate and the reference sentence, and we delineated them below:

- **Bilingual Evaluation Understudy (BLEU)** [67] is one of the first metrics for measuring the similarity between two sentences. The algorithm scales the proximity between an auto-

matic translation and a human reference sentence by considering the length and orderly of the generated translation.

In Natural Language Processing (NLP), an *n*-gram is a words sequence, where *n* is the number of words. In the literature, the metrics usually use up to 4-gram. To clarify this term, Table 2.1 exemplifies the following sentence separated in n-grams: “This is a simple sentence”.

Table 2.1: Example of a sentence separated in *n*-grams.

1-gram	2-gram	3-gram	4-gram
This	This is	This is a	This is a simple
is	is a	is a simple	is a simple sentence
a	a simple	a simple sentence	
simple	simple sentence		
sentence			

BLEU metric counts the number of *n*-grams matches between the candidate and the reference sentences. These matches are position-independent and similarly computed a modified unigram precision for any *n* to repetitions elimination, as Equation 2.1 shows:

$$CP_n(C,S) = \frac{\sum_i \sum_k \min(h_k(c_i), max_{j \in m}(h_k(s_{ij}))}{\sum_i \sum_k \ h_k(c_i)} \tag{2.1}$$

where: $h_k(c_i)$ = number of times an *n*-gram occurs in candidate(c_i) $\in C$.
$h_k(s_{ij})$ = number of times an *n*-gram occurs in reference (s_{ij}) $\in S$.
$max\ h_k(s_{ij})$ = maximum number of times an *n*-grams occurs in reference.
k = indexes the possible *n*-grams of length *n*.

However, the precision score favors short candidate sentences since it relies only on the candidate sentence’s length, meaning that a short candidate can acquire high precision even in few matches with the reference sentence. Thereby, a brevity penalty (Equation 2.2) compensates it:

$$b(C,S) = \begin{cases} 1 & \text{if } \ell_c > \ell_s \\ e^{1-\frac{\ell s}{\ell c}} & \text{if } \ell_c \leq \ell_s \end{cases} \tag{2.2}$$

where: ℓ_c = is the total length of the candidate sentence corpus.
ℓ_s = s the length of the corpus-level effective reference length.

Thereby, the computation of the BLEU score uses a weighted geometric mean of the individual *n*-gram precision (Equation 2.3):

$$BLEU_n(C,S) = b(C,S)\exp\left(\sum_{n=1}^{N}\omega_n \log CP_n(C,S)\right) \tag{2.3}$$

where: N = 1, 2, 3, 4.
ω_n = is typically held constant for all n and it is set to $\frac{1}{\omega_n}$.

• **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** [54] is a measure package for the automatic text summaries' evaluation and includes three metrics.

– **ROUGE-N: N-gram Co-occurrence Statistics** is an *n*-gram recall between a candidate sentence and a set of reference sentences. It computes the *n*-gram matches between a candidate and a reference sentence divided by the number of *n*-gram of the reference sentence (Equation 2.4):

$$ROUGE_N(ci,Si) = \frac{\sum_j \sum_k \ min(h_\kappa(c_i), h_\kappa(s_{ij}))}{\sum_j \sum_\kappa \ h_\kappa(s_{ij})} \tag{2.4}$$

where: $h_k(s_{ij})$ = number of times a *n*-gram occurs in reference.
$min\ (h_k(c_i), h_k(s_{ij}))$ = number of times that an n-gram occurs in a candidate sentence (c_i) and in a reference sentence (s_{ij}).

– **ROUGE-L:** is based on **Longest Common Subsequence (LCS)**. An LCS is a set of words shared between the candidate and the reference in the same order but not necessarily in sequence. The ROUGE-L results from an F-measure computation (Equations 2.5, 2.6, and 2.7):

$$R_\ell = max_j \frac{\ell(c_i, s_{ij})}{|s_{ij}|} \tag{2.5}$$

$$P_\ell = max_j \frac{\ell(c_i, s_{ij})}{|c_i|} \tag{2.6}$$

$$ROUGE_L(c_i, S_i) = \frac{1 + \beta^2 R_\ell P_\ell}{R_\ell + \beta^2 P_\ell} \tag{2.7}$$

where: $\ell(c_i, s_{ij})$ = is the length of the LCS between a pair of sentences.
β = is usually set to favor recall and its value is $\beta = 1$ or more.

– **ROUGE-S:** is based on **Skip-Bigram Co-Occurrence Statistics** which is any pair of ordered words but not necessarily in sequence, and it computes F-measure (Equations 2.8, 2.9, and 2.10). It is important to note that, in practice, arbitrary gaps are allowed by using a maximum distance of four words between the components of the skip bigrams.

$$R_s = max_j \frac{\sum \ min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(s_{ij})} \tag{2.8}$$

$$P_s = max_j \frac{\sum_k \ min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(c_i)} \tag{2.9}$$

$$ROUGE_S(c_i, S_i) = \frac{1 + \beta^2 R_s P_s}{R_s + \beta^2 P_s} \tag{2.10}$$

where: $f_\kappa(s_{ij})$ = is the skip bi-gram count for a reference sentence (s_{ij}).
$f_\kappa(c_i)$ = is the skip bi-gram count for a candidate sentence (c_i).
β = is usually set to favor recall and its value is β = 1 or more.

• **Metric for Evaluation of Translation with Explicit ORdering (METEOR)** [20] scores with an aim of 1:1 correspondence, thus matches between the reference and the candidate sentences are based on exact words, words of the same stem, synonyms, and paraphrase. The matches *m* is computed while minimizing the number of chunks *ch*. Chunk is a series of matches (correspondence) that is continuous and identically ordered in either reference and candidate sentences. The precision P_m (Equation 2.11) and the recall R_m (Equation 2.12) are calculated as follows:

$$P_m = \frac{|m|}{\sum_k h_k(c_i)} \tag{2.11}$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})} \tag{2.12}$$

Then, the harmonic mean of P_m and R_m is calculated (Equation 2.13):

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1 - \alpha) R_m} \tag{2.13}$$

A penalty *P* to account for the gaps and the differences in word order is scored, based on the total number of matched words (*m*) and the number of chunks (*ch*) (Equation 2.14):

$$Pen = \gamma \left(\frac{ch}{m} \right)^{\Theta} \tag{2.14}$$

Finally, METEOR score is computed (Equation 2.15):

$$METEOR = (1 - Pen) F_{mean} \tag{2.15}$$

where: α, γ, Θ = are parameters tuned to maximize correlation with human judgments and normally are set as α = 0.70, γ = 0.30 and Θ = 1.40.

• **Consensus-based Image De-scription Evaluation (CIDEr)** [92] is a specialized metric for image captioning evaluation, and it measures the consensus between a candidate and reference sentences. First, either sentence has all words mapped to their stem or root forms. Each *n*-gram ω_k contains one to four words, such as the BLEU metric. The *n*-grams common in all images are down-weighted since they are less informative. For each *n*-gram matched, a Term Frequency Inverse Document Frequency (TF-IDF) (Equation 2.16) weighting $g_k(s_{ij})$

is computed using the logarithm of the number of images in the dataset $|I|$ divided by the number of images for which ω_k occurs in any of their reference sentences:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{w_\ell \in \Omega} h_\ell(s_{ij})} \log\left(\frac{|I|}{\sum_{I_p \in I} min(1, \sum_q h_k(s_{pq}))}\right) \tag{2.16}$$

where: Ω = is the vocabulary of all *n*-grams.
I = is the set of all images in the dataset.

The first term measures the Term Frequency of each *n*-gram ω_k and the second term measures the rarity of the weight ω_k using its IDF. The TF sets a higher weight on *n*-grams that frequently occur in a reference sentence, while the IDF reduces the weight of the *n*-grams that commonly occur across all images in the dataset.

For *n*-grams of any length *n* the score is computed using the average cosine similarity between the candidate sentence and the reference sentences, which accounts for both precision and recall measures (Equation 2.17):

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i)\, g^n(s_{ij})}{\|g^n(c_i)\|\, \|g^n(s_{ij})\|} \tag{2.17}$$

where: $g^n(c_i)$ = is a vector formed by $g_k(c_i)$ corresponding to all *n*-grams of length *n*.
$g^n(s_{ij})$ = is a vector formed by $g_k(s_{ij})$ corresponding to all *n*-grams of length *n*.
$\|g^n(c_i)\|$ = is the magnitude of the vector $g^n(c_i)$.
$\|g^n(s_{ij})\|$ = is the magnitude of the vector $g^n(s_{ij})$.

The lengths of the *n*-grams can vary to capture grammatical properties and richer semantics information, so the scores from *n*-grams of varying lengths are combined (Equation 2.18):

$$CIDEr_(c_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i) \tag{2.18}$$

where: N = 1, 2, 3, 4.
ω_n = is typically held constant for all *n* and is set to $\frac{1}{\omega_n}$.

The term “gaming” refers to the phenomenon where a poorly sentence judged by humans may scores highly with an automated metric. To prevent its effects, a CIDEr-D modification is computed. First, the stemming is removed to ensure the use of correct forms of words. In long sentences, when words with higher confidence are repeated, the basic metric Cider produces a higher score, so a Gaussian penalty based on the difference between candidate and reference lengths was introduced. Finally, the number of an *n*-gram occurrence in a candidate sentence is limited to its occurrence in a reference sentence (Equation 2.19):

$$CIDErD_n(c_i, S_i) = \frac{10}{m} \sum_j e^{\frac{-(\ell(c_i)-\ell(s_{ij}))^2}{2*\sigma^2}} * \frac{min(g^n(c_i), g^n(s_{ij}))\ g^n(s_{ij})}{\|g^n(c_i)\|\ \|g^n(s_{ij})\|} \tag{2.19}$$

where: $\ell(c_i$ = denotes the lengths of the candidate sentences.
$\ell(s_{ij}$ = denotes the lengths of the reference sentences.
σ = a held constant, set to $\sigma = 6$.

The final CIDEr-D metric is computed as Equation 2.20 shows:

$$CIDErD_(c_i, S_i) = \sum_{n=1}^{N} w_n CIDErD_n(c_i, S_i) \tag{2.20}$$

where: N = 1, 2, 3, 4.
ω_n = is typically held constant for all n, and set to $\frac{1}{\omega_n}$.

2.2.3 Human Evaluation Measures

Quality evaluation of AI-generated image descriptions includes humans beings, which normally verify the consistency of automatic metrics output and the human rating [75, 41]. Evaluation metrics should satisfy two criteria [41]: (a) descriptive sentences considered good by human evaluators must achieve high scores in automatic metrics, and (b) high scored descriptive sentences must be considered good by them human evaluators. The referred evaluators include experts and untrained people by the crowdsourcing method. As Rashtchian *et al.* [75] state, crowdsourcing data collection is through an online framework, like Amazon's Mechanical Turk (MTurk); however, human-based evaluation creates additional costs since it is slow, subjective, and arduous to reproduce [41].

2.3 Human Image Descriptions Generation

Manually generated image descriptions are one of the most basic web accessibility criteria according to Web Content Accessibility Guidelines (WCAG) [96]. Usually, the professionals involved in website creation and its maintenance describe the imagery contents. World Wide Web Consortium (W3C) [95] provides guidelines for imagery descriptions' production to meet the criteria specified by WCAG, considering different image categories. Each imagery category has specific instructions for describing the visual content based on the image's purpose [95]. For example, informative imagery encompasses pictures, photos, and illustrations, and their descriptions should convey the meaning or visual content [95].

More recently, people who aren't involved in creating websites have the power to write their descriptive sentences for the content they produce. For example, in 2016, Twitter released a feature for entering manual descriptions [63]. Concerning the Brazilian context, Patricia Silva de Jesus created in 2012 a social media campaign to include image descriptions using the hashtag #PraCegoVer [19]. Floriano, Cassanego Junior, and Silva [27] investigated the adoption and acceptance of this campaign by sighted people and by public and private organizations analyzing more than 2500 posts. They reported a lack of discussions that could extend the understanding and adoption of #PraCegoVer [27]. Besides, the authors identified that only 73 public and private organizations use PraCegoVer in their posts, demanding initiatives to raise people's consciousness of the importance of image descriptions for visually impaired people [27].

3. RELATED STUDIES

This Chapter presents the related studies. Section 3.1 presents qualitative studies with visually impaired participants. Section 3.2 discusses studies whose research topic includes image descriptions from the perception of visually impaired people. Section 3.3 reports the Snowballing conduction and achieved results.

3.1 Qualitative studies' conduction with visually impaired participants

The studies reported in this section supported our research protocol. Thus, we present the **studies' conduction** rather than deeply report their limitations and results.

Chandrashekar *et al.* [12] conducted a simultaneous verbalization technique with six visually impaired students during a website's evaluation. The authors found it troublesome to use the simultaneous verbalization technique since they read the textual content through a screen reader. The participants expressed reluctance to pause and restart the speech for articulating their comments, resulting in confusion and discomfort. Thus, the authors performed interviews to collect the participants' experiences and the decision-making process.

Ferreira *et al.* [26] evaluated site accessibility with blind people and provided recommendations for observational studies with visually impaired people. The authors performed two observational studies with five participants in each one. Therefore, the authors observed the interactions of five users in their work environment (users' context of use) and five users in a lab (controlled context). For each study, they performed a pilot test checking possible issues and improvements in the evaluation structure.

In the controlled environment, participants received an accessible document containing the research objective, the methods adopted, and the software used. Then, the researchers collected users profiles data and presented the tasks individually. Through an adapted simultaneous verbalization technique, participants expressed their decision-making process and interaction strategies during the study. As a result, the authors provided recommendations as, for example, limiting the number of participants to five users of the same profile, performing a pilot test, applying a questionnaire to identify the users' profiles, among others.

Adams, Morales, and Kurniawan [2] developed an app to assist photography for visually impaired people. The authors performed an online survey with blind people to understand their photography habits, needs, and preferences, collecting 54 valid responses. The researchers asked participants' demographics data and their behavior of taking and sharing photos. Participants that did have this behavior answered their reasons in an open-

ended format. Through a series of Likert-scale questions, participants rated satisfaction with current tools for capturing, recognizing, and sharing photos. In an optional question, the participants answered the obstacles to taking pictures independently. Finally, the authors collected suggestions for the application in open-ended format questions.

Pereira, Ferreira, and Archambault [50] conducted a web accessibility evaluation with blind people focusing on the most critical problems previously classified by experts. The authors executed a pilot test, and the evaluations occurred with five participants in a familiar location, a computer lab in an association. The researchers applied a questionnaire to identify participants' profiles. Besides, the authors observed and encouraged participants to share their issues and difficulties during the interaction with the websites through simultaneous verbalization. The time limit for the task execution was 70 minutes, estimating 10 minutes for each task. As a result, the authors cite the difficulties encountered during participants' navigation in the selected sites, proving a set of critical points to ensure web accessibility. One of these recommendations refers to textual descriptions of images that should be equivalent to the visual content.

Choi *et al.* [16] proposed a prototype of a Google Chrome extension to extract data from charts and performed a qualitative evaluation with three visually impaired participants. The authors conducted structured interviews sessions to understand the participants' habits to access charts. Each interview lasted approximately 30 minutes. Participants reported rarely paying attention to charts or images due to lack of description, and they rely on the textual content to understand the visual content.

Once more, the authors recruited visually impaired people to evaluate the extension developed and did not specify the evaluation's location. The researchers observed the participants during the tasks and applied a questionnaire, followed by a feedback interview. Each session was conducted individually and lasted half an hour. According to participants, descriptive sentences of the tables were too long, slowing their navigation when read by the screen reader.

Table 3.1 summarizes the data extracted from the above studies. From this, we defined our research protocol, mentioned in the research methodology (Chapter 1, Section 1.2).

3.2 Image descriptions for people with visual impairments

Morris *et al.* [63] performed an online survey with 116 blind participants to collect their motivations and behaviors on Twitter. The authors also explored the Twitter profiles either 116 sighted people and the participants to compare their online activities, though the manual inspection method. Also, they investigated the prevalence of photos and videos from tweet metadata. Regarding interaction with visual content, 55.4% of the blind participants re-

Table 3.1: Research protocols of qualitative studies with visually impaired participants.

Study	Number of participants	Protocol	Method	Objective
[12]	Six (four blind and two low vision)	1) Participants selection; 2) Observations; 3) Data analysis.	1) Simultaneous verbalization; 2) Observation; 3) Interview.	Experiences and observations in using simultaneous verbalization during the evaluation of a website by blind users.
[26]	Ten (five for each observation method)	1) Participants selection; 2) Pilot test; 3) Study of tools and resources; 4) Observation in the users' context of use; 5) Observations in controlled environments; 6) Data analysis.	1) Questionnaire; 2) Simultaneous verbalization; 3) Observation (users' context of use and controlled environment).	Analyzing two observational methods involving visually impaired users to develop a protocol with recommendations.
[2]	Fifty-four	1) Development of the online survey; 2) Pilot test with screen reader user; 3) Questionnaire's launch; 4) Data analysis.	1) Questionnaire (structured and open questions).	Understanding the photography habits, needs, and preferences of blind people for blind-friendly app development.
[50]	Sixteen (first study); Three (second study)	1) Participants selection; 2) Questionnaire submission (email) or face-to-face interview; 3) Data analysis Study 1; 4) Interview; 5) Observation; 6) Data analysis Study 2.	1) Structured interview; 2) Remote structured questionnaire (email); 3) Face-to-face questionnaires; 4) Life story interview; 5) Observation.	Evaluating a conceptual design of an app for clothing.
[68]	Five	1) Participants selection; 2) Websites selection; 3) Pilot test; 4) Observation in a familiar IT laboratory; 5) Pilot test; 6) Data analysis.	1) Questionnaire; 2) Simultaneous verbalization; 3) Observation.	Evaluating websites' accessibility.
[16]	Three	1) Participants selection; 2) Questionnaire; 3) Structured interview; 4) Observation; 5) Data analysis.	1) Individual remote structured interviews; 2) Observation; 3) Questionnaire.	Understanding participants' visual data access. Evaluating a developed Google Chrome extension.

ported never asking other Twitter users to describe images and, 20.5% of them intentionally follow users who retweet imagery and add descriptions to them.

The vast majority of the participants expect imagery descriptions, including color, body language, action, clothing, among others. About tweet characteristics, 23.43% of sight users' tweets contained photos or videos, compared to 4.78% of blind users' tweets. Regarding the embedded multimedia in the tweet's metadata, 93.6% contained images and 6.4% enclosed videos. The authors evaluated how well a tweet text described the imagery content, finding that 61.8% of them would be poor descriptions, 27% would be minimally acceptable, and 11.2% would be good descriptions.

MacLeod *et al.* [57] explored, in two studies, how blind and visually impaired people experience automatic captions on Twitter. In the first study, six blind participants answered an interview about their experiences with the image descriptions. The authors identified that participants tended to trust in automatically generated sentences even when they do not make sense. The second study was through an online survey, and 100 participants answered it. The authors explored varied confident framing into positively (confidence level of the algorithm for an accurate description) and negatively (confidence level for an incorrect description). Their findings revealed that negative framing causes less trust in uncertain sentences than positive framing. Therefore, they suggest using negative framing to encourage distrust in incorrect automatic image descriptions.

Wu *et al.* [100] designed and evaluated an automatic alt-text feature for Facebook. The in-lab and interview study was performed with 4 participants and aimed to understand the benefit of image descriptions, the risk of providing incorrect ones, and how to mitigate it. The feature appeared to reduce the time consumption and the social cost to interact with imagery content. To validate on a larger scale, the authors conducted a field study with 9k visually impaired people for two weeks, separating them into control and test groups. Then, an online survey collected feedback about their experience, and some suggestions include extracting and recognizing text from imagery and more details about people, including their identity, age, gender, clothing, action, and emotional state.

Morris *et al.* [62] developed a series of prototypes to represent visually impaired people's interactions with visual content, and 14 blind participants evaluated three prototypes in a controlled environment. The authors identified that the progressive detail interaction contributed more to the participants' image understanding. This interaction type allowed participants to choose how many and what levels of detail they wanted to hear about imagery content. Besides, the authors asked how many detail levels should be as a standard, and most responses indicated that they should be at least three, but one participant reported that there should not be a standard number of levels, as it depends on each image. Moreover, another participant indicated that it is crucial to create guidelines on which details to include at each level.

Gleason *et al.* [31] investigated the dissemination of image descriptions on Twitter as since in 2016 it was available a feature for people to include posts' descriptions. They evaluated over one million imagery tweets, and their findings indicated that only 0.1% of the tweets had descriptive sentences for at least one image. Besides, the authors investigated popular Twitter accounts, and only 3 of the 50 most popular ones had used image descriptions, and the remaining never added them, including news organizations, politicians, and celebrities.

The authors also analyzed the Twitter profile of 94 blind people and collected tweets published by their friends. The authors found out that 18.4% of the tweets contained imagery, and only 4.6% had descriptions. Then, the authors interviewed 20 Twitter users who use to

write image descriptions in their posts. Some motivations cited by the participants were the impact on other people's lives and close connections with visually impaired people. Finally, participants suggested improving the Twitter feature to make it visible to sighted users rather than requiring manually able it and to provide a reminder to people describing the post's content.

Sacramento *et al.* [76] carried out three studies to investigate Brazilian imagery description practices in the social media context. The first study analyzed the interfaces of four social media platforms, YouTube, Facebook, WhatsApp, and Instagram, and evaluated their descriptions using screen readers. According to the results, all the platforms have emoticon descriptions, and WhatsApp and Instagram did not provide any description for stickers. None of the four described GIFs or videos, and only Facebook and Instagram provided guidelines on their help page.

The second study investigated description habits, and 333 people without visual impairment participated. The responses indicated that 228 people (68.47%) do not usually describe their posts, and the frequent reasons include unknown how to include imagery descriptions and how to write them. Seventy-six (19.82%) participants have writing image descriptions, and the reasons answered include the inclusion of visually impaired people and engagement with web accessibility context.

Finally, the third study had 100 screen readers users (78% blind and 18% low vision) and investigated their difficulties in visual content understanding. The majority (90%) indicated previously access to automatically described images, and 11.11% of them reported that image descriptions never meet their needs, expressing desiring for more detailed descriptions such as people's facial and body expressions.

Stangl *et al.* [84] investigated the imagery experience of 28 visually impaired participants across seven web sources: news, social networking, eCommerce, employment, online dating, productivity applications, and e-publications. The results demonstrated that participants always want to learn about people, objects, and the text present in images.

For the first source (news site), the participants indicated that their preferences depend on the image focus. For people focus, participants would like to know their physical appearance and whether they are famous or not. Otherwise, if the topic is events or sports, the description must report people's actions and image elements interaction. In the second source (social networking sites), participants wanted image descriptions that help them to understand the image's purpose. Also, they prefer full facial expressions and body language descriptions to assist their reaction/response decision for the image's content.

For the third source (eCommerce), participants' wanted to know clothes' colors and attributes, and for household and electronic items, they wanted material type, texts, symbols, and logos. For the fourth source (employment), neither participant remembered finding image descriptions on job sites or job boards. Their preferences refer to the people and work environment to learn more about the company offering the job.

Regarding the fifth source (online dating), none of the participants had access to it due to accessibility issues and reported preferences about physical characteristics and photography type. In the sixth source (productivity applications), participants shared not usually receive imagery documents, but they want to understand the visual contents and their importance. The authors received similar answers to the seventh source (e-publications), which refers to digital content.

3.3 Snowballing

We decided to use a snowballing approach to identify the relevant studies in the literature to answer the following Research Question: **"What are the current image descriptions' issues for people with visual impairments?"**. We followed snowballing procedures defined by Wohlin [99], and our snowballing results will be present at the 2021 Brazilian Symposium on Human Factors in Computing Systems (IHC) [47].

Snowballing technique requires an initial set of studies [99]; thus, our initial set are the studies formerly presented in Section 3.2. We selected those studies due to their scope, for answering the Research Question, their contributions and points of view reported. As Wohlin [99] explains the snowballing technique does not require searches in databases with a string since the search emerge from the references (Backward Snowballing), and citations (Forward Snowballing) of the initial set of studies. Therefore, we decided not to restrict the publication year of the potential studies. Besides, Snowballing technique is through iterations, *i.e.,* for each new selected study its references and citations are analyzed until no new study emerged [99].

- **Inclusion criteria:** we considered studies published in the following places, as they are related to the Human-Computer Interaction and Accessibility field:
  - 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).
- **Exclusion criteria:** we rejected the studies in the specified order:
  1. Duplicate or already selected studies;
  2. Titles outside the scope of the Snowballing;
  3. Summaries outside the scope of the Snowballing;

4. Introductions, results, and final considerations outside the scope of the Snowballing;

5. Did not respond to the Research Question.

- **Quality criteria:** to guarantee the quality of Snowballing, the following criteria were followed:

  – For each rejected study, we assigned a reason based on the defined exclusion criteria. Thus, it was possible to identify the number of studies rejected in each exclusion criteria;

  – The study under analysis must had discussed the issues imagery descriptions and not just mention them briefly, enabling an in-depth discussion of the extracted data.

We performed two interactions. Figure 3.1 presents a summary of the selection steps and the number of studies rejected at each step.

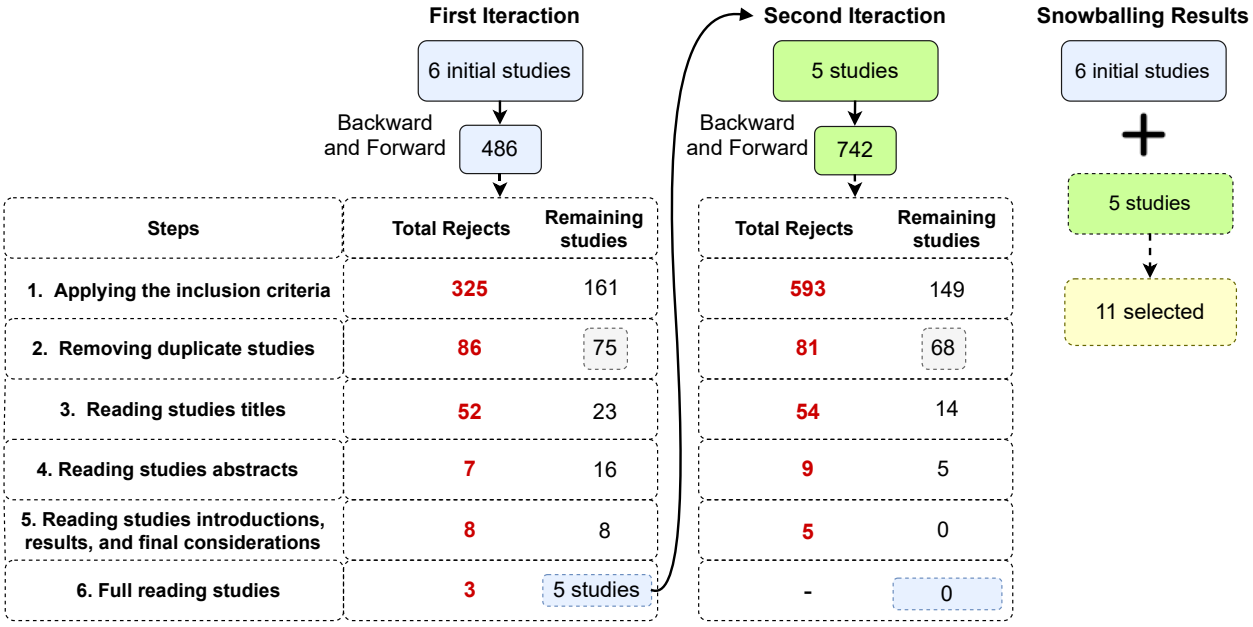


Figure 3.1: Summary of the steps for the studies' selection in both iterations with the respective numbers of rejects and remaining studies. Source: from the authors.

The initial set contained six studies, as we delineated in Section 3.2. In the First Iteration, the search procedures returned 486 potential studies. Of these, we rejected 325 after applying the inclusion criteria and another 86 after removing duplicate studies; thus, remaining 75 studies. Then, we excluded 52 studies after reading their titles, others 7 studies after reading study abstracts, and others 8 studies after reading their introductions, results, and final considerations. Lastly, we rejected 3 studies for not answering the Research Question, resulting in five new selected studies.

The Second Iteration refers to the five new selected studies. The search procedures returned 742 initial studies. Of these, we rejected 593 after applying the inclusion

criteria and another 81 after removing duplicate studies; thus, remaining 68 studies. Then, we removed 54 studies after reading their titles, others 9 studies after reading their abstracts, and others 5 studies after reading their introductions, results, and final considerations. Therefore, no new studies were selected.

After the Snowballing procedures, we selected eleven studies to answer the Research Question, which six of them belong to the initial set and the we selected five from the first interaction. Table 3.2 presents the eleven studies selected.

Table 3.2: Summary of the research contexts, the number of participants, and the data collection technique of the eleven selected studies.

Study	Context	Participants	Data collection
Morris *et al.* [63]	Twitter	112 blind people	→ Online survey → Manual inspection
Wu *et al.* [100]	Facebook	4 blind people	→ Controlled Observation → Interview
		375 screen reader users	→ Online survey
MacLeod *et al.* [57]	Twitter	6 visually impaired people	→ Observation → Interview
		100 visually impaired people	→ Online survey
Gleason *et al.* [31]	Twitter	20 sighted people	→ Manual inspection → Interview
Stangl *et al.* [84]	Web, social media, and productivity software	28 visually impaired people	→ Interview
Sacramento *et al.* [76]	Social media	333 sighted people e 81 screen reader users	→ Online survey
		19 screen reader users	→ Same survey but through phone
Lazar *et al.* [53]	Web	100 visually impaired people	→ Personal diary
Voykinska *et al.* [94]	Social media	11 visually impaired people	→ Interview
		60 visually impaired people	→ Online survey
Zhao *et al.* [102]	Facebook	12 visually impaired people	→ Interview
		6 people of 12 from the previous study	→ Personal diary → Interview
Gleason *et al.* [32]	Twitter	13 screen reader users	→ Interview
Bennett *et al.* [7]	Web	25 screen reader users	→ Interview

Nine studies comprises social media context such as Facebook [100, 84, 76, 94, 102] and Twitter [63, 57, 31, 94]. Four studies encompassed other platforms, such as YouTube [76], WhatsApp [76], Instagram [76, 94], LinkedIn [94], Snapchat [94], and Indeed [84]. One study also explored image descriptions in productivity applications [84], *e.g.,* Microsoft Word, Microsoft PowerPoint and Google Docs. Four studies explored the web context such as sites [84, 84], user's frustrations [53], and diversity issues in image descriptions [7].

Data collection techniques varied according to the number of participants. Five studies collected data through an online questionnaire, and the number of responses ranged from 1 to 100 [57, 76, 94], from 101 to 200 [63], and from 300 or more responses [100, 76]. Two studies used a questionnaire in the form of a personal diary: (1) an one-week study with six participants [102], and (2) a nine-month study with 100 participants [53]. Besides, one study reported some participants' preference to answer a questionnaire online by a phone [76]. Two studies also collected data through manual inspection [76, 63]. Eight studies used interviews and the number of participants ranged from 1 to 10 [100, 57, 102], from 11 to 20 [31, 94, 102, 32], and 21 or more participants [84, 7].

Based on data collected from the eleven studies, we identified issues in imagery descriptions for the visually impaired people, as presents the Subsection 3.3.1. Besides, we identified restrain factors for the generation of manual imagery descriptions in Subsection 3.3.2.

3.3.1 Image Descriptions Issues

Through the snowballing, we identified thirteen issues of image descriptions for visually impaired people, and we delineate each one below:

1. **Mediocre texts:** this issue refers to very generic image descriptions that, despite existing, do not provide enough data or detail to contextualize the scene [32]. Rich descriptive sentences are imperative to supply visually impaired people with the same or similar experiences as sighted people [84]. As Morris *et al.* [63] investigated, until June 2015, images represented 93.6% of Twitter content, and 61.8% of the texts associated with these contents would be poor descriptions, meaning that a blind person listening to it would not have any perception of what the image is conveying. Also, the authors considered 55.6% of the images as essential, *i.e.,* a person listening to the text without visualizing the image could not understand the tweet's purpose [63].

According to Voykinska *et al.* [94], the frequent lacking of helpful contextual descriptions on social media undermines the involvement of visually impaired people with most imagery content. Stangl, Morris, and Gurari [84] enhance that mediocre image descriptions or even their absence inhibit visually impaired people interaction on the social networks platforms, causing frustration and even confusion. Also, according to the authors, solely identifying the objects of an image is not enough to understand its representation, which can lead to misunderstanding [84].

As Wu *et al.* [100] noted, Facebook's automatic descriptions do not provide enough data for visually impaired people to feel encouraged to interact through comments or likes. Also, frequently, they do not generate any helpful information, requiring people close to visually impaired people to describe an image's content. The need to improve AI-generators

also emerged in the Brazilian context, as Sacramento *et al.* [76] observed in their study with 100 visually impaired participants. The authors identified that the second-highest difficulty faced by participants in accessing visual content in social media refers to over generic descriptions, which meet the participants' expectations only in some cases [76]. The authors also noted that visually impaired people look for alternatives to understand the visual content, such as looking for additional data in the comments or requesting help from a familiar sighted person [76].

2. **Missing texts:** visual contents without an associated description are inaccessible to visually impaired people, harming their digital inclusion and active participation on the Internet [94]. The absence of image descriptions causes feelings of isolation and frustration in visually impaired people, as they feel excluded or unable to participate in social media [94, 100].

In 2016, Twitter added a functionality to include manual image descriptions, and in 2018, Gleason *et al.* [31] investigated their prevalence. The authors collected over 9 million public tweets, of which above 1 million (11.84%) of them enclosed at least one image, and only 0.1% of these contained an alternative text [31]. According to Sacramento *et al.* [76], none of the four most popular social media in Brazil (WhatsApp, YouTube, Facebook, and Instagram) have automatic image descriptions in their mobile versions, and only Instagram allowed the manual insertion of imagery descriptions through the app mobile. Moreover, the absence of descriptions for visual content was the highest difficulty reported by participants in the four social media explored by the authors [76].

Lazar *et al.* [53] state that missing image descriptions can obstruct websites' access of visually impaired people that have security verification by typing embedded texts in imagery [53]. Furthermore, the absence of image descriptions was the fourth major cause of frustration reported by the 100 participants. Gleason *et al.* [32] also noted that the lack of accessible content is the main accessibility barrier on social media platforms.

3. **Inaccurate texts:** automatic image descriptions generators are still imperfect and can generate divergent texts from the images [57]. According to Gleason *et al.* [32], inaccurate image descriptions can lead to believe that an image contains something that is not present. Macleod *et al.* [57] observed that visually impaired people trust automatic image descriptions even when they present inconsistent with the respective images. Moreover, participants tried to justify or create scenarios to connect the unexpected descriptive sentence rather than suspecting that it was incorrect [57].

Wu *et al.* [100] observed, in their large-scale study, that participants demonstrated more intolerance for incorrect image descriptions, generating distrust in the algorithm and its output. Furthermore, Sacramento *et al.* [76] revealed that the presence of incorrect descriptions is the third greatest difficulty faced in the four social media most used by the 100 Brazilian participants.

4. **Descriptions not covering embedded texts:** this issue occurs when there are texts in images not included in the descriptive sentences. Including the embedded texts helps in understanding the images' contexts and purposes, for example, a news story containing a photo of a person holding a poster at a public demonstration [63]. According to Lazar *et al.* [53], images with embedded texts can cause frustration in people with visual impairments, which even an alternative audio version is not ideal due to the different pronunciations a word has.

Gleason *et al.* [32] observed that visually impaired people use strategies such as using applications that perform Optical Character Recognition. According to Morris *et al.* [63], several images categories contain textual information widely used, such as screenshots, motivational phrases, graphics, memes, and advertisements. The category variance and current limitations suggest that automatic description generators need to employ different approaches [63, 32].

5. **Confused texts:** this issue refers to incoherent and unclear image descriptions. According to Wu *et al.* [100], in the occurrence of confusing automatic image descriptions, visually impaired people guess the AI model's intention, trying to understand the imagery content. However, according to Sacramento *et al.* [76], this issue is also likely to occur in human image descriptions, as by professionals audio describers when using regionalism words, which causes discontent and confusion in those who are unaware of such expressions.

6. **Descriptions not covering humor and people's emotions:** Voykinska *et al.* [94] addressed the limitations of current technologies to cover the humor contained in images, *e.g.,* sarcasm, and comedy, as well as people's emotions, *e.g.,* anger, and sadness. Wu *et al.* [100] highlighted that people present in an image are considered the most intriguing image's element, and their humor is the most relevant characteristic followed by their action. However, automatic image descriptions are still limited in describing mood and emotions since it is highly complex to train algorithms to interpret these concepts [100].

7. **Scanty descriptions of the people's characteristics:** image descriptions are still limited in describing people's race, gender, disability, and other appearance feature [84, 7]. According to Stangl, Morris, and Gurari [84], describing people's appearance is a subjective task since they are not always implicit in images and can lead to incorrect descriptions. Furthermore, the authors noted that participants required people's race and gender descriptions in varied digital contexts; whereas, people's hairstyles, body height, weight, and eye color are preferred only in social and dating networking contexts [84]. Descriptions of people's appearance are highly relevant since sighted people can have these data from eyesight sense [84]. Bennett *et al.* [7] addresses the need to urgently adapting AI models' training to them consider non-binary, trans, and non-white people to ensure diversity.

8. **Scanty descriptions of the people's facial and body expressions:** this issue is similar to people's emotions, but describing facial expressions refers to facial features, *e.g.,* a

person looking seriously, showing the tongue, with a malicious look [84]. One of the demands reported in the study of Sacramento *et al.* [76] refers to more detailed descriptions of people's facial and body expressions. According to Stangl, Morris, and Gurari [84], facial expressions and body language are very relevant for visually impaired people in deciding how to react to imagery content. Furthermore, these people's features descriptions help visually impaired people to understand the images' purposes [84].

9. **Non-descriptive texts:** this issue refers to when an image description contains only the word "image" instead of describing its visual content. According to Lazar *et al.* [53] the presence of "non-descriptive texts" is the second biggest cause of frustration among screen reader users regarding image descriptions.

10. **Homogeneous texts for similar images:** As Zhao *et al.* [102] state, automatic image description generators are limited in providing more heterogeneous descriptions for similar images. This issue can hinder the imagination process of visually impaired people, besides obstructing the identification of diverse photos of the same event, *e.g.,* an album containing pictures of a family lunch. The authors observed that images' details or particularities are more relevant for low vision people than blind ones since they expect the sentences to convey more information than their visual acuity allows them to perceive.

11. **Erroneous texts for low quality images:** automatic generators execute their AI models regardless of the input images' quality; however, many photos present blur and fog, especially those captured by visually impaired people. According to Zhao *et al.* [102], a limitation of AI generators refers to them do not inform the images' conditions even when the inputs are insufficient for the object identification task. The authors argued that this information would be helpful so people could submit superior qualities imagery, avoiding the generation of incorrect or insufficient descriptions [102].

12. **Descriptions not covering the images' intention:** According to Stangl, Morris, and Gurari [84], an image's intention is very relevant in the social media contexts, especially when the comments do not reference the visual content. Also, the purposes of images impact the image descriptions preferences of visually impaired people.

13. **Scanty descriptions of the people's actions:** according to Wu *et al.* [100], describing what people are doing in the image, *i.e.,* their actions, are considered the second most intriguing people's features, and 26% of the participants in their large-scale study suggested more descriptive details about people's actions. Also, people's actions allow contextualization of the visual content.

Table 3.3 summarizes the frequency of the image descriptions' issues discussed in this Section. The three most frequent problems refer to "Mediocre texts" and "Missing texts", cited by six of the eleven selected studies, and "Inaccurate texts", cited by four of the eleven studies. It is worth noting that there are intersections between the studies as they reported more than one issue in an image description. Regarding the less frequent issues with only one occurrence: "Non-descriptive texts", "Homogeneous texts for similar images",

“Erroneous texts for low quality images”, “Descriptions not covering the images’ intention”, and “Scanty descriptions of the people’s actions”.

Table 3.3: Frequency of the image descriptions’ issues identified in the snowballing review.

Image descriptions’ issues	Frequency
Mediocre texts	6
Missing texts	6
Inaccurate texts	4
Descriptions not covering embedded texts	3
Confused texts	2
Descriptions not covering humor and people’s emotions	2
Scanty descriptions of the people’s characteristics	2
Scanty descriptions of the people’s facial and body expressions	2
Non-descriptive texts	1
Homogeneous texts for similar images	1
Erroneous texts for low quality images	1
Descriptions not covering the images’ intention	1
Scanty descriptions of the people’s actions	1

3.3.2 Inhibitor Factors

In addition to the issues in image descriptions, the snowballing identified contributing factors to the inhibition of manual image descriptions, as we present them below:

1. **Unknowing how to describe an image:** authors of manual image descriptions should follow guidelines recommendations; however, sighted people have difficulty describing imagery contents because they do not know how to elaborate the sentences [76, 31];

2. **Forgetfulness:** sighted people claim to forget to include imagery descriptions [76, 31]. According to Gleason *et al.* [31], sighted people forget the existence and the purpose of image descriptions. As Sacramento *et al.* [76] identified, forgetfulness was the third most frequent reason reported by sighted participants, even for those used to create image descriptions on social media;

3. **Lack of time:** despite its importance, sighted people claim not to have time to describe an image into words, as it requires time and energy [76, 31];

4. **Interface issues:** the features to add image descriptions on social media require manual configuration and are not widely publicized to their users [76, 31]. Thus, it takes several steps to enable it, and it is not easy to find in the account settings [31, 76]. Sacramento *et*

al. [76] reported that nearly 50% of the survey participants said they did not add image descriptions because they did not know how to add them into the posts. Also, the interfaces do not allow adding them to images in non-original posts [31, 76]. According to Sacramento *et al.* [76], the interfaces lack instructions since only the interfaces of Facebook and Instagram contain guidelines on how to add an image description;

5. **Ignore the image descriptions' relevance:** sighted people are unaware of the social importance of the imagery descriptions and how essential they are to ensure inclusion for visually impaired people [76]. The lack of interest of sighted people in developing accessible content highlight the need for actions to raise people's awareness [76].

Table 3.4 summarizes the reasons why sighted people do not provide their descriptions for the imagery content.

Table 3.4: Inhibitor factors of the manual image descriptions generation by sighted people.

Inhibitor factors	Frequency
Unknowing how to describe an image	2
Forgetfulness	2
Lack of time	2
Interface issues	2
Ignore the image descriptions' relevance	1

4. INTERVIEW STUDY

As we identified in the Snowballing, one of the two most frequent problems is insufficient image descriptions. In the interview study, we explored this "insufficiency", and we sought to collect characteristics of satisfactory and unsatisfactory image descriptions, as well as participants' expectations, to identify how to describe an image for people with visual impairment to understand its content. This Chapter has the following structure:

Section 4.1 explains our data collection method, and Section 4.2 expounds our data analysis approach. Section 4.3 presents the participants' demographic data, and Section 4.4 presents the interview results. Lastly, Section 4.5 discusses the convergence between the interviews and Snowballing data.

4.1 Data collection

As outlined in Chapter 1, our data collection method was semi-structured interviews. Each interview had last about 2 hours by the Zoom meeting tool. We read the Free and Clarified Consent Term (FCCT), presented in Appendix A, using the online Microsoft Word's text-to-speech feature, which data collection occurred after participants cleared questions and their Term acceptance. In total, eight people with low vision impairment had agreed to participate in our study.

The interview script is available in Appendix B. We started by asking participants' demographic data, *e.g.,* name, age, profession, diagnosis of visual condition, how long they live with visual impairment, visual acuity, and assistive accessories used to navigate. Then, we asked nine open questions: first, for an ice-breaking purpose, participants introduced themselves. Second, participants answered their familiarity with assistive technologies. The ensued four questions were to understand the participants' relationship with descriptive sentences, either textual and audio format. These questions were essential to understanding participants' previous experiences with image descriptions. Finally, we aimed to evaluate, individually, the images and their respective descriptions (automatic and human) through three questions.

The evaluations were delineated as follows: first, participants visualized the target image and answered what they thought was relevant in it, *i.e.,* what caught their attention. It aimed to engage participants' attention, identify their visual acuity, and grow their image descriptions' expectations. In cases of (very) low visual acuity, we offered automatic labels from the MS COCO dataset [14] that identified the main image elements. In these cases, we preferred not to collect this answer since they could not distinguish the image elements.

Second, we reproduced one of the image descriptions. We did not inform participants of the description type (automatic or human), as we did not want to influence their opinions. Furthermore, we randomly order the description type, *i.e.,* sometimes the first sentence was an automatic description, and sometimes it was a human description. After the first sentence, participants answered either it was or not satisfactory for their image understanding and the reason for their opinion, thus exploring its aspects. We then reproduced the second sentence and repeated the question. Finally, participants answered their expectations based on the previous image descriptions. In this way, we collected what is imperative in image descriptions to meet participants' expectations.

4.2 Data analysis

We used the **qualitative content analysis** method to analyze the interviews, following the procedures proposed by [25, 91]. The content analysis aims to identify and analyze the texts' characteristics [28], providing knowledge and understanding of the phenomenon under study [23]. The outcome is compressed text segments into content categories and codes [25, 86]. Thus, we defined this method to analyze the collected data, applying the **inductive coding** approach once the categories and respective codes are from the interviews data [25].

According to Thomas [91], inductive coding aims to assign meaningful text groups into categories to which they are relevant. Each category contains one or more codes that represent the essence of the textual data [77]. Figure 4.1 summarizes the procedures, and the content analysis process used is delineated as follows:

1. Preparation phase:
   (a) Data cleaning: it aimed to transcribe the recordings of the interviews;
   (b) Initial reading: it aimed to become immersed in the data through several readings.
2. Organizing phase:
   (a) Open coding: it aimed to create general notes from the data and possible categories;
   (b) Grouping: it aimed to arrange similar codes by grouping them into higher-order categories;
   (c) Categorization: it aimed to diminish the number of higher-order categories by deciding whether or not codes belong to the same category;
   (d) Abstraction: it aimed to name the categories using content-characteristic words.
3. Resulting phase:

(a) Categories map: it aimed to create an overview of the categories that originated from the data.

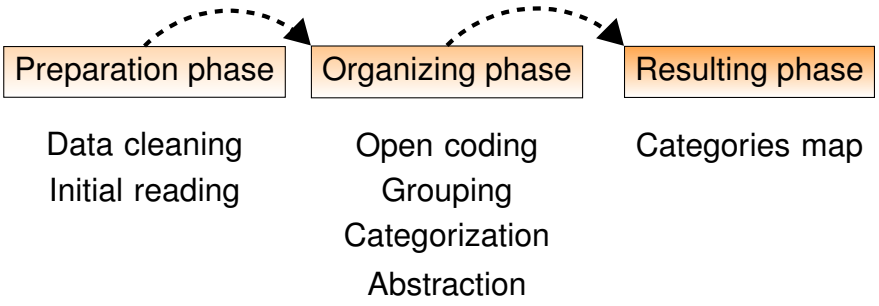


Figure 4.1: Qualitative Content Analysis Process.
Source: from the authors, based on [25].

Furthermore, we quantified each code to identify satisfactory and unsatisfactory image descriptions' characteristics. During the interviews, participants did not rank each sentence since we aimed to deepen understanding of their reasons than to elect the best image description in their opinion. We intended by quantifying the codes to point out high-priority aspects that may inspire IA researchers and sighted people to improve image descriptions quality. For example, by noting the frequent dissatisfaction reasons in automatic sentences, AI researchers could put effort into such features, also sighted people could pay attention to such characteristics to write more fitted sentences.

Regarding the trustworthiness or reliability of the analysis process, it must be clear to readers the coder's decision-making process, and the achieved results [25]. Therefore, for each defined code, we strongly support it by authoritative citations of the participants so the reader can understand its origin. Elo and Kyngäs [25] state that appendixes and tables can demonstrate links between the data and the results. Hence, Appendixes E, and F present in detail our analysis.

As Stemler [86] defends, when the coder gets the same results after repeating the coding and analysis, it indicates reliability in the process. Therefore, to ensure stability, the entire coding process was repeated three times by the study's principal researcher, using the QDA Miner Lite Tool. Because each researcher has a unique perspective of the collected data [86, 25], the study's principal researcher and their advisors had an agreement about the process.

Figure 4.2 presents a visual map of the content analysis process, showing the categories and the respective codes that lead the interviews' results.

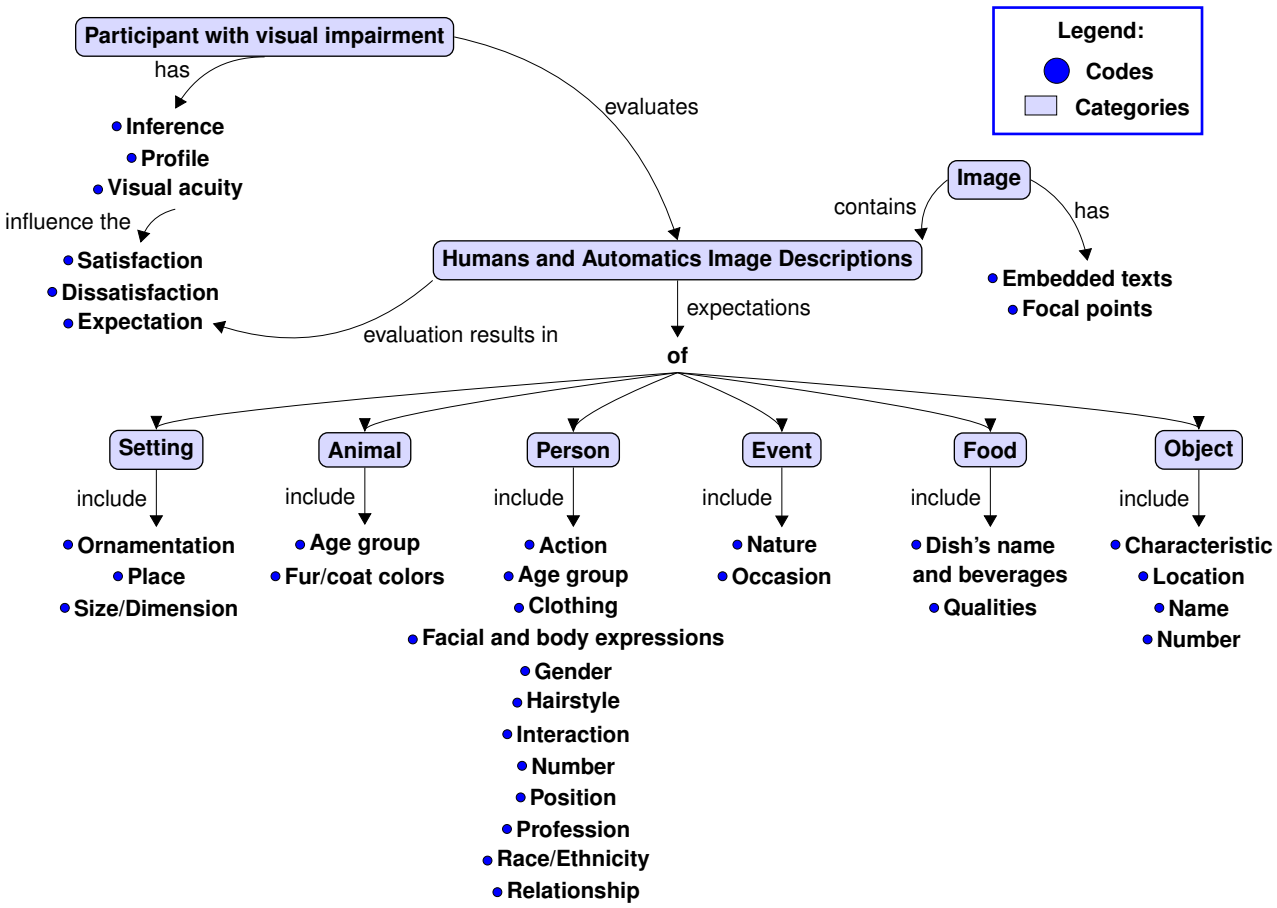


Figure 4.2: Codes and categories generated in the interview analysis process. Each small blue ball represents a code identified in the analysis, whereas the blue rectangles are their respective assigned categories. Source: from the authors.

4.3 Participants' Demographic data

Table 4.1 presents the demographic data of the eight volunteer participants in the interview study. The participants' average age is 43 years old, with the youngest person being 24 years old and the oldest being 56 years old. Regarding the professional activity, one participant (P4) works in an IT-related field, three participants (P2, P3, and P6) work as Public Servant, one participant (P5) is a clinical psychologist and also works professionally as an audio descriptor, one participant (P8) is a journalist and is currently studying to become a professional audio descriptor, and one participant (P1) works in the design field.

Concerning participants' visual acuity, all participants reported seeing contrasts and shapes, seven reported seeing colors, and one participant (P4) is colorblind. All participants have medical diagnoses of visual impairments, and all have lived with visual impairment for at least seven years. Concerning assistive technologies and devices, whereas six of the eight participants use a screen reader on personal computers, participants P2 and P7 reported only using it on mobile devices. Whereas five use a white cane, participants P2

and P3 prefer not to use it despite sharing its use would help them, and one participant (P8) uses a guide dog to navigate.

Table 4.1: Self-Reported Participant Demographics Data.

Participant	Age	Occupation	Diagnosis (Low vision causes)	Visual Acuity	Tech and devices
P1	50	Product designer	Glaucoma, since 2013. Monocular vision (sight in right eye)	Contrasts, shapes, shadows, and colors	Screen Reader, font increase, TalkBack, screen magnifier, and white cane
P2	52	Public Servant(Water Treatment Operator)	Undefined (5% vision in the left eye, and 4.5% vision in the right eye)	Contrasts, shapes, shadows, and colors	Screen Reader, TalkBack, and screen magnifier
P3	56	Public Servant	Glaucoma, since 2001. Monocular vision (sight in left eye)	Contrasts, shapes, shadows, and colors	Screen Reader, font increase, and TalkBack. She does not use white cane
P4	44	Computer Instructor for People with visual impairments	Diabetic retinopathy, since 2005	Strong contrasts, shapes, and shadows. Black and white vision (colorblind)	Screen Reader, voice recorder, TalkBack, and white cane
P5	41	Clinical psychologist and Professional audio describer	Pigmentary Retinitis. Significant sight loss since 2013	Contrasts, larger shapes, shadows, and colors.	Screen Reader, BeMyEyes, Lazarillo - Accessible GPS, CacheReader, VoiceOver, and white cane
P6	39	Public Servant (Admin Assistant)	Pigmentary Retinitis. Significant sight loss since 2011	Contrasts, some shapes, shadows, and colors (dark and light)	Screen Reader, CacheReader, VoiceOver, app for color, and white cane
P7	41	Self-employed	Glaucoma, since 1987	Contrasts, shapes, shadows, and colors	Screen magnifier and white cane
P8	24	Jornalist	Leber Congenital Amaurosis (15% residual vision)	Shapes, colors, and contrasts	Screen Reader, VoiceOver, and screen magnifier. P8 uses Guide dog to navigate

4.4 Interviews results

This section presents the interview results and has the following structure: Subsection 4.4.1 reveals what images' components caught participants' attention. Subsection 4.4.2 shows satisfaction and dissatisfaction causes in automatic image descriptions, and Subsection 4.4.3 presents the referred motives in human image descriptions. Finally, Subsection 4.4.4 reveals the participants' expectations regarding image descriptions, *i.e.,* what participants desire in a descriptive sentence.

4.4.1 Relevant images' elements

This subsection presents participants' answers regarding what caught their attention in images (see Question 7, in Appendix B). This question aimed to examine participants' image perceptions. In cases of (very) low visual acuity, participants asked for automatic tags that identified the main images' elements. We intended for participants to observe the target image, its details, and its relevant components, thus arousing their initial expectations. As we

expected, the participants' visual acuity varied and, due to the different colors and lighting of the images, they could not always identify or distinguish the images' elements. In these cases, the automatic tags aimed to provide them a minimal image conception. Therefore, this question was optional since participants could not answer it in (very) low visual acuity.

Of the 120 assessments, we collected 83 responses for this question, representing 69.16% of the evaluations. Figure 4.3 presents a word cloud with participants' answers regards to what they found relevant in images. The image shows that people, setting, person, table, and many were the most eye-catching images' attributes.

Figure 4.3: Word Cloud regarding relevant images' elements according to the participants' visual acuity. Source: from the authors.

4.4.2 Satisfaction and Dissatisfaction in Automatic Image Descriptions

We explored the limitations of automatic image descriptions for visually impaired people, and we identified unsatisfactory e satisfactory characteristics (see Question 8, in Appendix B). The most frequent **dissatisfaction** reason reported refers to **setting's place** description, present in 19 of the 25 image descriptions evaluated (76%). We identified that the setting's ornamentation assists participants to imagine its place, *i.e.,* where the scene happened:

- P1 — *"I do not know if they are on the street or at home. It is not satisfactory."*;
- P3 — *"The description stated it is a living room, but how are they? Are they sitting and talking, or are they waiting? In which living room? Is it a meeting, a formality? Saying a 'living room' is too generic."*.

Insufficient description of the **people's actions** was the second most frequent **dissatisfaction** reason, present in 18 of the 25 image descriptions evaluated (72%):

- P5 — *"It is not satisfactory because it does not describe the people's actions around the table. Are they eating? Are they playing any games? It does not say how many people there are either."*.

The third most frequent **dissatisfaction** reason, present in 17 of the 25 descriptions evaluated (68%), refers to the **number of people** in the images:

- P2 — *"When the description says a group of people, it just means that there are more than one or two. But it is not saying if there are three or twenty people in the image."*;
- P6 — *"It is not satisfactory because I am not seeing a group of people. There are just three people in the image. To be a group, it would have to be more."*;
- P8 — *"When the description says a group of people, I do not understand how many there are. It could be four or six people..."*.

Furthermore, **generic image descriptions**, *i.e.,* that do not describe the image in detail, was one of the main **dissatisfaction** reported:

- P2 — *"In my case, that I'm a low vision person and have some visual perception, the image description helps me. But assuming I only rely on it to understand the image, I would not say it is satisfactory since it does not describe anything. It does not contextualize or describes the image's setting. It does not give more detail about what people are doing."*;

- P5 — *“Partially satisfactory, this image description is too generic. It informs me of the scene, but it does not describe the image’s setting, neither how many people there are. It is a very brief and general description.”*;
- P8 — *“Not satisfactory as it does not answer what I want to know. I see a man, but where is the cat? Where is the eye-catching thing in the photo since the man is in the background? It is an over generic description.”*.

Participants reported dissatisfaction with **people’s facial and body expressions**’ descriptions, as well as their **clothing**:

- P3 — *“It is not satisfactory because it does not look like he is sitting on the couch. I need a little more description to know how he is, whether he seems relaxed at home with the dog or at someone’s house and the dog happens to be there.”*; and *“It’s not satisfactory because there is no detail. What are they doing, how are they dressed?”*;
- P5 —*“It is not satisfactory. In this case, the social clothing caught my attention, so it would be relevant to characterize these people, say how many people there are. Also, it needs to describe that they are adults drinking wine.”*;
- P6 —*“It would be better if saying it is an office or a library. Also, it should describe the person in the background. I think the person’s clothing is red.”*.

Participants expressed **dissatisfaction** about the **interaction** of the images’ elements (people, objects, and animals) and about what was happening (**event**) in the image:

- P2 — *“Not satisfactory. I am in doubt whether this person in the corner is interacting with the group or is there to serve, it is unclear.”*;
- P6 — *“It’s not satisfactory because the [image] description would have to express more about what’s going on, I can see they’re sitting at a table with laptops. It would have to be more descriptive about the image for me to understand better what is happening.”*.

We identified that participants created **scenarios** or **deductions** from insufficient descriptions and from their images **inferences**:

- P1 — *“It is satisfying as it says about the cake, so I imagine a birthday. It does not inform about the people’s faces, but it describes a characteristic that makes you understand what is happening in the image. If there is a cake, then I deduce that they are celebrating a birthday.”*;
- P7 — *“From this image description, you can imagine the context. Maybe they are executives, maybe employees of a company in a meeting room, so I think it is satisfactory.”*.

Regarding the frequent **satisfactory** characteristics, refer to the **people's position** in the images,*i.e.,* describing whether they are standing or sitting, the **objects' identification**, and the **people's gender**:

- P4 — *"It is satisfying because it brought me the details as, for example, the laptop and he is using the laptop."*;
- P6 — *"It is satisfying because people are sitting at desks working."*;
- P7 — *"It is satisfactory because you can understand the scene since people are sitting on the couch in a living room."*.

We identified that the **satisfaction and dissatisfaction** of the descriptions are related to the **participants' visual acuity**, *i.e.,* with what the participants were able to see in the images. Participants strongly criticized image descriptions that did not conform to their visual acuity, even though the image descriptions used in the interviews were not incorrect:

- P1 — *"It is pretty succinct, but I can imagine it. It is not what I was thinking of the image as it looked like something else to me I would like more details of the place so I can understand it better."*;
- P4 — *"It is not satisfactory as my view is completely different, it didn't help me at all. (...) I imagined something different."*;
- P6 — *"It is satisfactory because it is just saying what I thought I was seeing, but it lacks details."*.

4.4.3 Satisfaction and Dissatisfaction in Human Image Descriptions

Regarding the reasons for dissatisfaction and satisfaction of the manually generated image descriptions, we identified that the most reason for **dissatisfaction**, present in 17 of the 25 image descriptions evaluated (68%), refers to the insufficient **setting's ornamentation** description. According to the participants' opinion, the human image descriptions lacked details that would characterize the place, such as furniture pieces, accessories, and decorative objects:

- P2 — *"This image description is better as it described a long table, but it lacked about the place's characteristics."*;
- P5 — *"I missed a little about the characteristics of this place called 'a restaurant' since it can be a party hall and not a restaurant."*;

- P8 — *“Saying ‘an office’ gives me a lot of information. I can understand it, but I would add a few things about how this place looks like.”*.

The second most frequent **dissatisfaction** reason, present in 14 of the 25 images descriptions evaluated (56%), refers to the **setting’s place** identification:

- P3 — *“It’s not satisfactory, he is sitting on the couch, and there is someone behind him but is he at home? Where is this?”*;
- P7 — *“It is very vague, it does not relate to the computer, you can not know what context he is in if he is working or not. Although there is a desk with papers seeming to be a working environment, there is a cat. So this image description is very unclear.”*;
- P8 — *“Partially satisfactory because the image description says to me that people are working on multiple tables. But, for example, I have a question whether they are in an office or a classroom. Are they working individually or in a group sharing something? Are they in a coffee or in a co-working place where people are working in the same space?”*.

While most participants expressed satisfaction when the image description referred to the place as “a library”, or “an office”, one participant (P5) reported preferring the place/room details instead of the image description stating it directly, except in the cases where the place was explicit in the image:

- P5 — *“Well, I would prefer the place characteristics. Why does a sighted person conclude that it is a library unless it says ‘library’ somewhere? Otherwise, how do you reach that conclusion?”*, and *“What characterizes this place as an office? In a little while, it could be an Internet Cafe, with several people on computers in the same room, or a computer lab in a university.”*.

The third most frequent **dissatisfaction** reason in human image descriptions refers to the **number of people**, reported in 13 of the 25 image descriptions evaluated (52%):

- P3 — *“How many people are there?”*;
- P5 — *“I would not use the ‘a group of people’ expression unless it is many people. When it is less than ten people, I would always say a number.”*.

Moreover, participants were **dissatisfied** with the description of **people’s gender** and **people’s age**. Particularly, participant P5 drew attention to the generic words use to connote age group:

- P2 — *“I think it is important to describe that there are four people, two adults, and two children, it seems like two children, for me.”*, and *“I think that describing as ‘several people’ is unnecessary, there are not so many people in the image.”*;

- P4 — *“It does not inform the number of people, for example, eight women, eight men, or four women and four men, and so on”*, and *“I want to know how many people there are, how many boys, how many girls.”*;
- P5 — *“The image description uses expressions like ‘older people’, and I do not know if it means that they are adults or elderly.”*;
- P8 — *“What is lacking is how many people, adults, children, babies are in the kitchen or a dining room”*, and *“If they are many or few people if they are young, if they are all men or women.”*.

Some participants brought **dissatisfaction** with the **objects’ location**:

- P1 — *“It’s not satisfactory (...) it could not bring me a spatial notion, I do not know where they are, if they are in the middle of a room, I do not know where this sofa is. For me, it was too vague.”*;
- P2 — *“(...) This ‘surrounded by papers’ is also not nice it lacks to describe where the papers are because they could be on the desk and they are not”*.

As reported in the automatic image descriptions, participants expressed **dissatisfaction** with **generic image descriptions**:

- P3 — *“More or less satisfactory because actually, I don’t know how their position, I don’t know if they’re close together. Some details do not seem to be important, but for me, they are. It’s too generic for me. How are they? Are the tables next to each other? Is it a laboratory? How many people are there?”*

Some of the human image descriptions expressed **people’s relationships**, such as “friends” or “family”. Participants expressed **preferring** the use of these respective terms, for example:

- P7 — *“I find it satisfactory because it gives the context of a family celebrating around a cake, so you imagine a birthday.”*.

Participant P6 expressed discontent when in the absence of these terms:

- P6 — *“I don’t think it’s a group, it looks more like a family than a group.”*.

In contrast, other participants expressed that these words are **generic**:

- P2 — *“The description says it is a family, but ‘family’ does not tell me much.”*;
- P8 — *“A family can be an elderly person and an adult, so it’s very unspecific.”*.

Participants also reported **dissatisfaction** with the use of **generic words** to express **people's actions**, such as "celebrating" or "giving a present/gift":

- P1 — *"It could be an anniversary, it could be an award, anything like that."*;
- P2 — *"It's not satisfactory. Okay, the description said me what people are doing, but saying 'celebrating' is very vague because there are so many things that celebration can mean."*;
- P4 — *"The image description should describe to me whether it is a birthday gift or it was delivered after an event, celebration, solemnity or something like that."*;
- P5 — *"Why are [the image descriptions] saying that people are working and studying? What is the difference? Are people working because they are writing something in the notebook? Are people texting on their cell phones or working on the computer? the image description needs to describe to me what they are doing."*

The word "electronic" was also a **dissatisfaction** reason:

- P5 — *"This image description gives rise to many possible interpretations, there are several situations in which you can work with electronics. It could be people working digitally on tablets, cell phones, and computers, or it could be people who are Electronics Technicians and who are working on electronic boards, like in an electronic class where they are repairing a cellphone or computer board."*;
- P8 — *"Not satisfactory because it is too generic. A lot of the other images you showed me had a group of people with electronics. What electronics are these? It could be a robot!"*.

We also identified **dissatisfaction** with the expression "open" to characterize the **setting's dimension**:

- P5 — *"This 'open office' at the end was weird. It could be an airy room or the room could have an open window..."*.

Similarly, participant P3 reported confusion regarding this term's use in image description:

- P3 — *"But what time is it? Is this room closed? Because [the image description] is saying that [the room] is open, but I do not see lights in the environment."*.

Participant P5 addressed, once more, personal interpretations in image descriptions.

- P5 — *“This sentence has a problem that I consider very serious in image descriptions. It says that people are friends, but those who see the picture has no way of knowing if they are friends or not. It is a personal interpretation.”*, and *“The image descriptions need to be linked to the principle of autonomy, of ‘not interpreting for me’, I want the information so I can conclude what it is. Of course, there is always a limit the idea is never to play around with the person with visual impairment because this is very irritating.”*.

Moreover, participant P1 addressed the generation of these expressions in automatic image description models, despite not having a background in the technology area:

- P1 — *“It gave an extra feature by saying ‘a family’ (...) But this is interesting, how could a machine interpret that they are a family? What would be the elements that the machine would consider?”*.

Regarding the **satisfaction reasons**, the description of **people’s actions** was the most frequent reason, followed by the description of **people’s gender** and the **people’s position** in the image:

- P3 — *“I consider satisfactory, now I know that people are in the kitchen drinking coffee and that someone is using the laptop.”*;

- P7 — *“It is satisfactory because it describes that she is watching television and sitting next to two men so you can understand the context, that they are in a living room.”*.

Once more we identified that **visual acuity** is related with **satisfaction and dissatisfaction** of image descriptions:

- P2 — *“It is satisfactory, I had not seen the open laptop on the coffee table.”*;

- P4 — *“The image description brought a detail that I had not noticed, I can imagine that they are playing a game. I had only noticed the table.”*;

- P6 — *“It is not satisfactory, it did not help me much. At first, I thought it was a law firm I never imagined it was an Indian classroom.”*.

Figure 4.4 summarizes the reasons for dissatisfaction and satisfaction reported by the participants in automatic and human images descriptions.

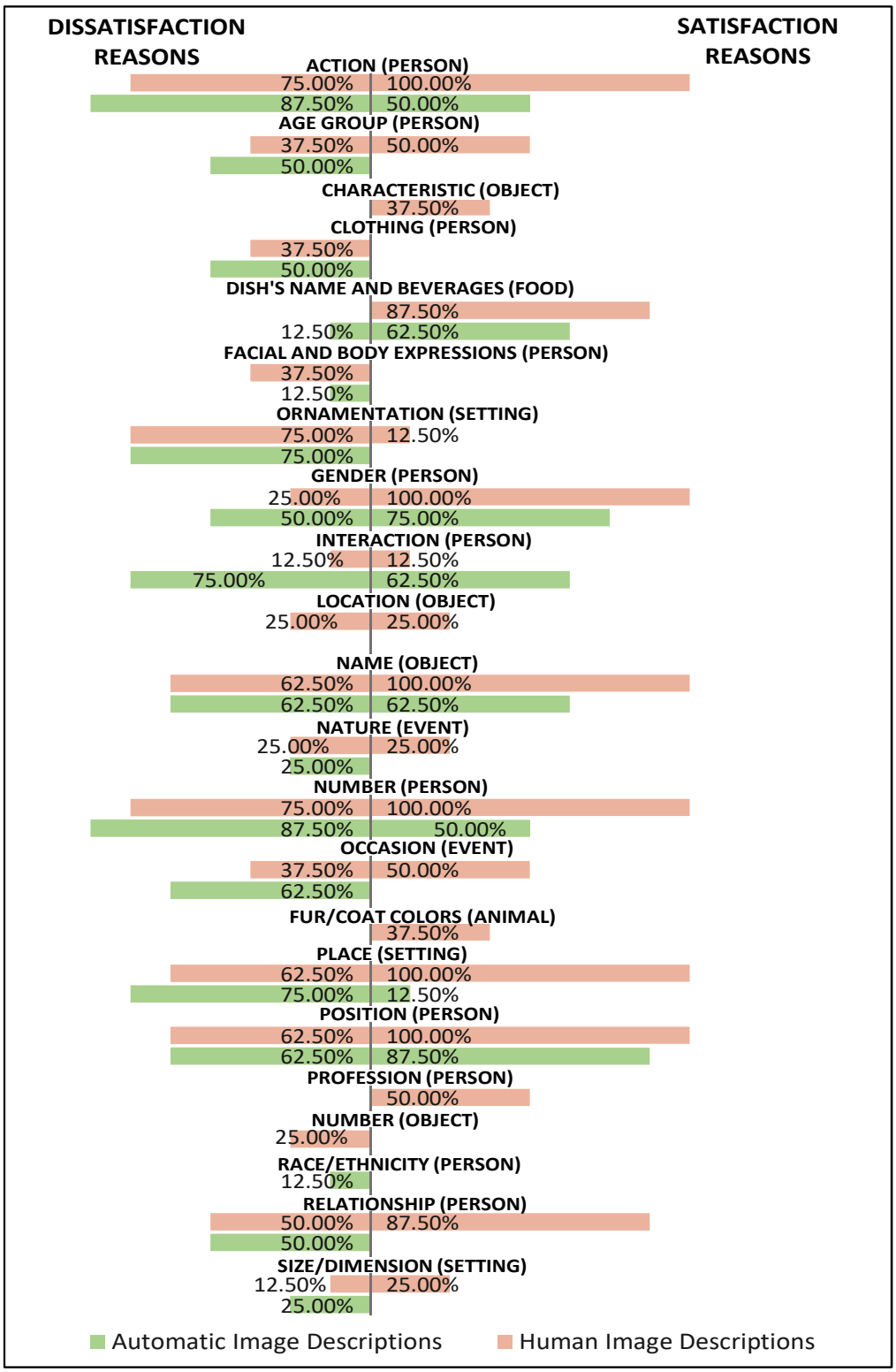


Figure 4.4: Satisfaction and dissatisfaction reasons collected sorted in alphabetical order. The total percentage (100%) means the eight participants of the study. The figure shows that the highest satisfaction reason in the human image descriptions was descriptions of the people's actions, whereas the highest dissatisfaction cause was descriptions of the settings' ornamentation. Regards the automatic image descriptions, the highest satisfaction reason refers to the people's positions, whereas the dissatisfaction reason was descriptions of the settings' places.

4.4.4 Image Descriptions' Expectations

At the end of the human and automatic image descriptions evaluation, we asked participants their image descriptions expectations, *i.e.,* what they would like a descriptive sentence to inform them about the image. The percentages presented contain intersections between codes, as one expectation contains one or more codes.

Participants' expectations regarding the **setting's ornamentation** represents 51.67% of the expectations reported. This expectation was the only one announced in all images by at least one participant, meaning that none of both evaluated image descriptions, human and automatic, described the environment's ornamentation in a way that met the participants' expectations. Participants highlighted its relevance in image descriptions:

- P2 — *"(...) Let's say the room has a old decor, something unusual about this room."*;
- P5 — *"Details concerning the room's decoration, furniture, how the environment is...I missed these things in the image descriptions."*;
- P8 — *"I see there is decoration, shelves above the man, so there are many things that are in my expectation. I have the impression that there is a mess in this environment, and it changes a lot how we look at the picture."*.

Participant P1 brought up that **setting's ornamentation** description helps people with visual impairments to understand what is happening **(event)** in the image:

- P1 — *"If the image description gives more details and information about the environment, [for example], 'There are balloons attached to the ceiling', I can understand it is someone's birthday. As much detail as possible about the environment will allow a better understanding of it. What matters for a person with visual impairments is to be included by understanding the environment's appearance and where this happened."*.

Image descriptions about the setting's ornamentation include, for example, describing the **objects' names**, and represents 20.00% of the participants' expectations:

- P5 — *"It would like a description about the environment, the sofa, the bag, even if it was something slight. A description about the characteristics and the objects of this environment."*;
- P6 — *"[I expect a]Description about all these objects behind them, the sofa back there, the cabinet, the bookcase, back [of the room] there are some decorative objects (...). More details for a person to understand where it is."*.

Another participants' expectation refers to the **objects' characteristics** (10.83%). We identified it includes describing the objects' colors, shapes, and sizes:

- P3 — *“I would like to know more details about the sofa, to know if it is a formal sofa or not...”*;
- P6 — *“There is another sofa with a red toss pillow, it would be interesting to say that in the image description...”*;
- P7 — *“A description that described me (...) the gift’s size if it is a small box. I think it would give a little more context to the setting”*, and *“It is important to say whether they are on a rounded or a squared table. It is a short piece of information, but it helps to give context for the image.”*;
- P8 — *“What stands out for me in the picture is something very colorful, the table looks yellow or orange, the wall seems to be red (...) [I expect] Describing the colors, it is color photography so the naming the colors is important.”*.

Also, participants reported expectations regarding the **objects’ location** (8.33%). We identified these expectations include informing the objects’ location in the image *e.g.,* [object] in the background, top, left side, etc., and the objects’ location in the setting relative to others image elements *e.g.,* [object] above [another element], between, etc., as expressed by:

- P1- *“I expect the description says me more about how the organization of the space, where this sofa is in the room.”*;
- P5 - *“The description must inform about the sofa, the chairs, the backpack, that the remote is between the person and the TV...”*;
- P6 - *“It should describe that there is an open laptop on the desk, there is a paper pad with a pen, all of it would have to be in the image description.”*, and *“The image description should say that there is an easel board in the back of the room...”*.

Participant P1 highlighted the importance for people with visual impairments to understand the setting’s physical space and the objects’ location:

- P1 — *“I want more [description] about the setting’s space (...), if the person is next to a window, for example. (...) For those with visual impairments is very important to understand the room’s physical space since we are used to orienting ourselves in an unfamiliar environment by its objects, to navigate or to identify objects and obstacles, so it is essential to understand the physical elements.”*.

We also identified expectations about the **number of objects**, representing 4.17% of the participants’ expectations:

- P4 — *“I want to know how many objects, how many computers, how many chairs, how many tables...”*;

- P7 — *“I want the description to inform me if all people have laptops or if it is just one laptop, such as in a solo presentation.”*;
- P8 — *“I want to know if it is an office of a bank or an advertising agency, for example, [because] there is a lot of difference. Also, approximately, the number of desks in this office room.”*.

Descriptions about the **setting’s place** (where) represent 27.50% of the expectations:

- P2 — *“These people are in a workplace, using computers. When I say ‘apparently’ it means that I have questions and I would like the image description to enlighten me.”*;
- P3 — *“I would like the image description to inform me more details, what is around, where it happened. For me, it is a room, but I do not define whether they are in an office, in a house, or a leisure area.”*.

Participants reported expectations regarding the **setting’s dimension** (7.50%):

- P3 — *“I would like to know more about the size or dimension of this room, whether it is an opened or closed room.”*;
- P7 — *“A image description about the room, the environment you are in, the room’s appearance so I can have an idea if it is a meeting room if it is small or large”*, and *“I want to know the size of this room.”*;
- P8 — *“I want in the description whether it is a big place or a small one. I want detailed and objective image descriptions, it is not supposed to be a long description, but two or three more words to improve image understanding.”*.

Setting’s dimension description includes informing **structural elements** such as windows and pillars, as well as lighting:

- P5 — *“I would like a description about the ornamentation of this room, if it has windows...”*;
- P6 — *“What I think is important and missing in the image description is about the pillar dividing the room...”*;
- P8 — *“I want to know If the room is well-lighted.”*.

Participants reported expectations regarding the **number of people** in the image (39.17%), and they noted it is relevant even in the cases where people appear partially in the image:

- P2 — *“It is important to know how many people there are, even this person who does not appear in the image, only appearing her/him plate or knees.”*;
- P8 — *“I would like the description says me whether she is alone in the picture. I do not know, for example, if she is doing a presentation on the laptop, because she could be in a video conference, or if she is in a room where you expected to have a lot of people there, but the framing shows only a woman in the scene.”*.

Descriptions of **people’s clothing** represents 26.67% of the participants’ expectations:

- P6 — *“The image description has to say to me about the color of the person’s clothes so I can have this extra detail.”*;
- P8 — *“It needs to describe whether people are dressed too formally or not, or if it looks like something more humble. I would add information that gives me information about social class, the context in which these people are inserted.”*.

We identified that **people’s clothing** descriptions help people with low vision to understand **people’s profession**:

- P8 — *“It is always important to say people’s gender and their clothing. [For example] people [in the image] could be doctors or students.”*.

We identified that **people’s clothing** descriptions help people with low vision to understand the **event’s occasion** and the **event’s nature**:

- P3 — *“Is it a formal dinner? Is it Christmas? Is it a birthday or something like that? How are they dressed? Are they well-dressed or not? The waiter is serving wine. Is he a relative of someone, or is he wearing a uniform?”*;
- P7 — *“I want to know more or less their dressing...whether they are formally or informally. It does not need to go into too much detail, but to get an idea of the context, the scene, the event that is happening there.”, and “I expect if it is not a very large number of people, [to describe] their clothing, so I can know if they are wearing clothes for a party, a larger production, or if it is more an informal event.”*;
- P8 — *“It is a difference if the woman is wearing a pink suit or a casual white blouse, it is a big difference...”*.

Participants reported that descriptions of **people’s clothing** also help them to understand the **people’s interaction** with other image elements, such as people, animals, objects, etc.:

- P5 — *“The image description should describe me the clothing of these people, for example, whether they are in social clothing or sportswear, to understand what they could be doing on the laptop.”*;

- P7 — *“It is important to describe the outfit and the setting, so I can imagine what he might be doing on the laptop. For example, if he is wearing more formal clothes, it means he is working, or if it is more informal, then he is not working, it is entertainment activity.”*.

Descriptions of **people’s actions** represent 22.50% of the participants’ expectations. Participants highlighted the relevance of describing different actions that are happening in the image:

- P2 — *“I would expect to know what each person is doing because each one of them is doing something different”*;

- P5 — *“I want to know what they are doing if they are typing on the computer keyboard or reading books (...). One of the audio description’s principles is ‘Do not summarize and avoid wordiness’. You need to be objective when describing details. You must be able to think about what is relevant in an image and do not extend yourself. But this is not synonymous with summarizing, because when we do it, we cut details, and this is usually a problem.”*.

Descriptions concerning **people’s interaction** represents 14.17% of the participants’ expectations. We identified this characteristic to help people with visual impairment to understand what people are doing, *i.e.,* the actions in images:

- P2 — *“The image description should inform me more about the man, what he is doing on the laptop. It should describe the coffee table because he put his feet on it. I do not have this habit. But, when someone’s feet are on the coffee table, it seems relaxed. I do not know if he is watching television because the TV does not show up in the image. I don’t know if he is talking to someone.”*;

- P4 — *“I want to know what they are doing on the laptop if they are reading, doing researches, or watching something...”*;

- P8 — *“I want to understand what is happening there, if he is petting the cat or if the cat was in the way and the man was taking the cat away.”*.

Descriptions of **people’s positions** in images represent 15.00% of participants’ expectations, including the descriptions’ expectations of different people’s positions in the same image:

- P1 — *“The description needs to say about the person’s appearance, or highlight something specific such as that person standing. The image descriptions did not say that someone is standing, [this is important] because maybe that standing person is the one being honored.”*;
- P3 — *“I want to know how he is on the couch as I can not see him on the sofa. It looks like he is sitting, but I do not know if he is sitting on a sofa with crossed legs or not, or if he is more like lying on the couch.”*;
- P5 — *“I expect to know how they are sitting, describing that they are sitting in different sofas or sitting in pairs.”*.

Similar to what participants reported in the objects’ location expectations, we identified expectations regarding the **people’s position** relative to other image elements:

- P1 — *“The ideal is a physical description of the setting, [for example] ‘people sitting at different tables, one after another’, and what is in front of them...”*;
- P6 — *“(...) There are several people, and a woman is standing. It would be interesting these people’s clothing descriptions. The image description has to say that there are children at the table, and I think there is a child in the room’s back.”*.

Other participants’ expectations about the people in images include **age group** (25.83% of the expectations), **gender** (21.67%), **race or ethnicity** (7.50%), and **hairstyle** (2.50%):

- P3 — *“Does this conference have more women than men? Because this is my universe, I work with women.”*;
- P5 — *“The ideal is to say (...) whether they are young or adult people. It would be much better if the description says people’s gender, whether they are men or women.”*, and *“I think the description must inform some characteristics of this woman, [for example] if she is a white woman, a black woman, her clothing if she is young or not, something like this.”*;
- P6 — *“It would have to describe their clothing, the colors of clothing. Also, their hairstyles, if their hair up into a ponytail or a down style, whether their hairs are short or long because this would be something that people who are blind or low vision would begin to imagine their outfit and their hairstyles. The image description would be more interesting if it had these two details, this extra thing to hold the person’s attention.”*;
- P8 — *“I expect a description gives me a bit of a sense of Social Class, I identify everyone as white and blonde, and information about people’s gender, ethnicity, and Social Class is important for me.”*.

Descriptions of **people's facial and body expressions** represent 16.67% of participants' expectations:

- P1 — *"When you are going to describe an image, if you could convey the person's countenance, describing whether it is a moment of joy, sadness, or worry, would be very important for visually impaired people to understand what is happening there."*;
- P8 — *"People's facial expressions is what makes me most curious about an image. For example, if it looks like they are happy, relaxed, or in a hurry."*, and *"I want to know people's countenance, whether it seems they are smiling or laughing at something, or if they are having a disagreement or talking amicably."*.

Participants reported expectations about the **people's professions** in the images:

- P1 — *"I want more qualities about the teacher, [for example] 'the teacher looks carefully at the students'..."*;
- P8 — *"I expect to know how many students there are, whether there are two or ten students because that makes a difference. Also, if there is a teacher there with the students."*

Descriptions of **event's occasion** represent 13.33% of the participants' expectations, whereas **events' nature** descriptions reflect 9.17% of the expectations. We considered "event's occasion" when participants mentioned expectations of, for example, birthdays, parties, meetings, Christmas, and general celebrations. Events' nature encompasses participants' expectations regarding the context of the event and actions presented in images, such as informal events (*e.g.,* social gatherings with friends), formal events (*e.g.,* meetings), and familiar events (*e.g.,* birthday parties):

- P2 — *"It is always important to contextualize the image and provide the main information about it. In the case of this image, (...) they are in a library eating pizza, but are they studying, or are they gathering there? Maybe something more descriptive about it."*;
- P3 — *"Is this a formal conference, or is it one of those courses that you go informally on weekends? Is it a study group?"*, and *"I would like to know if it is an anniversary, a gathering."*;
- P4 — *"What was the birthday cake be a celebration? Because sometimes the celebration could be a formal event, such as a corporate anniversary. Whose birthday is it?"*, and *"The image description should inform the reason for the meeting, and the number of people. For example, 'it is a business meeting' or 'it is a relaxed meeting', something like this. Their clothing already informs me it is a work-related thing, but it could be a moment after the work, like a happy hour with coworkers."*;

- P6 — *“(...) I think it is Christmas [because] I think there is a small tree on the table, so the image descriptions should have described this.”*.

Descriptions of the **food** represent 8.33% of the participants’ expectations, including the **dish’s name and beverages** and the **food’s qualities**:

- P3 — *“What is she drinking? Is she drinking a [alcoholic] drink, or is she drinking a coffee and she is talking to her friends?”*;
- P5 — *“what does this cake look like? Is it rectangular or circular? Is it a tall or a small cake?”*;
- P7 — *“I want table’s details, what food and drinks are laid on the table.”*;
- P8 — *“I want to know the dish names.”*.

Participants also reported that **image’s focal point**, *i.e.,* focus of the image, influences their image descriptions expectations:

- P1 — *“In my opinion, the description did not need to say the appearance of the man since he is alone in the background of the image. I want a little description about the surrounding objects, the setting appearance, if there are many objects...”*;
- P5 — *“Detailing the cake is certainly important because it is in the center of the image”*, and *“As it is only her in the scene, the description should say some characteristics of her.”*;
- P8 — *“I identify several objects and a man in the background of the image, so I expect to know what is in its foreground.”*.

Participants rely on the **objects’ location** and **people’s position** to understand the **focus of an image**:

- P1 — *“As the person is in the center of the image, the description would have to follow, [for example], ‘a person sitting in front of a laptop, holding a cup of coffee, there is a turned on TV in the background or an image [on TV screen]...’ so you have a complete image description.”*;
- P2 — *“I would expect the description to tell me the environment they are in and what the main characters are doing, these three people mentioned [in the image descriptions]. The description should say if more than two or three people are present in the image. It should say the focus is on these three people in the foreground. I want to understand what these three people are doing, what is going on with these two men over there.”*;

Regarding the descriptions of the **animals** present in images, the participants’ expectations were about their **age group** and their **fur**/**coat colors**:

- P5 — *“I want a description about the cat. If it is an adult cat, then it is enough to say ‘it is a cat’, but it could be a kitten. I think the fur color of the cat would be nice to describe too.”*.

Lastly, although the images selected for the study did not present **embedded texts**, participant P7 expect descriptions of the visible texts in images:

- P7 — *“If what she is presenting appears in the image as, for example, if it has a huge title, and it is important for the imagery contexts, then it is important to report it in the description.”*.

4.5 Convergence between Interviews and Snowballing

Although the satisfaction and dissatisfaction reasons identified are more punctual than those identified through Snowballing. We observed convergence such as mediocre descriptions of people’s physical appearance, race/ethnicity, and gender. We identified, in the Snowballing, that image descriptions of these personal characteristics is a matter of fairness for people who cannot use the sense of sight [84]. As reported in the interviews, people’s gender descriptions are the fourth highest participant’s expectations of the category “Person”, and four of the eight interviewed participants brought expectations about people’s race or ethnicity, in at least one of the images they evaluated. Images descriptions of people’s physical appearance is not a common practice, especially about people’s race or ethnicity, as P7 states:

- P7 — *“It is subjective to detail whether they are white or black people because we are used to not having any description of it. We ended up contending with poor descriptions, but ideally, yes, I would like to have gender details and more or less the physical characteristics of these people, of each one of them.”*.

Another convergence result refers to the people’s actions descriptions. In Snowballing, Wu *et al.* [100] reported that people’s action description was considered the second most intriguing feature by their participants with visual impairments. We identified that people’s actions descriptions were the second biggest cause of dissatisfaction in the automatic-generated sentences. Other convergences between Snowballing and the interview study include descriptions of people’s facial and bodily expressions. Also, texts that do not convey the image’s intention are similar to the occasion and nature of the event. Moreover, one of the problems in images descriptions identified during the Snowballing concerns embedded text in imagery identifying this expectation even though the selected images did not contain explicit texts.

As pointed out in Snowballing, insufficient images descriptions are generic sentences that, although existing, do not provide enough information or detail to contextualize images. We identified in both types of image descriptions participants' speeches over generic descriptions. Analogously, participants pointed out that the evaluated descriptions were very similar and mostly homogeneous:

- P6 — *"If this is a library, should not the description have mentioned books? I do not see books in the image, so that is why it did not occur to me that this place could be a library. The main characteristic of a library, which are books, was not in neither image descriptions."*;
- P8 — *"This description fits for three other images you have shown me. Saying 'A bunch of people, tables, and laptops' does not describe me almost anything. It does not tell me if it is a familiar or a professional setting if the place is messy or organized. I see a coffee mug, and the description is not giving me this information."*.

Participants reported dissatisfaction to regard mediocre image descriptions and, similar to we found in the Snowballing, the lack of accessibility causes feelings of isolation and frustration *et al.* [100], as shared by P3:

- P3 — *"I need detailed descriptions because I see these details in the image (...) It is like everyone has sight because they will talk [describe] to those who do not have a vision [blind people], but I have low vision, I see something, and I can not forget that I can see! Sometimes, when I see detail and the image description does not inform me, it confuses me. (...) My low vision gives me more possibility to see [compared to blind people]. But, at the same time, I need to be cautious because it may be an erroneous vision, or the description does not meet what I am seeing. It makes me sad, the description is not the way I want, and I can see that it can be better!"*.

Furthermore, in Snowballing, we identified that the unknown relevance of image descriptions is one of the factors that inhibit the human image descriptions generation [76]. Similarly, in the interviews, participants P2 and P6 highlighted that even the people close to those with visual impairments, such as family members and friends, do not include images descriptions in their personal social media posts:

- P2 — *"In [Facebook and WhatsApp] groups I participate in, people do not collaborate. Most people still lack its perception, including family members and people with visual impairments, especially people with low vision. Sometimes they are participating in a group with people who have no vision, and even so, they do not understand the need to include image descriptions."*;
- P6 — *"Many people live with a person with visual impairment, and they do unknown the importance [of image descriptions] and publish a photo without describing it."*.

We identified that, in general, participants' dissatisfaction was more accentuated in the automatic image descriptions. Even though the objective of the interviews was not to point out the best kind of image descriptions (human or automatic), we cannot avoid discussing the limitation regarding the classification and relationship of objects in automatic generators. Even the manual (human) image descriptions used in the study, from the MSCOCO dataset [14], are limited to 80 object categories. To exemplify this limitation of automatic models, we share participant P7's speech about Facebook's automatic descriptions in photos of "Beaches". The participant shared the following at the end of the interview:

- P7 — *"When the Facebook's image description of a beach photo says you 'This image may contain an ocean' is the worst kind of description for me because it does not tell me anything at all."*.

Lastly, we identified that, just as mediocre image descriptions are unsatisfactory for participants image's understanding, wordy ones are also, since they are exhaustive:

- P5 — *"One thing you must be cautious about an image description is overloading the two extremes. Summarizing it too much is troublesome because it ends too generic, and filling in details is weary. Thus, describe what is important for that occasion and context."*;
- P7 — *"When the description mentioned 'a long table', we imagine a lot of people, so it would be better and more detailed saying how many people there are. Mainly, the more people, there is not a need to describe each one clothing and hairstyle because this would be a long and tiresome sentence, and we want a description that gives the context of what is happening there."*.

Based on the Snowballing results and the data collected in the interviews, we have identified a set of best practices for image descriptions as discussed in the next Chapter.

5. GOOD PRACTICES FOR WRITING IMAGE DESCRIPTIONS

This Chapter introduces and discusses best practices for writing image descriptions identified in the interview study and based on the problems found in the Snowballing. As formerly reported, the image descriptions manually generated are inhibited, among other reasons, due to unknowing the descriptive sentences' importance and their creation for people with visual impairments [76, 31]. Therefore, we aim to contribute to good practices to power more acceptable images descriptions for those with visual impairments. We want to emphasize that our results do not represent all people with visual impairments, and it is not our intention to generalize the collected opinions nor treat this audience as a homogeneous group. This Chapter is structured as follows: a discussion precedes each identified good practice, and Table 5.1 summarizes the study's contribution.

According to Frota [29] and Costa [17], we must abandon the belief that there is always an ideal, unique, and correct descriptive sentence. To produce suitable descriptions, we must foremost consider that either sighted and visually impaired people have distinct images relationships which are interfered with by the person's individuality and bonded to the visual impairment type [17]. Kastrup [48] complements that "*blind people do not have an inferior perception, but different from ours*" (sighted people).

Some participants, for example, P1, P5, and P8 expressed concern about **personal interpretations** in the appraised sentences, a characteristic mainly present in humans images descriptions. Conforming to Pettersson [70], an image description remarkably impacts our image perception, and we *"see what we are told to see"* in an image. Costa [17] performed an in-depth investigation regards interpretation in descriptive sentences, stating that a description's author (also called describer), by choosing to *describe* an image allows the other to infer its scene, whereas choosing to *interpret* does an explanation from your perspective, providing a point of view instead of an accurate description.

Moreover, Costa [17] investigated whether preferences for more subjective (interpretive) or objective descriptions were related to the visual impairment type, either congenital or acquired, either low vision or total blindness. The analysis reports that although it seems that there would be specific needs for each impairment type, this idea could not be stated in the research [17]. However, the author identified that the participants' choices (between interpretive and objective descriptions) were impacted by their knowledge of this accessibility resource [17]. Similarly, we observed in the interviews that participants with professional training in audio description (P5 and P8) were concerned with personal interpretation issues in a more expressive way than other participants without such training.

According to Praxedes Filho and Arraes [74], a descriptive sentence must be exempt from any personal opinion, thus respecting the people with visual impairments rights to create their judgments. In this statement, we perceive the similarity with the P5's speech

regards respecting autonomy and not interpret an image in the person's place. As Costa [17] defended, it is not possible to achieve total objectivity or subjectivity in descriptions, but it should not be an excuse for not reduce personal interpretation when describing an image. Magalhães and Praxedes Filho complement that we must delineate what we see in an image, never what we think we are seeing. [58].

When describers incorporate what an image is not representing, superfluous information is added in the descriptive sentences [58], which emphasizes the pertinence of keeping objectivity in mind when generating image descriptions [17]. Besides, as participants reported, overly long sentences are exhausting, evidencing the relevance of pragmatical image descriptions. From this reflection, we identified the following good practice (GP):

GP−1: when writing a descriptive sentence, be objective and avoid adding personal interpretations about the image. Answer to the question: **"What do I see in the image?"**

As our participants reported, **image's focal points** play a substantial role in a descriptive sentence. Its relevance is such that automatic image descriptions models seek to consider images' focal points (also called regions of interest) through the attention mechanism [85], as we explained in Chapter 2. According to Sá, Campos, and Silva [88], the visual system of a sighted person instantly and immediately detects and integrates more than 80% of the visual stimuli, taking about 2-3 seconds to recognize an image's content [71]. Therefore, when describing the visual content, we must prioritize what caught our attention.

Since sighted people can have different narratives on what an image is about or what is the image's focal point [45], it is clear, once more, the influence of personal inference in a descriptive sentence. For this reason, Costa [17] states that describers must constantly seek strategies to minimize personal inferences. Among the guiding elements that describers can use are identifying the relevant image's aspects and assigning a priority to describe them [17]. Similarly, Tang [90] suggests considering the context and what people want to know about the image.

We must remember that providing to people with visual impairments the same information that sighted people perceive is a matter of equity [84] and one of the recommendations of Accessibility Guidelines [96]. Therefore, the use of selective vision does not mean there is no need to describe what did not catch our attention, but rather consider the purpose and context of images to identify what to convey and assess its importance [90]. As Costa [17] and Tang [90] state, describers must first present the crucial content, followed by the least important information if there is time and space available. Based on the discussion presented, we identified the following good practice (GP):

GP−2: when writing a descriptive sentence, consider the image's purpose (why) and its focal points (what). Answer to the questions: **"Why is the image being presented or shared?"**, **"What are the salient regions or focal points?"**

We identified that **embedded texts** in images are relevant for understanding the context and purpose of the image's event. Gleason *et al.* [32] observed that to overcome the limitations of automatic descriptions models in identifying such texts, people with visual impairments use applications that perform Optical Character Recognition (OCR). Therefore, the visible texts in images must be reported in descriptions since a sighted person can access the referred information. In this way, we identified the following good practice (GP):

GP−3: when writing a descriptive sentence, include the visible text in the image. Answer to the question: **"Are any visibly embedded texts in the image?"**

An image is the visual representation of an **event**, *i.e.,* a moment recorded in a photograph. We identified that describing the image's context encompasses informing the **occasion** and the **nature** of the event. For example, a birthday party (occasion) can be of a person (informal nature) or a company (formal nature). In this way, we identified the following good practice (GP):

GP−4: when writing a descriptive sentence, consider the type and nature of what is happening in the image. Answer to the question: **"What is the event's context represented in the image?"**

The image event always takes **place** somewhere. As we previously discussed, even when image descriptions had informed a specific location, such as a library or a restaurant, expectations were perceived regarding more descriptive details of the settings. The previous discussion about personal interpretation in image descriptions led us to realize that stating a place without certainty, *i.e.,* based on what we think of the image, can result in erroneous descriptive sentences, which in turn, is one of the top two problems in images descriptions identified in the Snowballing. As shared by participant P5, the idea is not to make a person with visual impairment guess the image's place; however, we must observe the veracity of the information described. To avoiding erroneous information when writing a descriptive sentence, we can use reference points that appear in an image [90]. For example, the Eiffel Tower (reference point) denotes Paris (place), and Christ the Redeemer (reference point) denotes Rio de Janeiro (place). This discussion resulted in the following good practice (GP):

GP−5: when writing a descriptive sentence, avoid stating the setting's place unless it is explicit in the image, giving preference to characterizing it. Answer to the question: **"Which landmarks are visible in the image that indicates the setting as a [place]?"**

In the case of implicit places, it is critical to characterize the image's setting. As we identified, it includes describing the setting's **size or dimension** and its **ornamentation**, allowing the imagination of people with visual impairments about the image's place. For example, instead of informing that the image's setting is a *living room* (place), prefer describing the furniture items and other setting's objects that characterize it as a *living room*. We must pay attention to what participants P5 e P7 shared: same as poor image descriptions are lousy, very long descriptions are. Therefore, when characterizing the setting, we must be cautious to the sentence does not become exhaustive. For example, in environments with many objects, we should evaluate the need to describe each one of them. Thus, we must identify what is relevant to the image's understanding, observing its purpose and its context, as we exposed in good practices GP−2 and GP−3, respectively. We identified the following good practice regarding the setting's furnishing:

GP−6: when writing a descriptive sentence, include characteristics of the setting that lead a person to imagine where the scene is taking place. Answer to the question: **"What are the furnishings, the objects, and the decorative items that identify the setting as a [place]?"**

According to the participants' expectations, descriptions regards the setting's **size or dimension** refers to describing its physical space, including its **structural elements** as, for example, windows, stairs, and pillars, besides the setting's lighting and the elements' location. As we stated in good practice GP−1, we must be objective, so for example, in the sentence *"The room is well-lighted."*, we are not sure if the room is well-lighted because the lights are on or because there are large windows which sunlight radiates through it. Thus, we understand that the reasons why the room is well-lighted (or not) are more appropriate and should be used to characterizes it. In other words, we must ask ourselves why the setting is or not well-lighted. Concerning the setting's size or dimension, we identified the following good practice (GP):

GP−7: when writing a descriptive sentence, describe the setting's physical space. Answer to the questions: **"What is the setting's size/dimension?"**, **"What structural elements appear in the image?"(***e.g.,* **windows, stairs, and pillars.)**, **"How is the setting's lighting, and why is it well-lightened or not?"**

As discussed in good practice GP−5, a description of the image's setting includes its ornamental elements, *e.g.,* furniture items, decorative artifacts, and other objects. Moreover, we identified expectations regarding ornamental elements descriptions such as **colors**,

sizes, and **shapes**, besides informing their number and location. However, participants did not report a term preference or expectation to refer to the objects' size. Since expressions such as large, small, or big are relative to the describer's subjectivity (what is small for one may not be for another), the American Council of the Blind suggests using comparisons to describe an object's size [80]. According to the guideline provided, the comparisons need to include terms that are common knowledge and which people are familiar with, *e.g., "The small snake is as long and thick as a pencil."* and *"The puma is the size of a large dog."*. Through this discussion, we identified the following good practice (GP):

GP−8: when writing a descriptive sentence, describe the physical attributes of the setting's elements. Answer to the questions: **"What are the colors?"**, **"What are the sizes?"**, **"What are the shapes?"**

As previously related, a descriptive sentence must inform the **location** of the structural and ornamental elements. We identified that participants' expectations include the element's location in the image and the element's location in the setting relative to other image elements. To refer to such expectations, participants used expressions such as *background*, *foreground*, *behind*, among others. In this sense, we identified the following words to describe the element's location in images: *top*, *bottom*, *left/left upper/right upper*, and *right/right upper/right lower*. Concerning element's location in the setting relative to another, Tang [90] suggests the following expressions: *above*, *below*, *to the right/left of*, *in front of*, *behind*, *touches*, *crosses*, *overlaps*, *contains*, and *within*.

Furthermore, the image's elements must be described logically and consistently, for example, from left to right or from top to bottom [90]. Besides, the descriptive sentence must be organized, avoiding leaving loose elements, and trying to describe all the element's attributes before starting to describe another [89]. We want to emphasize that the term element refers to any image's component, *i.e.,* person, object, animal, food, among others, as their locations are essential to construct the mental imagery of the scene described. Besides, as participant P1 shared, this information is valuable for visually impaired people as they navigate according to the setting's spatial organization. The discussion presented resulted in the following good practices (GP):

GP−9: when writing a descriptive sentence, describe the element's location in the image. Answer to the question: **"Where is the [element] located in the image?"** (*e.g.,* top, bottom, right, left, among others.)

GP−10: when writing a descriptive sentence, describe the element's location in the setting relative to other elements. Answer to the question: **"Where is the [element] located in the setting relative to [another element]?"** (*e.g.,* above of, below of, to the right of, among others.)

A descriptive sentence must also inform the **number** of each image's elements. Regarding the people element, as participant P5 indicated, when there are less than ten people in images, the ideal is always to inform the respective number. As participant P7 shared, when images contain several people, there is no need to relate each person's clothes and hairstyles since it leads to a long and tiresome sentence. As explained by participant P2, it is also relevant to count people who appear partially in the image, informing this aspect in the descriptive sentence. Through this, we identified the following good practice (GP):

GP−11: when writing a descriptive sentence, describe the number of each element present in the image, including those partially appear. Answer to the question: **"How many of each element appears?"** (*e.g.,* people, objects, animals, among others.)

According to the participants, they expect **people's clothing** descriptions. We identified that this information assists people with visual impairments to understand the image's context, that is, the event represented. When describing people's clothing, we must start with the larger pieces and follow the top-down orientation [1]. As participant P7 explained, it is critical to inform whether people follow a dress code or a standard attire, for example, an image in which everyone is wearing a school uniform or suits. Besides, the image description should mention a stand-out outfit color, *e.g.,* a red coat, since it helps people with low vision to differing the image's elements [1]. We identified the following good practice (GP):

GP−12: when writing a descriptive sentence, describe the people's clothing in the image. Answer to the questions: **"How are people dressed?"**, **"Is there any dress code or standard garments?"** (*e.g.,* suit, long dress, school uniform, among others.), **"What are the clothes' colors?"**

Another characteristic of people refers to their **age group**. According to the participants, generic expressions are unhelpful, so for example, the expression older people could mean either adults or elderly. To refer to their expectations, participants used expressions as *children*, *teens*, *adults*, and *elderly*. In this sense, Hutchinson, Thompson, and Cock [43] recommend describing people's age by ranges of decades as, for example, *"Their ages range from 40 to early 50s."*. Through this discussion, we identified the following good practice (GP):

GP−13: when writing a descriptive sentence, describe the people's age group. Answer to the question: **"What range of age do people look to be?"** (*e.g.,* babies, teenagers, in their 70s, in their 20s, mid-40s, among others.)

Participants reported that an image description should inform the **people's actions** since this information is highly relevant to them understand an image's content. We identified an association between the image elements' actions and their interactions, including people, animals, and objects. Examples of image elements' interactions are two people having a conversation, a person playing a videogame, and a dog chasing a cat, among others. Regarding descriptions about the actions and interactions, we identified the following good practice (GP):

GP−14: when writing a descriptive sentence, describe what each image's element is doing and its interaction with others elements. Answer to the questions: **"What is the [element] doing, or what is the action it performs?"**, **"How are the image's elements interacting with each other?"**

Furthermore, we identified expectations regarding **people's facial and body expressions**. Participants reported that such descriptive sentences are vital to understanding the emotions and the humor represented in images. According to Costa [17], we must use descriptive sentences rather than interpretive ones as, for example, instead of relating *"The person is sad."*, we can describe it as *"The person has tears in the eyes and dropped mouth corners."*. Naves *et al.* [64] state that there are circumstances in which impersonal descriptions of people's gestures and expressions do not lead to a proper understanding and may lose their meaning in the sentence. In this way, the authors advise, if necessary, to describe the expression's meaning [64] as, for example, the sentence *"A person with a hand on the chin and a worried expression."*. Through this, we identified the following good practice (GP):

GP−15: when writing a descriptive sentence, describe the people's emotions in images. Answer to the question: **"What are people's facial and body expressions?"** (*e.g.,* wide eyes, raised mouth corners, arched eyebrows, among others.)

We identified several expectations regarding **people's appearance**, including race or ethnicity, gender, and hairstyles. Since such descriptions are a sensitive topic [7, 35], we will thoroughly discuss it. Some guidelines seek to orient the image descriptions' production, either in textual [96] or audio format [80]. Concerning the Brazilian context, in 2010, was released the first book about audio/textual description; however, it does not guide gender descriptions [4]. Regards people's race or ethnicity, it only informs that such physical attributes must be present during appearance descriptions [4].

Moreover, in August 2016, the Ministry of Culture released a guideline for accessible audiovisual productions, which briefly orient to describe people's physical characteristics

in the following sequence: gender, age group, ethnicity, skin color, height, physique, eyes, hair, and other outstanding attributes [64]. Similarly, the Brazilian Association of Technical Standards mentions this sequence [1]; however, neither the sources guide *how* or *what* terms fittingly describe people's characteristics. Thus, there is still an orientation deficiency relating to people's descriptions [7].

It is not possible to perceive, in an image, a **person's gender** since it is a complex relationship between three dimensions: body, identity, and social gender [81]. According to Bennett *et al.* [7], physical appearance and identity descriptions are distinct. The authors state that physical appearance description includes skin tone, hairstyle, clothing, and accessories, whereas an identity description refers to race, gender, and disability [7]. Descriptive sentences of a person's skin tone and hairstyle avoid race assumption, while attire, accessories, and hairstyle descriptions may avoid gender premise, and devices and technologies descriptions may prevent guessing a person's disability [7]. Moreover, a describer should opt for a physical appearance description in cases when a person's identity is unknown [7].

To refer to people's gender, in the English language, the pronoun *they* is used as a neutral pronoun [87]. Regarding the language's grammar, despite the existence of binary form words (female and male), *e.g.,* actor/actress, this phenomenon is less frequent than in the Portuguese language, which in turn, uses -a for female and -o for male words, *e.g., a atriz*/*o ator* [72]. Recently, the use of *non-binary* or *inclusive* language has emerged to avoid gender discrimination and include using, for example, the sign "@" and the letter "X" [65, 72]. However, such characters discriminate screen reader users, imposing an obstacle to accessing the content [8, 61]. For example, a screen reader would read the sentence *"tod@s @s alun@s"* (all students) as follows: *"tod arroba s arroba s alun arroba s."*. Would this be the best option?

According to Bennett *et al.* [7], erroneous gender assumptions made by AI models cause more harm than humans mistakes, mainly because a person can be aware of the discrimination and learn from it to avoid future prejudice, whereas automatic generators are inflexible. Besides, automatic generators would be more respectful if they use appearance descriptive language rather than identity presumptive language [7]. Keyes [49] exposed that AI models assume gender as a concept containing two and only two categories (man and woman), meaning that automatic gender identifications neither include trans and non-binary people.

Although some social media platforms allow gender customization in the user's profile [98], their automatic image descriptions still classify a person's gender only as female or male. This limitation is due to the dataset used in the models' training, which only contains binary terminology references [7, 14, 73]. Despite AI researchers are aware that gender identity classification is not necessarily binary, they treat gender inference as such [42]. In the literature, recent studies seek to include other gender terminologies in automatic models, which may lead to more inclusive image descriptions [101]. For example, Wu *et*

al. [101] collected public photos of famous people who identify themselves as non-binary or transgender, seeking to share an inclusive-gender dataset. However, extensive studies are required to improve the detection made by AI models since erroneous gender identification has significant ethical implications [101, 7].

Regarding a **person's race or ethnicity**, we must not dismiss it in images descriptions since its visual perception is a social phenomenon [43, 7]. Bittner [10] critically analyzed the governmental guidelines of audio description from Australia, France, Germany, Ireland, the United Kingdom, and the United States, from which only Australia has recommendations about ethnicity wherein merely propose describing it if relevant for the context, and none of them advises how properly refer to a person's gender. Once more, personal interpretation is evident since a describer needs to decide whether descriptions of ethnicity and gender are relevant or not to a listener [43]. It is intrinsic to us the complexity of describing people's appearance. Thus, it is inherent that describers understand the social implications of such descriptions as they convey gender and ethnicity's representativity in images [43, 7].

Gzara [35] advises avoiding the connotation *"native"* since it is impracticable to deduce a person's origin based only on a specific ethnic trait or skin color. Also, according to the author, there is no right or wrong answer, nor a solution that fits all circumstances, but we must always keep a non-offensive vocabulary and prefer the terminologies used by the ethnic's groups [35]. Concerning Brazilian terminologies, the Brazilian Institute of Geography and Statistics has used, since 1991, five categories to express the race or skin color's groups: white, black, yellow, pale brown, and Brazilian Indian (or Indigenous) [36, 56]. The referred Brazilian Institute considers race and skin color similarly, once it collects the population's demographic data through the question *"What is your skin color or race?"* [56], whereas Bennett *et al.* [7] treat such attributes distinctly, considering skin color as a person's physical appearance characteristic and race as a person's identity characteristic.

Hutchinson, Thompson, and Cock [43] investigated ethnicity and gender descriptions in theater performances which extensive study brings significant contributions that seem to apply to image descriptions. The authors suggest avoiding the words *diverse* and *multi-ethnic* and describing every single person. Besides, in a homogeneous group, first describe their ethnicity, with no need to repeat it on the individual description [43]. For example, if there are only white people or only black people, the suggestion is to refer to the group as *"all-white"* or *"all-black"*. In groups where there is race heterogeneity, the authors suggest describing it for every person or none, but never just for some of the group as it implies a standard's ethnicity assumption [43].

Regardless, claiming a person's ethnicity based on physical appearance can lead to misunderstandings and the assignment of a race other than the one the person identifies with it [43]. To overcome this issue, Snyder [80] suggests using skin color to refer to a person's ethnicity or race and cites expressions as, for example, light-skinned, dark-skinned,

and olive-skinned. However, this approach provides unspecific descriptions and may obstacle the listener's imagination [43]. Bennett *et al.* [7] argues that describing skin color seems more appropriate than determining a person's race, but this action can encourage discrimination. Moreover, the variety of skin tones could lead to more confusion in those who do not have visual color memories [7]. Thereby, there is no common consensus yet about adequately describe people's ethnicity. Although the obstacles previously pointed out, skin tones are less harmful than untrue race assumptions; thus, describers should prioritize this option when ethnicity or race is unknown [7].

We did not explore what terms the interview's participants prefer or consider appropriate to describe a person's ethnicity and gender. Therefore, are required more in-depth and specific studies on this matter. We identified, however, that participants used binary terms (female and male) to refer to their gender description's expectations. Nevertheless, as we discussed, such terminologies omit non-binary or transgender people. In this vein, we understand that describers should refer to a person's gender exclusively when it is known, *i.e.,* when a describer is familiar with the people in the image or when the gender is publicly known as, for example, famous people who expressed their pronoun preference.

In images where describers are familiar with the image's individuals, a suggestion is to employ the same terminology the individuals used when describing themselves [43]. However, we must be sensitive and note that a person may prefer not to share this personal information. In these cases, describers should delineate the person's physical appearance rather than the identity-related characteristics [7]. Furthermore, describers must keep abreast with the terminologies used by non-white and non-binary people for a more adequate and inclusive image description [7].

Conclusively, we must not describe what is not visibly perceive in images. Elucidating this idea, we bring what Magalhães and Praxedes Filho [58] exemplified regarding the use of adjectives in image descriptions: *"Beautiful solely says that something is not ugly. But what exactly makes it beautiful?"*. It does not mean avoiding adjectives in image descriptions, but we must observe the context and limit their use to occasions when there is temporal restraining [17]. Concerning people's physical appearance descriptions, we identified the following good practice (GP):

GP−16: when writing a descriptive sentence, describe the people's appearance. Avoid expressions that convey only your point of view, and use a non-offensive vocabulary. Answer to the question. **"What are the people's physical characteristics in the image?"** (*e.g.,* skin tone, hairstyles, and other accessories such as glasses and earrings.)

Other expectations regard people in images refer to their **professions** and **interactions** between them. We identified that people's professions are related to their outfits which, in turn, help people with visual impairments to understand the image's context. Describers should, however, only indicate the people's professions if evident in images as, for example,

people wearing firefighter outfits or medical scrubs. Moreover, describers must point people's relationships solely when they are familiar to describers or when their relationship is public-known, as famous people. Thereby, we identified the following good practice (GP):

GP−17: when writing a descriptive sentence, describe what implies labors activities in the image. Answer to the questions: **"What explicit occupational duties do people perform?"**,**"Are they wearing a uniform or accessories?"** (*e.g.,* firefighter outfits, medical scrubs, among others.)

Participants also reported expectations regards the **dish's names** and **beverages** in images, including their **qualities** as size and appearance. As participants shared, for example, in a picture whose focal point is a cake, it is essential to delineate it, providing more descriptive details than just *"It is a cake."*. Besides, image descriptions should include kitchen utensils as, for example, plates, cutlery, glasses, and cups, if applicable to the image's context. Hence, we identified the following good practice (GP):

GP−18: when writing a descriptive sentence, describe the dish's names, beverages, including their qualities, as well as kitchen utensils. Answer to the questions: **"What food and drinkable appear in the image?"**,**"What are their size and appearance?"**, **"Are there any kitchen-related elements relevant to the image's context?"**

Lastly, we identified that participants' expectations about **animals** include their **age groups** and **fur/coat colors**. The images used for the interviews did not comprise many animals since their descriptions were not our primary objective. Therefore, the collected data may not include all relevant aspects about animals descriptions. Thereby, we identified the following good practice (GP):

GP−19: when writing a descriptive sentence, describe the physical characteristics of animals in the image. Answer to the questions: **"What age groups do animals resemble?"** (*e.g.,* puppy, kitten, adult, elderly, among others.), **"What are the animal's coats colors?"**

Table 5.1 summarizes the nineteen good practices in image descriptions outlined in this chapter. We hope our findings contribute to future images descriptions generation to more acceptable and fair descriptive sentences for people with visual impairments.

Table 5.1: Good Practices (GP) in image descriptions for People with visual impairments.

Good Practice (GP)	Questions
GP-1	What do I see in the image?
GP-2	Why is the image being presented or shared? What are the salient regions or focal points?
GP-3	Are any visibly embedded texts in the image?
GP-4	What is the event's context represented in the image?
GP-5	Which landmarks are visible in the image that indicates the setting as a [place]?
GP-6	What are the furnishings, the objects, and the decorative items that identify the setting as a [place]?
GP-7	What is the setting's size/dimension? What structural elements appear in the image?(*e.g.,* windows, stairs, and pillars.) How is the setting's lighting, and why is it well-lightened or not?
GP-8	What are the colors? What are the sizes? What are the shapes?
GP-9	Where is the [element] located in the image? (*e.g.,* top, bottom, right, left, among others.)
GP-10	Where is the [element] located in the setting relative to [another element]? (*e.g.,* above of, below of, to the right of, among others.)
GP-11	How many of each element appears?
GP-12	How are people dressed? Is there any dress code or standard garments? (*e.g.,* suit, long dress, school uniform, among others.) What are the clothes' colors?
GP-13	What range of age do people look to be?(*e.g.,* babies, teenagers, in their 70s, in their 20s, mid-40s, among others.)
GP-14	What is the [element] doing, or what is the action it performs? How are the image's elements interacting with each other?
GP-15	What are people's facial and body expressions? (*e.g.,* wide eyes, raised mouth corners, arched eyebrows, among others.)
GP-16	What are the people's physical characteristics in the image?(*e.g.,* skin tone, hairstyles, and other accessories such as glasses and earrings.)
GP-17	What explicit occupational duties do people perform? Are they wearing a uniform or accessories? (*e.g.,* firefighter outfit, medical scrubs, among others.)
GP-18	What food and drinkable appear in the image? What are their size and appearance? Are there any kitchen-related elements relevant to the image's context?
GP-19	What age group do animals resemble?(*e.g.,* puppy, kitten, adult, elderly, among others.) What are the animal's coats colors?

6. FINAL CONSIDERATIONS

This Chapter presents the final considerations of this study. Section 6.1 shows its limitations, and Section 6.2 presents future work.

Three Research Questions (RQ) guided this study's execution. The first RQ sought to identified issues in image descriptions for people with visual impairments. We answered it through an investigation in the literature using the snowballing technique, and we identified thirteen issues from eleven studies. Despite being a recommendation of the most basic level of accessibility, people with visual impairments still face mediocre and missing descriptive texts, which are the top issues we identified.

The second guiding Research Question refers to the characteristics of image descriptions, which investigation occurred through semi-structured interviews with eight low vision participants. We identified reasons for satisfaction and dissatisfaction in image descriptions, either automatic and human-generated. Satisfactory characteristics of the automatic-generated sentences include descriptions of the people's position in the images, objects' identification, and people's gender. Dissatisfaction reasons include mediocre descriptions of the settings' places, people's actions, and the number of people in images.

Concerning human-generated sentences, participants expressed satisfaction when they described people's actions, gender, and position, whereas dissatisfaction reasons include setting's ornamentation, setting's place, and the number of people in images. In either image descriptions type, generic sentences confused and did not help the participants' image understanding. Also, mediocre descriptive texts cause frustration and sadness.

The third guiding Research Question sought to collect the participants' expectations of image descriptions. We have identified 26 attributes of image descriptions to meet participants' expectations. Based on the interviews data and the snowballing results, we identified nineteen good practices in image descriptions for people with visual impairments, which seek to guide the development of more satisfactory sentences, facilitating the image understanding process.

Through the obtained results, we noticed improvements' demand for AI models to generate higher quality descriptive sentences, especially about people's characteristics, including their clothing, facial expressions, actions performed. Besides, we identified that the absence of descriptions inhibits the inclusion of visually impaired people, causing frustration and isolation feelings, exposing the need and relevance of accessible imagery contents.

Furthermore, it is inherent to raise sighted people's awareness about image descriptions' importance and how to describe imagery content for visually impaired people. We identified that descriptions of people's ethnicity and gender are an equality matter for visually impaired people; however, AI models still fail to describe them, preferring to omit such information and remain impartial. This impartial behavior strengthens the need for further

studies about how politely inform people's physical appearance to enhance the inclusion of visually impaired people without denouncing others such as non-binary, trans, and non-white people.

Finally, this study sought to contribute with a group of good practices in image descriptions, and we expect to support future generations of human and automatic descriptive sentences. From these lessons, we hope to contribute to the scientific community of the CHI and other research's fields by identifying current limitations of images descriptions produced manually and by automatic models from the perspectives of people with visual impairments, as well as their satisfaction and dissatisfaction with image descriptions, and their descriptions' expectations. We expect our contribution will inspire future research to minimize the limitations of image descriptions and enhance their quality to improve the image understanding of people with visual impairments.

6.1 Limitations

Regarding the limitations of this study, although we did not specifically search for the social media context, we observed that nine of the eleven studies selected in snowballing referred to this context and, therefore, may not encompass all issues in image descriptions. Moreover, the snowballing did not identify whether image descriptions issues are related to visual impairments types as many studies used terms as "people with visual impairments" or "screen readers users." Furthermore, the snowballing does not include all relevant studies due to the limited scope of the publications' places we considered. Therefore, we understand that the snowballing results do not contain all image descriptions' issues for visually impaired people.

Regarding the online survey's limitation, the sighted people's rating may not represent the best choice in the opinion of people with visual impairments. Besides, responses may present bias among the respondents. About the interview study's limitation, we recruited only low vision people; thus, future studies need to investigate the perspectives of blind people. Despite paying attention to the coding process and remaining faithful to the participants' speeches, we understand that the coder's interpretation can influence the coding process. Furthermore, the selected images represented indoor environments and do not encompasses all the relevant contexts for visually impaired people.

6.2 Future Work

Perspectives for future studies include: investigating the relation between image descriptions' issues and visual impairment type, *i.e.,* if they occur exclusively for blind peo-

ple or people with low vision; analyzing non-internal image contexts to identify participants' expectations in these contexts; in-depth and specific studies on descriptions of people's physical appearance, especially issues related to gender and ethnicity and; evaluating the efficiency of the good practices we identified, seeking to assess if they meet the instructions needs of sighted people about how to describe imaginary content for people with visual impairments.

REFERENCES

[1] ABNT. Associação Brasileira de Normas Técnicas/Brazilian Association of Technical Standards. NBR 16452:2016. Comitê Técnico de Acessibilidade. “Accessibility in communication — audio description”. Source: https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/ABNT%20-%20Acessibilidade.pdf, Jul 2021.

[2] Adams, D.; Morales, L.; Kurniawan, S. “A qualitative study to support a blind photography mobile application”. In: Proceedings of the International Conference on PErvasive Technologies Related to Assistive Environments, 2013, pp. 1–8.

[3] Amirian, S.; Rasheed, K.; Taha, T. R.; Arabnia, H. R. “A short review on image caption generation with deep learning”. In: Proceedings of the International Conference on Image Processing, Computer Vision, Pattern Recognition, 2019, pp. 10–18.

[4] Araújo, V. L. S. “A formação de audiodescritores no ceará e em minas gerais: Uma proposta baseada em pesquisa acadêmica”. In: *Audiodescrição: transformando imagens em palavras*, Motta, L. M. V. M.; Filho, P. R. (Editors), Secretaria dos Direitos da Pessoa com Deficiência do Estado de São Paulo, 2010, pp. 93–105.

[5] Aung, S. P. P.; Pa, W. P.; Nwe, T. L. “Automatic myanmar image captioning using cnn and lstm-based language model”. In: Proceedings of the Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages, 2020, pp. 139–143.

[6] Bai, S.; An, S. “A survey on automatic image caption generation”, *Neurocomputing*, vol. 311, May 2018, pp. 291–304.

[7] Bennett, C. L.; Gleason, C.; Scheuerman, M. K.; Bigham, J. P.; Guo, A.; To, A. ““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, 2021, pp. 1–19.

[8] Bernardo, V. U. C. “A referenciação a partir da concordância de gênero no português brasileiro: uma abordagem da linguística cognitiva”, Master’s Thesis, Faculdade de Filosofia, Letras e Ciências Humanas – Universidade de São Paulo, São Paulo, SP, Brasil, 2019, 141p.

[9] Biswas, R.; Barz, M.; Sonntag, D. “Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking”, *KI-Künstliche Intelligenz*, vol. 34, Jul 2020, pp. 1–14.

[10] Bittner, H. “Audio description guidelines: a comparison”, *New Perspectives in Translation*, vol. 20, Jan 2012, pp. 41–61.

[11] Brannen, J. “Mixing methods: The entry of qualitative and quantitative approaches into the research process”, *International Journal of Social Research Methodology*, vol. 8–3, Feb 2005, pp. 173–184.

[12] Chandrashekar, S.; Stockman, T.; Fels, D.; Benedyk, R. “Using think aloud protocol with blind users: A case for inclusive usability evaluation methods”. In: Proceedings of the International Conference on Computers and Accessibility, 2006, pp. 251–252.

[13] Chen, S.; Jin, Q.; Wang, P.; Wu, Q. “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, 2020, pp. 9959–9968.

[14] Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C. L. “Microsoft coco captions: Data collection and evaluation server”. In: Proceedings of the European Conference on Computer Vision, 2014, pp. 740–755.

[15] Chiarella, D.; Yarbrough, J.; Jackson, C. A. “Using alt text to make science twitter more accessible for people with visual impairments”, *Nature communications*, vol. 11–1, Nov 2020, pp. 5803.

[16] Choi, J.; Jung, S.; Park, D. G.; Choo, J.; Elmqvist, N. “Visualizing for the non-visual: Enabling the visually impaired to use visualization”, *Computer Graphics Forum*, vol. 38–3, Jul 2019, pp. 249–260.

[17] Costa, L. M. “Audiodescrição em filmes: história, discussão conceitual e pesquisa de recepção”, Ph.D. Thesis, Departamento de Letras–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brasil, 2014, 401p.

[18] C.Rowe, N. “Marie-4: A high-recall, self-improving web crawler that finds images using captions”, *IEEE Intelligent Systems*, vol. 17–4, Jul. 2002, pp. 8–14.

[19] CRP-PR. Conselho Regional de Psicologia do Paraná. “Você sabe o que é #pracegover?” Source: https://crppr.org.br/pracegover/, Dez 2020.

[20] Denkowski, M.; Lavie, A. “Meteor universal: Language specific translation evaluation for any target language”. In: Proceedings of the Workshop on Statistical Machine Translation, 2014, pp. 376–380.

[21] Diamant, E. “Unveiling the mystery of visual information processing in human brain”, *Brain Research*, vol. 1225, Aug 2008, pp. 171–178.

[22] Dognin, P.; Melnyk, I.; Mroueh, Y.; Ross, J.; Sercu, T. “Adversarial semantic alignment for improved image captions”. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2019, pp. 10455–10463.

[23] Downe-Wamboldt, B. “Content analysis: Method, applications, and issues”, *Health Care for Women International*, vol. 13–3, Nov 1992, pp. 313–321.

[24] eMAG. Modelo de Acessibilidade em Governo Eletrônico. “v3.1. 2014”. Source: http://emag.governoeletronico.gov.br/, Jan 2021.

[25] Elo, S.; Kyngäs, H. “The qualitative content analysis process”, *Journal of Advanced Nursing*, vol. 62–1, Nov 2008, pp. 107–115.

[26] Ferreira, S. B. L.; da Silveira, D. S.; Capra, E. P.; Ferreira, A. O. “Protocols for evaluation of site accessibility with the participation of blind users”. In: Proceedings of the International Conference on Software Development for Enhancing Accessibility and Fighting Info-exclusion, 2012, pp. 47–55.

[27] Floriano, M. D. P.; Junior, P. V. C.; Silva, A. H. “#pracegover: uma discussão da inclusão digital e social sob a ótica da pesquisa transformativa do consumidor”, *CTS: Revista iberoamericana de ciencia, tecnología y sociedad*, vol. 15–45, Oct 2020, pp. 183–207.

[28] Frey, L. R.; Botan, C. H.; Kreps, G. L. “Investigating Communication: An Introduction to Research Methods”. Boston, MA, United States: Allyn Bacon, 1999, second ed., chap. 9, pp. 139–161.

[29] Frota, M. P. “Erros e lapsos de tradução: um tema para o ensino”, *Cadernos de Tradução*, vol. 1–17, Apr 2006, pp. 141–156.

[30] Gaur, S. “Generation of a short narrative caption for an image using the suggested hashtag”. In: Proceedings of the International Conference on Data Engineering Workshops, 2019, pp. 331–337.

[31] Gleason, C.; Carrington, P.; Cassidy, C.; Morris, M. R.; Kitani, K. M.; Bigham, J. P. ““it’s almost like they’re trying to hide it”: How user-provided image descriptions have failed to make twitter accessible”. In: Proceedings of the The World Wide Web Conference, 2019, pp. 549–559.

[32] Gleason, C.; Pavel, A.; McCamey, E.; Low, C.; Carrington, P.; Kitani, K. M.; Bigham, J. P. “Twitter a11y: A browser extension to make twitter images accessible”. In: Proceedings of the Conference on Human Factors in Computing Systems, 2020, pp. 1–12.

[33] Guinness, D.; Cutrell, E.; Morris, M. R. “Caption crawler: Enabling reusable alternative text descriptions using reverse image search”. In: Proceedings of the Conference on Human Factors in Computing Systems, 2018, pp. 1–11.

[34] Gurari, D.; Zhao, Y.; Zhang, M.; Bhattacharya, N. “Captioning images taken by people who are blind”. In: Proceedings of the European Conference on Computer Vision, 2020, pp. 417–434.

[35] Gzara, N. “Audio description and ethnicity”. In: Proceedings of the Swiss Conference on Barrier-free Communication, 2020, pp. 87–94.

[36] Haag, C. “Brazilian diversity”, *Pesquisa Fapesp*, vol. 1–173, Jul 2010, pp. 141–156.

[37] Hochreiter, S.; Schmidhuber, J. “Long short-term memory”, *Neural computation*, vol. 9–8, May 1997, pp. 1735–1780.

[38] Hodosh, M.; Young, P.; Hockenmaier, J. “Framing image description as a ranking task: Data, models and evaluation metrics”, *Journal of Artificial Intelligence Research*, vol. 47–1, May 2013, pp. 853–899.

[39] Hollink, L.; Schreiber, A. T.; Wielinga, B. J.; Worring, M. “Classification of user image descriptions”, *International Journal of Human-Computer Studies*, vol. 61–5, Nov 2004, pp. 601–626.

[40] Hossain, M. Z.; Sohel, F.; Shiratuddin, M. F.; Laga, H. “A comprehensive survey of deep learning for image captioning”, *ACM Computing Surveys*, vol. 51–6, Feb 2019, pp. 118–154.

[41] Hrga, I.; Ivašić-Kos, M. “Deep image captioning: An overview”. In: Proceedings of the International Convention on Information and Communication Technology, Electronics and Microelectronics, 2019, pp. 995–1000.

[42] Hu, Y.; Hu, C.; Tran, T.; Kasturi, T.; Joseph, E.; Gillingham, M. “What’s in a name? – gender classification of names with character based machine learning models”. 2102.03692, Source: https://arxiv.org/abs/2102.03692, Jul 2021.

[43] Hutchinson, R.; Thompson, H.; Cock, M. “Describing diversity: an exploration of the description of human characteristics and appearance within the practice of theatre audio description”, Technical Report, Describing Diversity project in partnership of VocalEyes with Royal Holloway, University of London, 2020, 84p.

[44] IBM. “Ibm developer model asset exchange: Image caption generator”. Source: https://developer.ibm.com/technologies/artificial-intelligence/models/max-image-caption-generator/, Jan 2021.

[45] Ilinykh, N.; Dobnik, S. “When an image tells a story: The role of visual and semantic information for generating paragraph descriptions”. In: Proceedings of the International Conference on Natural Language Generation, 2020, pp. 338–348.

[46] Jaimes, A.; Chang, S.-F. “A conceptual framework for indexing visual information at multiple levels”, *Electronic Imaging*, vol. 3964, Jan 2000, pp. 2–15.

[47] Jandrey, A. H.; Ruiz, D. D. A.; Silveira, M. S. “Image descriptions’ limitations for people with visual impairments: Where are we and where are we going?” In: Proceedings of the Brazilian Symposium on Human Factors in Computing Systems, 2021, pp. 1–11.

[48] Kastrup, V. “Atualizando virtualidades: construindo a articulação entre arte e deficiência visual”. In: *Exercícios de ver e não ver: arte e pesquisa com pessoas com deficiência visual*, Moss, A.; Rodrigues, S. (Editors), Rio de Janeiro, RJ, Brazil: Nau editora, 2010, chap. 2, pp. 32–46.

[49] Keyes, O. “The misgendering machines: Trans/hci implications of automatic gender recognition”, *Proceedings of the ACM on Human-Computer Interaction*, vol. 2–CSCW, Nov 2018, pp. 1–22.

[50] Kinoe, Y.; Noguchi, A. “Qualitative study for the design of assistive technologies for improving quality of life of visually impaired”. In: Proceedings of the International Conference on Human Interface and the Management of Information. Information and Knowledge in Applications and Services, 2014, pp. 602–613.

[51] Kothari, C. R. “Research Methodology Methods and Techniques”. New Delhi, Delhi, India: New Age International, 2013, second ed., 418p.

[52] Kougia, V.; Pavlopoulos, J.; Androutsopoulos, I. “A survey on biomedical image captioning”. In: Proceedings of the Workshop on Shortcomings in Vision and Language, 2019, pp. 26–36.

[53] Lazar, J.; Allen, A.; Kleinman, J.; Malarkey, C. “What frustrates screen reader users on the web: A study of 100 blind users”, *International Journal of Human–Computer Interaction*, vol. 22–3, Dec 2007, pp. 247–269.

[54] Lin, C.-Y. “Rouge: A package for automatic evaluation of summaries”. In: Proceedings of Workshop on Text Summarization Branches Out, 2004, pp. 74–81.

[55] Liu, S.; Bai, L.; Hu, Y.; Wang, H. “Image captioning based on deep neural networks”. In: Proceedings of the International Conference on Electronic Information Technology and Computer Engineering, 2018, pp. 1052–1058.

[56] Loveman, M.; Muniz, J. O.; Bailey, S. R. “Brazil in black and white? race categories, the census, and the study of inequality”, *Ethnic and Racial Studies*, vol. 35–8, Sep 2012, pp. 1466–1483.

[57] MacLeod, H.; Bennett, C. L.; Morris, M. R.; Cutrell, E. “Understanding blind people’s experiences with computer-generated captions of social media images”. In:

Proceedings of the Conference on Human Factors in Computing Systems, 2017, pp. 5988–5999.

[58] Magalhães, C. M.; Praxedes Filho, P. H. L. “Neutrality in audio descriptions of paintings: an appraisal system-based study of corpora in english and portuguese”, *Revista da Anpoll*, vol. 44–1, Feb-Apr 2018, pp. 279–298.

[59] Makav, B.; Kılıç, V. “A new image captioning approach for visually impaired people”. In: Proceedings of the International Conference on Electrical and Electronics Engineering, 2019, pp. 945–949.

[60] Marchal, N.; Neudert, L.-M.; Kollanyi, B.; Howard, P. “Investigating visual content shared over twitter during the 2019 eu parliamentary election campaign”, *Media and Communication*, vol. 9–1, Feb 2021, pp. 158–170.

[61] Mello, A. G.; Fernandes, F. B. M. “Guia de orientações básicas sobre gênero, deficiência e acessibilidade”, Technical Report, Cartilha da Comissão de Acessibilidade do Congresso Mundos de Mulheres junto ao Seminário Internacional Fazendo Gênero, Florianópolis, SC, Brasil, 2017, 23p.

[62] Morris, M. R.; Johnson, J.; Bennett, C. L.; Cutrell, E. “Rich representations of visual content for screen reader users”. In: Proceedings of the Conference on Human Factors in Computing Systems, 2018, pp. 1–11.

[63] Morris, M. R.; Zolyomi, A.; Yao, C.; Bahram, S.; Bigham, J. P.; Kane, S. K. “"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, 2016, pp. 5506–5516.

[64] Naves, S. B.; Mauch, C.; Alves, S. F.; Araújo, V. L. S. “Guia para produções audiovisuais acessíveis”, Technical Report, Ministério da Cultura por meio da Secretaria do Audiovisual, Rio de Janeiro, RJ, Brasil, 2016, 80p.

[65] Oliveira, E. R.; Amorim, E. C. F.; Ueda, A. H.; Rodrigues, P. R. “Computação para tod@s: criação, planejamento e realização de um evento sobre equidade de gênero”. In: Proceedings of the Latin American Women in Computing Congress hold as part of the Latin American Computing Conference, 2018, pp. 1–10.

[66] Palmer, C.; Bolderston, A. “A brief introduction to qualitative research”, *Canadian Journal of Medical Radiation Technology*, vol. 37–1, Mar 2006, pp. 16–19.

[67] Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. “Bleu: A method for automatic evaluation of machine translation”. In: Proceedings of the Annual Meeting on Association for Computational Linguistics, 2002, pp. 311–318.

[68] Pereira, L. S.; Ferreira, S. B. L.; Archambault, D. “Preliminary web accessibility evaluation method through the identification of critical items with the participation of visually impaired users”. In: Proceedings of the International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, 2015, pp. 77–86.

[69] Petrie, H.; Harrison, C.; Dev, S. “Describing images on the web: a survey of current practice and prospects for the future”. In: Proceedings of the Human Computer Interaction International, 2005, pp. 1–10.

[70] Pettersson, R. “Interpretation of image content”, *Educational Technology Research and Development*, vol. 36–1, Mar 1988, pp. 45–55.

[71] Pettersson, R. “Views on visual literacy”, *Journal on Images and Culture*, vol. 1–1, Feb 2013, pp. 1–9.

[72] Pio, C.; Silva, E. V. “An inclusionary open access textbook for portuguese”. In: *New case studies of openness in and beyond the language classroom*, Comas-Quinn, A.; Beaven, A.; Sawhill, B. (Editors), Research-publishing.net, 2019, pp. 11–22.

[73] Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; Lazebnik, S. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models”, *International Journal of Computer Vision*, vol. 123–1, May 2017, pp. 74–93.

[74] Praxedes Filho, P. H. L.; de Albuquerque e Arraes, D. “To appraise or not to appraise, that is the question. the state of the art in appraisal research in audio description”, *Trabalhos em Linguística Aplicada*, vol. 56–2, May-Aug 2017, pp. 379–415.

[75] Rashtchian, C.; Young, P.; Hodosh, M.; Hockenmaier, J. “Collecting image annotations using amazon’s mechanical turk”. In: Proceedings of the Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, 2010, pp. 139–147.

[76] Sacramento, C.; Nardi, L.; Ferreira, S. B. L.; ao Marcelo dos Santos Marques, J. “#pracegover: Investigating the description of visual content in brazilian online social media”. In: Proceedings of the Brazilian Symposium on Human Factors in Computing Systems, 2020, pp. 1–10.

[77] Saldaña, J. “The Coding Manual for Qualitative Researchers”. London, UK: SAGE Publications Ltd, 2013, second ed., 328p.

[78] Sharma, H.; Agrahari, M.; Singh, S. K.; Firoj, M.; Mishra, R. K. “Image captioning: A comprehensive survey”. In: Proceedings of the International Conference on Power Electronics IoT Applications in Renewable Energy and its Control, 2020, pp. 325–328.

[79] Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning”. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2018, pp. 2556–2565.

[80] Snyder, J. “Audio description guidelines and best practices”, Technical Report, American Council of the Blind’s Audio Description Project, USA, 2010, 98p.

[81] Spectrum, G. “Understanding gender”. Source: https://genderspectrum.org/articles/understanding-gender, Jul 2021.

[82] Splendiani, B.; Turró, M. R.; García, R.; Salse, M. “An interdisciplinary approach to alternative representations of images”. In: Proceedings of the International Conference on Computers Helping People with Special Needs, 2012, pp. 153–158.

[83] Srinivasan, L.; Sreekanthan, D.; A.L, A. “Image captioning - a deep learning approach”, *International Journal of Applied Engineering Research*, vol. 13–9, Sep 2018, pp. 7239–7242.

[84] Stangl, A.; Morris, M. R.; Gurari, D. “’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, 2020, pp. 1–13.

[85] Staniūtė, R.; Šešok, D. “A systematic literature review on image captioning”, *Applied Sciences*, vol. 9–10, May 2019, pp. 2024–2043.

[86] Stemler, S. “An overview of content analysis”, *Practical Assessment, Research, and Evaluation*, vol. 7–1, Nov 2000, pp. 1–10.

[87] Sun, T.; Webster, K.; Shah, A.; Wang, W. Y.; Johnson, M. “They, them, theirs: rewriting with gender-neutral english”. 2102.06788, Source: https://arxiv.org/abs/2102.06788, May 2021.

[88] Sá, E. D.; Campos, I. M.; Silva, M. B. C. “Inclusão escolar de alunos cegos e com baixa visão”. In: *Atendimento educacional especializado em deficiência visual*, Brasilia, DF, Brazil: SEESP/SEED/MEC, 2007, chap. 1, pp. 13–40.

[89] Sá, L. R. S.; Hubert, L.; Nunes, J. S. “Técnicas de audiodescrição aplicadas à internet e sites”, Technical Report, Fundação Escola Nacional de Administração Pública, Brasília, DF, Brasil, 2020, 47p.

[90] Tang, L. “Producing informative text alternatives for images”, Ph.D. Thesis, University of Saskatchewan, Saskatoon, SK, Canada, 2012, 286p.

[91] Thomas, D. R. “A general inductive approach for analyzing qualitative evaluation data”, *American Journal of Evaluation*, vol. 27, Jun 2006, pp. 237–246.

[92] Vedantam, R.; Zitnick, C. L.; Parikh, D. “Cider: Consensus-based image description evaluation”. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[93] Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. “Show and tell: Lessons learned from the 2015 mscoco image captioning challenge”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39–4, Sep 2017, pp. 652–663.

[94] Voykinska, V.; Azenkot, S.; Wu, S.; Leshed, G. “How blind people interact with visual content on social networking services”. In: In Proceedings of the Conference on Computer-Supported Cooperative Work Social Computing, 2016, pp. 1584–1595.

[95] W3C.World Wide Web Consortium. “Understanding success criterion 1.1.1: Non-text content”. Source: https://www.w3.org/WAI/WCAG21/Understanding/non-text-content, Jul 2021.

[96] WCAG.Web Content Accessibility Guidelines. “v2.0 (2008)”. Source: https://www.w3.org/TR/WCAG20/, Jan 2021.

[97] WebAIM. “The webaim million”. Source: https://webaim.org/projects/million/, Jul 2021.

[98] Weber, P. “Facebook offers users 56 new gender options: Here’s what they mean”. Source: https://theweek.com/articles/450873/facebook-offers-users-56-new-gender-options-heres-what-mean, Jul 2021.

[99] Wohlin, C. “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, 2014, pp. 1–10.

[100] Wu, S.; Wieland, J.; Farivar, O.; Schiller, J. “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, 2017, pp. 1180–1192.

[101] Wu, W.; Protopapas, P.; Yang, Z.; Michalatos, P. “Gender classification and bias mitigation in facial images”. In: Proceedings of the Conference on Web Science, 2020, pp. 106–114.

[102] Zhao, Y.; Wu, S.; Reynolds, L.; Azenkot, S. “The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments”, *ACM Transactions on Computer-Human Interaction*, vol. 1–CSCW, Dec 2017, pp. 1–22.

APPENDIX A – FREE AND CLARIFIED CONSENT TERM (FCCT)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Nós, Alessandra Helena Jandrey (aluna de Mestrado), Duncan Dubugras Alcoba Ruiz (professor orientador) e Milene Selbach Silveira (professora co-orientadora), responsáveis pela pesquisa **Um estudo qualitativo sobre a percepção de descrições de imagens por deficientes visuais**, estamos fazendo um convite para você participar como voluntário nesse estudo.

Esta pesquisa pretende avaliar em quais contextos e os motivos pelos quais uma descrição de imagem é satisfatória para o deficiente visual. Acreditamos que essa pesquisa seja importante porque seus resultados poderão ser usados como base para elaborações de descrições de imagens automáticas e humanas no futuro.

Para a realização da pesquisa serão seguidas as seguintes etapas: apresentação de imagens digitais, suas respectivas descrições, e discussão sobre as descrições apresentadas. Sua participação constará na análise destas imagens e nas respostas às perguntas de uma entrevista conduzida pelo pesquisador. A duração prevista é de, no mínimo, 30 minutos e, no máximo, 1 hora e 30 minutos, com possibilidade de intervalos para descanso a cada 30 minutos.

É possível que aconteçam os seguintes desconfortos ou riscos como dor de cabeça e cansaço ou aborrecimento durante a realização das atividades ou da entrevista. Desconforto, constrangimento ou alterações de comportamento durante as gravações de áudio e vídeo podem acontecer. Além disso, divulgação de dados confidenciais ou quebra de sigilo podem ocorrer. Além dos desconfortos que você possa sentir em virtude das respostas a este questionário, é possível que, infelizmente, sua conexão falhe, ou apresente certa lentidão, ou que você tenha dúvidas em como salvar suas respostas. O pesquisador tomará notas durante a entrevista e, portanto, você não precisará digitar ou salvar as respostas. Em caso de dúvidas ou esclarecimentos, não hesite em contatar o pesquisador Duncan Dubugras Alcoba Ruiz (duncan.ruiz@pucrs.br) no telefone [redacted] a qualquer hora. Você tem o direito de pedir uma indenização por qualquer dano que, comprovadamente, resulte da sua participação no estudo.

Os benefícios que esperamos do estudo são: identificação de quais informações são importantes para os deficientes visuais considerando os diferentes contextos; quando cada descrição é satisfatória; e quais são as expectativas de descrições nos contextos representados pelas imagens, podendo assim apoiar futuros pesquisadores de descritores automáticos, e também pessoas videntes, para que as perspectivas e as opiniões dos deficientes visuais sejam consideradas. Durante todo o período da pesquisa você tem o direito de esclarecer qualquer dúvida ou pedir qualquer informação sobre o estudo, bastando para isso entrar em contato, com Duncan Dubugras Alcoba Ruiz (duncan.ruiz@pucrs.br) no telefone [redacted] a qualquer hora.

Em caso de algum problema relacionado com a pesquisa você terá direito à assistência gratuita que será prestada pelos pesquisadores a partir do contato citado anteriormente. Você tem garantido o seu direito de não aceitar participar ou de retirar sua permissão, a qualquer momento, sem nenhum tipo de prejuízo ou retaliação, pela sua decisão. Se por algum motivo você tiver despesas decorrentes da sua participação neste estudo com transporte e/ou alimentação, você será reembolsado adequadamente pelos pesquisadores.

As informações desta pesquisa serão confidenciais, e serão divulgadas apenas em eventos ou publicações científicas, não havendo identificação dos participantes, a não ser entre os responsáveis pelo estudo, sendo assegurado o sigilo sobre sua participação.

Caso você tenha qualquer dúvida quanto aos seus direitos como participante de pesquisa, entre em contato com Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Rio Grande do Sul (CEP-PUCRS) em (51) 33203345, Av. Ipiranga, 6681/prédio 50, sala 703, CEP: 90619-900, Bairro Partenon, Porto Alegre – RS, e-mail:

cep@pucrs.br, de segunda a sexta-feira das 8h às 12h e das 13h30 às 17h. O Comitê de Ética é um órgão independente constituído de profissionais das diferentes áreas do conhecimento e membros da comunidade. Sua responsabilidade é garantir a proteção dos direitos, a segurança e o bem-estar dos participantes por meio da revisão e da aprovação do estudo, entre outras ações.

Ao aceitar este termo de consentimento você não abre mão de nenhum direito legal que teria de outra forma. Não aceite este termo de consentimento a menos que tenha tido a oportunidade de fazer perguntas e tenha recebido respostas satisfatórias para todas as suas dúvidas.

Se você concordar em participar deste estudo, você fará o aceite oralmente informando seu nome completo, e o áudio com a gravação da leitura e do aceite deste termo de consentimento serão enviados por e-mail. Ao ler e aceitar todas as páginas deste documento, você de forma voluntária e esclarecida, nos autoriza a utilizar todas as informações de natureza pessoal que constam nas respostas de sua entrevista, bem como as imagens da tela do computador durante a coleta de dados, para finalidade de pesquisa e realização deste estudo. Você receberá a gravação e um documento acessível deste termo para seus registros e os mesmos serão arquivados pelo responsável pelo estudo.

Eu, (nome completo do participante), após a leitura (ou a escuta da leitura) deste documento, e de ter tido a oportunidade de conversar com o pesquisador responsável para esclarecer todas as minhas dúvidas, acredito estar suficientemente informado, ficando claro para mim que minha participação é voluntária e que posso retirar este consentimento a qualquer momento sem penalidades ou perda de qualquer benefício. Estou ciente também dos objetivos da pesquisa, dos procedimentos aos quais serei submetido, dos possíveis danos ou riscos deles provenientes e da garantia de confidencialidade e esclarecimentos sempre que desejar.

Diante do exposto expresso minha concordância de espontânea vontade em participar deste estudo, autorizando o uso, compartilhamento e publicação dos meus dados e informações de natureza pessoal para essa finalidade específica.

_____

(Nome completo do participante da pesquisa ou de seu representante legal)

DECLARAÇÃO DO PROFISSIONAL QUE OBTEVE O CONSENTIMENTO

Expliquei integralmente este estudo ao participante. Na minha opinião e na opinião do participante, houve acesso suficiente às informações, incluindo riscos e benefícios, para que uma decisão consciente seja tomada.

Data: ____ de _____ de 2021

_____

Alessandra Helena Jandrey (Aluna PPGCC/PUCRS)

APPENDIX B – INTERVIEW GUIDE

Procedimento de introdução e apresentação da pesquisa:

(Nome do participante), queremos te agradecemos pela participação e disponibilidade. Eu vou me apresentar agora. Meu nome é Alessandra, sou mestranda em Ciência da Computação da PUCRS e atualmente venho desenvolvendo uma pesquisa que busca explorar descrições de imagens com pessoas que possuem baixa visão.

Procedimento de leitura e assinatura do Termo de Consentimento Livre e Esclarecido:

Eu lhe encaminharei por email o Termo de Consentimento Livre e Esclarecimento em um documento word acessível, por favor me confirme o recebimento. Como eu preciso do registro do teu consentimento, eu preciso realizar a leitura do Termo. Eu compartilharei minha tela com você e o sintetizador de voz do word realizará a leitura do mesmo Termo que você recebeu por email. Podemos começar?

Por questões éticas, precisamos do seu consentimento verbal para iniciarmos a entrevista. Preciso que você diga seu nome completo e se você concorda ou não em permitir o uso dos dados coletados nesta entrevista para nossa pesquisa.

Procedimento de coleta de dados:

Eu gostaria de coletar suas informações de perfil.

Nome:
Email (TCLE):
Idade:
Profissão:
Diagnóstico (CID):
Há quanto tempo possui deficiência visual:
O que enxerga:
Utiliza bengala ou cão guia:

1. Você poderia se apresentar para mim? Comece falando sobre você, onde você trabalha e o que você gosta de fazer no seu tempo livre.

Queremos entender como é seu contato com Tecnologias Assistivas.

2. Quais tecnologias você usa no seu dia a dia para lhe auxiliar como, por exemplo, os leitores de tela e aplicativos?

Queremos entender como é sua relação com elementos visuais e sua experiência com descritores de imagens.

3. Como é o seu contato com imagens compartilhadas em redes sociais ou em sites de notícias?
4. Quando você se depara com uma imagem na Internet e pede para alguém lhe descrever ela, o que você quer saber a respeito?
5. Como é a sua experiência com audiodescrição e legendas de imagens?
6. Qual é sua opinião sobre as descrições de imagens que as pessoas colocam, como o uso da hashtag #PraCegoVer?

Nós temos um conjunto de imagens, e caso você prefira, eu posso lhe falar quais são os objetos identificados automaticamente em cada imagem.

7. O que você considera importante nessa imagem, ou seja, o que lhe chama a atenção?

Nós temos 2 descrições para essa imagem. Você receberá uma descrição e lhe faremos uma pergunta. Depois, vamos lhe mostrar a segunda descrição e repetimos a pergunta. Podemos começar?

8. A descrição demonstrada é suficientemente clara, ou seja, é satisfatória para você entender a cena? Por quê?

Imagine que você pode ter uma descrição desse ambiente com o uso de qualquer Tecnologia que você goste.

9. O que você gostaria que esse descritor lhe falasse sobre essa imagem, ou seja, qual seria sua expectativa de descrição?

APPENDIX C – IMAGE DESCRIPTIONS RATING FOR SIGHTED PEOPLE

Qual é a sua área de atuação?	Qual é o seu nível de escolaridade?
professor	Pós-graduação completa
Engenharia de Software	Pós-graduação incompleta
EDUCAÇÃO	Superior completo
Matemática	Superior incompleto
Psicologia	Pós-graduação completa
TI	Superior incompleto
TI	Superior completo
Tecnologia da informação	Superior incompleto
TI	Pós-graduação incompleta
Educação	Pós-graduação completa
TI	Superior completo
Estudante	Superior incompleto
comunicação e design	Pós-graduação completa
TI- Desenvolvimento	Superior incompleto
Computação	Superior completo
Educação	Pós-graduação incompleta
Professor	Pós-graduação incompleta
Computação	Pós-graduação completa
Computação.	Superior incompleto
TI	Superior incompleto
TI	Pós-graduação completa
Engenharia de dados	Pós-graduação incompleta
Tecnologia de Informação	Superior incompleto
Educação superior	Pós-graduação completa
Educação	Pós-graduação completa
TI	Pós-graduação completa
Saúde	Superior incompleto
Engenharia de Software	Pós-graduação completa
Comunicação	Superior completo
Informática	Superior completo
Mestrando em Ciências da Computação	Pós-graduação incompleta
TI	Pós-graduação completa
TI	Pós-graduação completa
TI	Superior incompleto
Analista de sistemas - computação	Superior completo
Desenvolvimento de software	Pós-graduação incompleta
UX	Pós-graduação completa
Mestranda Ciência da Computação	Pós-graduação incompleta
RH	Superior incompleto
Tecnologia da Informação	Pós-graduação incompleta
Engenharia de Software	Superior completo
Educação	Pós-graduação incompleta
Ciência de dados e direito	Pós-graduação incompleta
Administrativo	Pós-graduação completa
Ciência da computação	Pós-graduação completa
Sistemas de Informação	Pós-graduação incompleta
Ciencia de dados	Pós-graduação completa
Física Médica	Pós-graduação incompleta
Percepção, Computação Gráfica, Visão Computacional	Pós-graduação completa
Ciência da Computação	Pós-graduação completa
Analise de Sistemas	Pós-graduação completa
Ciência da computação	Pós-graduação incompleta
Desenvolvimento de Software	Pós-graduação incompleta
UI Design	Pós-graduação incompleta
Nenhuma	Superior incompleto
TI	Superior incompleto
TI	Pós-graduação incompleta

RATING LIKERT SCALE	IMAGE ID 22411								IMAGE ID 24436								IMAGE ID 38336							
COUNT "1" LIKERT	18	9	20	2	3	2	1	2	6	30	30	0	4	4	3	13	25	5	5	2	3	3	11	3
COUNT "2" LIKERT	12	9	14	3	5	6	9	3	5	13	14	3	6	9	1	14	16	3	8	4	1	1	28	1
COUNT "3" LIKERT	16	14	14	7	14	7	16	17	24	12	11	6	12	15	5	19	11	10	19	10	10	8	14	13
COUNT "4" LIKERT	8	12	6	22	22	19	13	25	11	2	1	18	18	17	18	5	4	17	7	21	22	21	4	13
COUNT "5" LIKERT	3	13	3	23	13	23	18	10	11	0	1	30	17	12	30	6	1	22	18	20	21	24	0	27
Weighted Average	2.4	3.19	2.26	4.07	3.65	3.96	3.67	3.67	3.28	1.75	1.75	4.32	3.67	3.42	4.25	2.6	1.95	3.84	3.44	3.93	4	4.09	2.19	4.05

RATING LIKERT SCALE	IMAGE ID 125909								IMAGE ID 137724								IMAGE ID 153231							
COUNT "1" LIKERT	4	1	8	12	14	5	6	2	43	30	18	2	0	22	20	10	22	17	38	0	19	3	2	5
COUNT "2" LIKERT	6	5	5	9	7	4	5	1	9	19	12	4	1	11	13	7	8	16	13	0	14	5	4	2
COUNT "3" LIKERT	9	3	11	11	18	13	10	5	5	6	6	6	2	10	12	15	18	13	4	7	16	12	16	3
COUNT "4" LIKERT	26	9	21	7	11	25	25	11	0	2	10	14	12	6	6	18	4	6	1	22	4	18	23	6
COUNT "5" LIKERT	12	39	12	18	7	10	11	38	0	0	11	31	42	8	6	7	5	5	1	28	4	19	12	41
Weighted Average	3.63	4.4	3.42	3.18	2.82	3.54	3.53	4.44	1.33	1.65	2.72	4.19	4.67	2.42	2.39	3.09	2.33	2.4	1.49	4.37	2.3	3.79	3.68	4.33

RATING LIKERT SCALE	IMAGE ID 188239								IMAGE ID 259060								IMAGE ID 261843							
COUNT "1" LIKERT	12	9	9	0	8	12	1	0	11	9	17	1	0	3	11	8	2	5	7	0	0	2	2	7
COUNT "2" LIKERT	11	7	7	3	7	14	2	2	8	9	8	2	2	6	11	12	7	2	1	2	4	3	2	8
COUNT "3" LIKERT	22	9	11	10	11	19	10	1	12	11	11	7	8	6	12	16	11	13	14	4	7	8	9	20
COUNT "4" LIKERT	6	20	17	13	22	4	25	12	17	18	12	24	15	20	13	10	18	20	14	13	15	10	17	10
COUNT "5" LIKERT	6	12	13	31	9	8	19	42	9	10	9	23	32	22	10	11	19	17	21	38	31	34	27	12
Weighted Average	2.7	3.33	3.32	4.26	3.3	2.68	4.04	4.65	3.09	3.19	2.79	4.16	4.35	3.91	3	3.07	3.79	3.74	3.72	4.53	4.28	4.25	4.14	3.21

RATING LIKERT SCALE	IMAGE ID 358921								IMAGE ID 367095								IMAGE ID 429416							
COUNT "1" LIKERT	27	20	22	4	12	1	22	4	7	9	2	7	7	3	4	6	46	11	7	1	2	8	0	3
COUNT "2" LIKERT	15	10	8	8	10	0	8	3	6	7	3	8	6	7	3	6	9	19	7	6	7	14	3	5
COUNT "3" LIKERT	9	15	16	12	15	10	13	9	17	14	7	12	8	17	15	24	2	14	13	13	12	18	10	5
COUNT "4" LIKERT	6	10	8	20	13	15	12	18	21	21	12	18	14	13	22	12	0	10	22	17	15	9	24	21
COUNT "5" LIKERT	0	2	3	13	7	31	2	23	6	6	33	12	22	17	13	9	0	3	8	20	21	8	20	23
Weighted Average	1.89	2.37	2.33	3.53	2.88	4.32	2.37	3.93	3.23	3.14	4.25	3.35	3.67	3.6	3.65	3.21	1.23	2.56	3.3	3.86	3.81	2.91	4.07	3.98

RATING LIKERT SCALE	IMAGE ID 495243								IMAGE ID 499966								IMAGE ID 509718							
COUNT "1" LIKERT	6	37	5	3	7	17	3	3	3	34	31	0	4	2	1	0	10	12	10	4	4	2	2	3
COUNT "2" LIKERT	13	13	11	4	8	11	4	5	1	16	18	2	9	7	5	5	12	7	11	8	10	5	4	9
COUNT "3" LIKERT	22	5	15	19	13	19	8	8	14	6	7	4	19	18	13	10	21	20	16	10	19	17	16	17
COUNT "4" LIKERT	8	2	14	14	18	8	19	24	22	1	1	13	17	15	16	14	10	14	15	17	11	12	22	18
COUNT "5" LIKERT	8	0	12	17	11	2	23	17	17	0	0	38	8	15	22	28	4	4	5	18	13	21	13	10
Weighted Average	2.98	1.51	3.3	3.67	3.32	2.42	3.96	3.82	3.86	1.54	1.61	4.53	3.28	3.6	3.93	4.14	2.75	2.84	2.89	3.65	3.33	3.79	3.7	3.4

RATING LIKERT SCALE	IMAGE ID 69700								IMAGE ID 80714								
COUNT "1" LIKERT	24	25	38	1	1	0	5	9	4	5	40	4	9	2	0	5	
COUNT "2" LIKERT	20	14	13	6	1	2	4	9	7	5	6	4	7	6	0	1	
COUNT "3" LIKERT	8	13	4	11	3	3	17	15	12	11	5	14	22	7	7	8	
COUNT "4" LIKERT	4	3	2	21	15	14	14	7	15	16	4	17	12	11	22	19	
COUNT "5" LIKERT	1	2	0	18	37	38	17	17	19	20	2	18	7	31	28	24	
Weighted Average	1.91	2	1.47	3.86	4.51	4.54	3.6	3.25	3.67	3.7192982	1.63	3.72	3.02	4.11	4.37	3.98	

RATING LIKERT SCALE	IMAGE ID 154911								IMAGE ID 156497								
COUNT "1" LIKERT	28	5	10	0	11	6	6	3	1	5	9	14	7	1	2	11	
COUNT "2" LIKERT	10	9	10	6	27	7	8	2	6	3	2	11	6	3	6	8	
COUNT "3" LIKERT	14	13	14	11	13	16	22	10	20	7	6	20	10	11	13	14	
COUNT "4" LIKERT	5	17	16	21	4	12	11	16	17	27	6	11	20	16	18	15	
COUNT "5" LIKERT	0	13	7	19	2	16	10	26	13	15	34	1	14	26	18	9	
Weighted Average	1.93	3.42	3	3.93	2.28	3.44	3.19	4.05	3.61	3.7719298	3.95	2.54	3.49	4.11	3.77	3.05	

RATING LIKERT SCALE	IMAGE ID 332158								IMAGE ID 355440								
COUNT "1" LIKERT	33	32	7	0	5	0	2	1	17	6	7	15	3	0	1	0	
COUNT "2" LIKERT	7	8	5	6	10	1	5	1	12	7	7	21	4	1	6	3	
COUNT "3" LIKERT	13	12	24	8	22	14	17	6	20	14	15	16	13	11	13	15	
COUNT "4" LIKERT	3	5	11	16	8	22	19	16	7	21	18	4	17	20	15	16	
COUNT "5" LIKERT	1	0	10	27	12	20	14	33	1	9	10	1	20	25	22	23	
Weighted Average	1.81	1.82	3.21	4.12	3.21	4.07	3.67	4.39	2.35	3.3508772	3.3	2.21	3.82	4.21	3.89	4.04	

RATING LIKERT SCALE	IMAGE ID 464831								IMAGE ID 482728								
COUNT "1" LIKERT	10	4	5	7	1	2	1	1	4	6	10	2	7	2	2	14	
COUNT "2" LIKERT	1	5	2	4	1	11	7	4	5	5	11	2	12	2	7	11	
COUNT "3" LIKERT	15	15	16	12	3	13	11	18	10	7	23	2	19	5	15	17	
COUNT "4" LIKERT	19	22	17	22	12	22	21	15	23	21	7	11	14	26	22	9	
COUNT "5" LIKERT	12	11	17	12	40	9	17	19	15	18	6	40	5	22	11	6	
Weighted Average	3.39	3.54	3.68	3.49	4.56	3.44	3.81	3.82	3.7	3.7017544	2.79	4.49	2.96	4.12	3.58	2.68	
										SELECTED							

RATING LIKERT SCALE	IMAGE ID 528018								IMAGE ID 572173								
COUNT "1" LIKERT	9	8	6	1	3	34	7	0	6	16	15	3	2	2	3	9	
COUNT "2" LIKERT	17	15	11	1	3	8	19	2	9	9	11	4	3	2	4	12	
COUNT "3" LIKERT	15	16	20	6	8	12	18	11	12	17	17	10	17	12	13	19	
COUNT "4" LIKERT	9	12	7	15	20	2	3	15	15	6	6	19	24	13	11	6	
COUNT "5" LIKERT	7	6	13	34	23	1	10	29	15	9	8	21	11	28	26	11	
Weighted Average	2.79	2.88	3.18	4.4	4	1.74	2.82	4.25	3.42	2.7017544	2.67	3.89	3.68	4.11	3.93	2.96	

APPENDIX D – IMAGES USED IN THE INTERVIEW STUDY AND THEIR RESPECTIVE AUTOMATIC AND HUMAN-GENERATED DESCRIPTIONS

Images	Descriptions
	(A) Um grupo de pessoas sentadas ao redor de uma mesa comendo pizza. (H) Alguns alunos ou amigos estão comendo pizza em uma biblioteca.
	(A) Um grupo de pessoas sentadas ao redor de uma mesa. (H) Várias pessoas ao redor de uma mesa com cartas de baralho.
	(A) Um grupo de pessoas sentadas em mesas trabalhando em laptops. (H) Em uma sala de aula na Índia, os alunos fazem seus trabalhos em computadores.

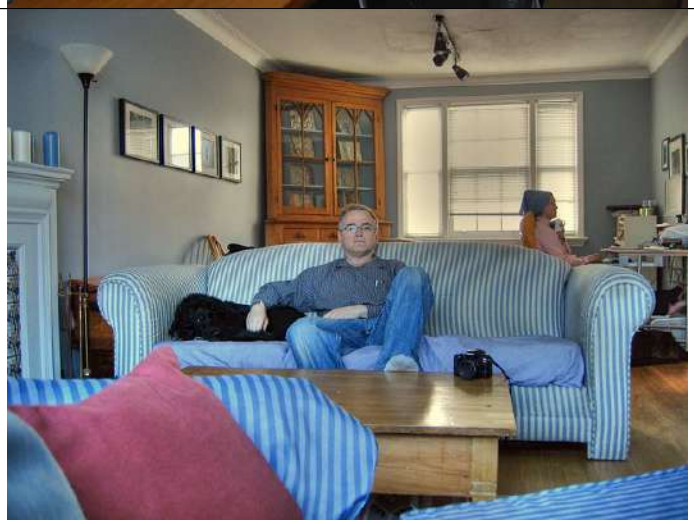
Images	Descriptions
	(A) Um grupo de homens em pé próximos um do outro em uma sala. (H) Dois homens apertam as mãos enquanto um deles dá um presente.
	(A) Uma mulher sentada à mesa com um computador portátil. (H) Uma mulher está fazendo uma apresentação com um laptop.
	(A) Um grupo de pessoas sentadas ao redor de uma mesa comendo pizza. (H) Um grupo de pessoas está comendo pizza em uma mesa comprida.
	(A) Um grupo de pessoas sentadas à mesa com laptops. (H) Pessoas sentadas em diferentes mesas trabalhando em laptops.

Images	Descriptions
	(A) Um grupo de pessoas sentadas ao redor de uma mesa com comida. (H) Uma mulher com dois filhos assistindo TV enquanto come.
	(A) Um grupo de pessoas sentadas à mesa com laptops. (H) Dois homens e uma mulher sentados próximos um do outro usando computadores portáteis.
	(A) Um homem sentado em um sofá com um laptop no colo. (H) Um homem está sentado em um sofá numa sala de estar com uma mesa de centro que tem um laptop aberto.

Images	Descriptions
	(A) Um grupo de pessoas sentadas ao redor de uma mesa com laptops. (H) Uma família sentada à mesa da cozinha com alguns laptops e um pouco de café.
	(A) Um grupo de pessoas sentadas ao redor de uma sala de estar juntas. (H) Uma mulher bebendo sentada ao lado de dois homens enquanto assiste à televisão.
	(A) Um grupo de pessoas sentadas ao redor de uma mesa de jantar. (H) Idosos sentados juntos, desfrutando do jantar e do vinho.

Images	Descriptions
	(A) Um grupo de pessoas sentadas ao redor de uma mesa em uma sala. (H) Um grupo de pessoas trabalhando em alguns eletrônicos em uma mesa.
	(A) Um grupo de pessoas sentadas ao redor de uma mesa com um bolo. (H) Uma família comemorando um aniversário em volta de um bolo.
	(A) Um homem sentado em uma escrivaninha com um laptop e um computador. (H) Adultos estão trabalhando sentados em suas mesas em um escritório aberto.

Images	Descriptions
	(A) Um grupo de pessoas em pé ao redor de uma mesa com taças de vinho. (H) Um grupo de pessoas estão ao lado de uma mesa cheia de bebidas e de um garçom.
	(A) Um grupo de pessoas sentadas à mesa com laptops. (H) Pessoas com laptops sentadas em um semicírculo ao redor de uma sala de conferências.
	(A) Um grupo de pessoas sentadas ao redor de uma mesa com taças de vinho. (H) Uma mesa em um restaurante com várias pessoas mais velhas sentadas e um homem está servindo uma garrafa de vinho em uma taça.

Images	Descriptions
	(A) Um grupo de pessoas em pé ao redor de uma mesa com comida. (H) Pessoas de um escritório reunidas comendo diferentes tipos de bolo.
	(A) Um homem sentado no sofá em uma sala de estar. (H) Um homem sentado no sofá enquanto uma mulher está sentada em uma mesa de computador atrás dele.
	(A) Um grupo de pessoas sentadas ao redor de uma mesa com laptops. (H) Pessoas estão na biblioteca trabalhando e estudando.

Images	Descriptions
	(A) Um homem sentado em uma cadeira diante de um computador. (H) Um homem em sua escrivaninha cercado por papéis está segurando o seu gato.
	(A) Um homem sentado em uma cadeira diante de um computador. (H) Um homem em sua escrivaninha cercado por papéis está segurando o seu gato.

Table D.1: Images used in the interview study with their respective automatic and human-generated image descriptions.

APPENDIX E – LOW VISION PARTICIPANTS' SATISFACTION AND DISSATISFACTION REASONS IN IMAGE DESCRIPTIONS

REASONS	DISSATISFACTION (AUTOMATIC)	SATISFACTION (AUTOMATIC)	DISSATISFACTION (HUMAN)	SATISFACTION (HUMAN)
ACTION (PERSON)	16.96%	10.00%	11.67%	15.76%
AGE GROUP (PERSON)	5.26%	0.00%	4.17%	1.97%
CHARACTERISTIC (OBJECT)	0.00%	0.00%	0.00%	1.48%
CLOTHING (PERSON)	2.34%	0.00%	3.33%	0.00%
DISH'S NAME AND BEVERAGES (FOOD)	0.58%	7.50%	0.00%	9.36%
FACIAL AND BODY EXPRESSIONS (PERSON)	0.58%	0.00%	4.17%	0.00%
ORNAMENTATION (SETTING)	10.53%	0.00%	16.67%	0.49%
GENDER (PERSON)	4.09%	15.00%	4.17%	14.29%
INTERACTION (PERSON)	4.68%	6.25%	0.83%	0.49%
LOCATION (OBJECT)	1.00%	0.00%	1.67%	0.00%
NAME (OBJECT)	5.26%	18.75%	5.83%	9.85%
NATURE (EVENT)	1.17%	0.00%	1.67%	0.00%
NUMBER (PERSON)	16.96%	7.50%	14.17%	8.37%
OCCASION (EVENT)	5.26%	0.00%	4.17%	2.46%
FUR/COAT COLORS (ANIMAL)	0.00%	0.00%	0.00%	1.48%
PLACE (SETTING)	18.13%	2.50%	15.00%	10.84%
POSITION (PERSON)	4.09%	32.50%	4.17%	11.82%
PROFESSION (PERSON)	0.00%	0.00%	0.00%	1.97%
NUMBER (OBJECT)	0.00%	0.00%	2.50%	0.00%
RACE/ETHNICITY (PERSON)	0.58%	0.00%	0.00%	0.00%
RELATIONSHIP (PERSON)	2.34%	0.00%	5.00%	6.40%
SIZE/DIMENSION (SETTING)	1.17%	0.00%	0.83%	0.99%

APPENDIX F – LOW VISION PARTICIPANTS' EXPECTATIONS IN IMAGE DESCRIPTIONS

CODES/IMAGES ID	22411	24436	38336	69700	80714	125909	137724	153231	154911	156497	188239	259060	261843	332158	355440	358921	367095	429416	464831	482728	495243	499966	509718	528018	572173	TOTAL
FOOD							**1**					**3**	**2**		**1**	**3**									**1**	**11**
QUALITIES																1										1
DISH'S NAME AND BEVERAGES (FOOD)							1					3	2		1	2									1	10
SETTING	**6**	**7**	**2**	**3**	**3**	**5**	**6**	**3**	**7**	**5**	**3**	**5**	**5**	**5**	**3**	**4**	**5**	**4**	**2**	**1**	**2**	**3**	**3**	**7**	**5**	**104**
SIZE/DIMENSION (SETTING)	1				1	1			1					1					1				1	1	1	9
PLACE (SETTING)	1	3		1			2	2	4	1	2	2	2	2	1	2	2	2				1	1	1	1	33
ORNAMENTATION (SETTING)	4	4	2	2	2	4	4	1	2	4	1	3	3	2	2	2	3	2	1	1	2	2	1	5	3	62
ANIMAL																								**2**		**2**
FUR/COAT COLORS (ANIMAL)																								2		2
EVENT	**2**	**3**		**3**	**3**	**3**	**1**			**1**		**2**	**2**		**2**	**2**		**1**	**1**	**2**						**28**
NATURE (EVENT)		2		3		1	1			1		1							1	1						11
OCCASION (EVENT)	2	1			2	2						1	2		2	2		1		1						16
EMBEDDED TEXTS IN IMAGES					1																					1
OBJECT	**3**	**2**	**2**	**1**	**3**	**4**	**1**	**1**	**1**	**4**		**4**	**2**	**4**	**2**	**2**	**2**		**2**	**2**	**1**	**4**	**1**	**2**	**2**	**52**
CHARACTERISTIC (OBJECT)	1	1		1		1			1				1		2				1	2		2				13
NAME (OBJECT)	2	1	1			2	1	1		2		2	1	3		2					1	1		2	2	24
LOCATION (OBJECT)			1		2	1				1		2		1			1					1				10
NUMBER (OBJECT)					1					1							1		1				1			5
PERSON	**7**	**7**	**12**	**9**	**14**	**8**	**8**	**10**	**10**	**7**	**12**	**10**	**5**	**11**	**11**	**12**	**11**	**11**	**9**	**4**	**8**	**6**	**15**	**9**	**11**	**237**
ACTION (PERSON)	1	1	1		1		2	4	3		1	1	1		1	1		2	2			1	3	1		27
HAIRSTYLE (PERSON)	1							1				1														3
"HOW PEOPLE LOOK LIKE"											1				1									1	1	4
FACIAL AND BODY EXPRESSIONS (PERSON)		1	1	2	2			1	2	1	2	1				1	2	1			1	1			1	20
AGE GROUP (PERSON)	1	2	1	3	1	2	1	1		1	2	2		2		2		1	1		1		2	2	3	31
GENDER (PERSON)	2	1	2		1	2	1	1	1		1			1	2	2	1	1	3		2		1		1	26
INTERACTION (PERSON)									1	2	2	1		4		1	1					2		3		17
POSITION (PERSON)		1	2	1	3		1	1	1			1	1		1		1					1	1	1	1	18
PROFESSION (PERSON)			1																							1
NUMBER (PERSON)	1		3	2	2	3	2		1	1		1	3	4	2	3	3	3	3	2	1		4		3	47
RACE/ETHNICITY (PERSON)		1			2			1			1	1			1						1		1			9
RELATIONSHIP (PERSON)											1				1											2
CLOTHING (PERSON)	1		1	1	2	1	1		1	2	1	1			2	2	3	3		2	2	1	3	1	1	32
TOTAL	**18**	**19**	**16**	**16**	**22**	**20**	**17**	**14**	**18**	**17**	**15**	**24**	**16**	**20**	**19**	**23**	**18**	**16**	**14**	**9**	**11**	**13**	**19**	**20**	**19**	**433**

NUMBER OF PARTICIPANTS 8
NUMBER OF EVALUATIONS 120

	22411	24436	38336	69700	80714	125909	137724	153231	154911	156497	188239	259060	261843	332158	355440	358921	367095	429416	464831	482728	495243	499966	509718	528018	572173	%of evaluations that "cited" each code	Number of images that cited each code	% of images
EVALUATED BY X NUMBER OF PARTIC.	5	5	5	5	5	5	5	5	5	5	4	5	5	5	6	5	5	4	4	4	4	4	5	5	5			
FOOD	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	60%	40%	0%	17%	60%	0%	0%	0%	0%	0%	0%	0%	0%	20%			
QUALITIES	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0.83%	1	4.00%
DISH'S NAME AND BEVERAGES (FOOD)	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	0%	60%	40%	0%	17%	40%	0%	0%	0%	0%	0%	0%	0%	0%	20%	8.33%	6	24.00%
SETTING	120%	140%	40%	60%	60%	100%	120%	60%	140%	100%	75%	100%	100%	100%	50%	80%	100%	100%	50%	25%	50%	75%	60%	140%	100%			
SIZE/DIMENSION (SETTING)	20%	0%	0%	0%	20%	20%	0%	0%	20%	0%	0%	0%	0%	20%	0%	0%	0%	0%	25%	0%	0%	0%	20%	20%	20%	7.50%	9	36.00%
PLACE (SETTING)	20%	60%	0%	20%	0%	0%	40%	40%	80%	20%	50%	40%	40%	40%	17%	40%	40%	50%	0%	0%	0%	25%	20%	20%	20%	27.50%	19	76.00%
ORNAMENTATION (SETTING)	80%	80%	40%	40%	40%	80%	80%	20%	40%	80%	25%	60%	60%	40%	33%	40%	60%	50%	25%	25%	50%	50%	20%	100%	60%	51.67%	25	100.00%
ANIMAL	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	40%	0%			
FUR/COAT COLORS (ANIMAL)	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	40%	0%	1.67%	1	4.00%
EVENT	40%	60%	0%	60%	60%	60%	20%	0%	0%	20%	0%	40%	40%	0%	33%	40%	0%	25%	25%	50%	0%	0%	0%	0%	0%		14	
NATURE (EVENT)	0%	40%	0%	60%	0%	20%	20%	0%	0%	20%	0%	20%	0%	0%	0%	0%	0%	0%	25%	25%	0%	0%	0%	0%	0%	9.17%	8	32.00%
OCCASION (EVENT)	40%	20%	0%	0%	40%	40%	0%	0%	0%	0%	0%	20%	40%	0%	33%	40%	0%	25%	0%	25%	0%	0%	0%	0%	0%	13.33%	10	40.00%
EMBEDDED TEXTS IN IMAGES	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0.83%	1	4.00%
OBJECT	60%	40%	40%	20%	60%	80%	20%	20%	20%	80%	0%	80%	40%	80%	33%	40%	40%	0%	50%	50%	25%	100%	20%	40%	40%			
CHARACTERISTIC (OBJECT)	20%	20%	0%	20%	0%	20%	0%	0%	20%	0%	0%	0%	20%	0%	33%	0%	0%	0%	25%	50%	0%	50%	0%	0%	0%	10.83%	10	40.00%
NAME (OBJECT)	40%	20%	20%	0%	0%	40%	20%	20%	0%	40%	0%	40%	20%	60%	0%	40%	0%	0%	0%	0%	25%	25%	0%	40%	40%	20.00%	15	60.00%
LOCATION (OBJECT)	0%	0%	20%	0%	40%	20%	0%	0%	0%	20%	0%	40%	0%	20%	0%	0%	20%	0%	0%	0%	0%	25%	0%	0%	0%	8.33%	8	32.00%
NUMBER (OBJECT)	0%	0%	0%	0%	20%	0%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	20%	0%	25%	0%	0%	0%	20%	0%	0%	4.17%	5	20.00%
PERSON	140%	140%	240%	180%	280%	160%	160%	200%	200%	140%	300%	200%	100%	220%	183%	240%	220%	275%	225%	100%	200%	150%	300%	180%	220%			
ACTION (PERSON)	20%	20%	20%	0%	20%	0%	40%	80%	60%	0%	25%	20%	20%	0%	17%	20%	0%	50%	50%	0%	0%	25%	60%	20%	0%	22.50%	17	68.00%
HAIRSTYLE (PERSON)	20%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2.50%	3	12.00%
"HOW PEOPLE LOOK LIKE"	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	17%	0%	0%	0%	0%	0%	0%	0%	0%	20%	20%	3.33%	4	16.00%
FACIAL AND BODY EXPRESSIONS (PERSON)	0%	20%	20%	40%	40%	0%	0%	20%	40%	20%	50%	20%	0%	0%	0%	20%	40%	25%	0%	0%	25%	25%	0%	0%	20%	16.67%	15	60.00%
AGE GROUP (PERSON)	20%	40%	20%	60%	20%	40%	20%	20%	0%	20%	50%	40%	0%	40%	0%	40%	40%	0%	25%	25%	0%	25%	0%	40%	40%	60%	25.83%	19	76.00%
GENDER (PERSON)	40%	20%	40%	0%	20%	40%	20%	20%	20%	0%	25%	0%	0%	20%	33%	40%	20%	25%	75%	0%	50%	0%	20%	0%	20%	21.67%	18	72.00%
INTERACTION (PERSON)	0%	0%	0%	0%	0%	0%	0%	0%	20%	40%	50%	20%	0%	80%	0%	20%	20%	0%	0%	0%	0%	50%	0%	60%	0%	14.17%	9	36.00%
POSITION (PERSON)	0%	20%	40%	20%	60%	0%	20%	20%	20%	0%	0%	20%	20%	0%	17%	0%	20%	0%	0%	0%	0%	25%	20%	20%	20%	15.00%	15	60.00%
PROFESSION (PERSON)	0%	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0.83%	1	4.00%
NUMBER (PERSON)	20%	0%	60%	40%	40%	60%	40%	0%	20%	20%	0%	20%	60%	80%	33%	60%	60%	75%	75%	50%	25%	0%	80%	0%	60%	39.17%	20	80.00%
RACE/ETHNICITY (PERSON)	0%	20%	0%	0%	40%	0%	0%	20%	0%	0%	25%	20%	0%	0%	17%	0%	0%	0%	0%	0%	25%	0%	20%	0%	0%	7.50%	8	32.00%
RELATIONSHIP (PERSON)	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	25%	0%	0%	0%	17%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1.67%	2	8.00%
CLOTHING (PERSON)	20%	0%	20%	20%	40%	20%	20%	0%	20%	40%	25%	20%	0%	0%	33%	40%	60%	75%	0%	50%	50%	25%	60%	20%	20%	26.67%	20	80.00%

ATTACHMENT A – PROJECT'S APPROVAL OPINION GENERATED BY THE RESEARCH ETHICS COMMITTEE

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL - PUC/RS

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Um estudo qualitativo sobre a percepção de descrições de imagens por deficientes visuais

Pesquisador: DUNCAN DUBUGRAS ALCOBA RUIZ

Área Temática:

Versão: 1

CAAE: 42963621.0.0000.5336

Instituição Proponente: UNIÃO BRASILEIRA DE EDUCAÇÃO E ASSISTENCIA

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.575.486

Apresentação do Projeto:

As informações elencadas nos campos "Apresentação do Projeto", "Objetivo da Pesquisa" e "Avaliação dos Riscos e Benefícios" foram retiradas do arquivo Informações Básicas da Pesquisa (PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1690762.pdf, de 22/01/2021) e/ou do Projeto Detalhado. Conforme descrito pelo pesquisador, "Os avanços tecnológicos mudaram significativamente a forma de se consumir informação. Na última década, ocorreu um aumento exponencial de conteúdos produzidos, em especial os multimídias como imagens e vídeos (López-Sánchez, Arrieta e Corchado, 2018). Esses conteúdos devem ser acessíveis para todas as pessoas, incluindo aquelas que possuem algum impedimento físico, como é o caso dos deficientes visuais. Os leitores de tela são um meio de promover o acesso e a inclusão digital desse público e consistem em ler em voz alta as informações textuais disponíveis (Morris et al., 2018). Para leitura de imagens, os leitores de tela dependem que elas contenham textos alternativos. Um texto alternativo, chamado também de descrição, é uma recomendação das diretrizes de acessibilidade e refere-se a transcrever em palavras o conteúdo visual da imagem (Sacramento et al., 2020).Diversas subáreas da Inteligência Artificial propõem-se a oferecer soluções para a geração de descrições automáticas. Tais soluções buscam identificar os elementos da imagem, como as pessoas e os objetos presentes, para extrair informações e gerar uma sentença em linguagem natural, como o português (Bernardi et al., 2018). O principal desafio de construir descritores automáticos está em

Endereço: Av.Ipiranga, 6681, prédio 50, sala 703
Bairro: Partenon **CEP:** 90.619-900
UF: RS **Município:** PORTO ALEGRE
Telefone: (51)3320-3345 **Fax:** (51)3320-3345 **E-mail:** cep@pucrs.br

projetar um modelo que possa gerar descrições semelhantes às humanas (Liu et al., 2018). Apesar de todo o avanço que a Inteligência Artificial apresentou nas últimas décadas, as descrições automáticas são imperfeitas (Macleod et al., 2017). Humanos costumam ter facilidade em descrever uma imagem ou uma cena, oferecendo maior nível de detalhamento (Elliott e Keller, 2013). Entretanto, as descrições humanas podem facilmente tornar-se longas, o que acarreta excesso de informação e cansaço às pessoas que dependem de leitores de tela para consumir a informação.No caso dos deficientes visuais, que dependem dessas descrições para consumir informação visual, descrições detalhadas aparentemente são sempre ideais (Stangl, Morris e Gurari, 2018). Todavia, as preferências, pensamentos e comportamentos das pessoas dependem de seus objetivos que, por sua vez, mudam de acordo com suas experiências e os diferentes contextos vivenciados (Warren, McGraw e Van Boven, 2011). Para alguns deficientes visuais, disponibilizar o máximo de informação é uma questão de justiça e inclusão, para outros, informações demais podem ser distrativas, tediosas e até mesmo inúteis (Stangl, Morris e Gurari, 2018). Dessa forma, é necessária uma investigação das perspectivas dos deficientes visuais, buscando compreender o que estes desejam saber sobre os diferentes contextos em que eles podem estar inseridos. Portanto, o objetivo principal dessa pesquisa é identificar quais informações são importantes em descrições de imagens para os deficientes visuais em diferentes contextos.

As perguntas que motivam o estudo são: Quando uma descrição automática ou humana é considerada satisfatória e por quê? O contexto influencia na quantidade de informação necessária? De que forma as descrições ajudam os deficientes visuais a entender uma imagem?

A presente pesquisa qualitativa busca investigar quando e o porquê uma descrição é satisfatória ou não, e quais informações são julgadas importantes ao considerar diferentes contextos. Além disso, buscamos entender como as descrições ajudam os deficientes visuais na compressão de imagem, e quais são suas expectativas e opiniões sobre as descrições. Assim, foram estudados os conceitos técnicos sobre a geração de descrição automática, quais as arquiteturas e os conjuntos de dados utilizados, e quais as métricas quantitativas de avaliação. Essa etapa nos permitiu entender como as descrições automáticas são produzidas e avaliadas por computadores. Foram coletados trabalhos que buscaram analisar as experiências de deficientes visuais com descrições automáticas, assim como trabalhos que investigaram descrições produzidas por humanos e como vem ocorrendo sua disseminação. Além disso, foi investigado, na literatura, quais são os procedimentos e técnicas utilizadas em pesquisas qualitativas com participantes deficientes visuais. Esta etapa nos forneceu conhecimento sobre como conduzir a pesquisa com esse público

e possibilitou definir os protocolos necessários. As etapas previstas para esta pesquisa são:1)Fundamentação teórica: buscamos compreender os geradores de descrição automáticos, identificando as abordagens de aprendizado de máquina, quais os conjuntos de dados utilizados no treinamento das redes neurais, e quais as métricas utilizadas para avaliação quantitativa. Com isso, foram definidas as fontes de dados necessárias para a realização do trabalho, isto é, as imagens e as descrições automáticas. Além disso, buscamos trabalhos que avaliaram a experiência de deficientes visuais com descrições automáticas.2) Estudo sobre protocolos e condutas qualitativas com deficientes visuais: buscamos entender como conduzir um estudo qualitativo com a participação de deficientes visuais. Portanto, foram levantados trabalhos que tiveram esse público em particular e que especificaram a conduta utilizada, o número de participantes, técnica de coleta de dados e demais procedimentos. Com isso, pudemos definir nosso protocolo, que se dará por meio de entrevistas com 5 a 10 participantes.3) Desenvolvimento da técnica de coleta de dados: uma vez identificada a conduta, elaboramos um roteiro para a entrevista semiestruturada que nos permitirá executar a pesquisa e coletar os dados necessários para posterior análise.4) Execução da entrevista com os participantes: com o roteiro da entrevista definido, iremos aplicá-la ao público alvo da pesquisa, a fim de identificar quais informações são importantes para os deficientes visuais em diferentes contextos, coletar suas expectativas de descrições e como elas ajudam na compreensão de imagem, e entender a relação entre contexto e quantidade de informação necessária. As entrevistas serão realizadas durante a performance das tarefas. Acreditamos que os resultados que serão obtidos podem contribuir para que os modelos de descrições automáticas considerem as necessidades dos deficientes visuais, auxiliando também pessoas com visão normal a construírem suas descrições de imagens considerando as perspectivas e as opiniões desse público".

Objetivo da Pesquisa:

O objetivo primário proposto pelos pesquisadores é o seguinte: "Identificar quais informações são importantes em descrições de imagens para os deficientes visuais em diferentes contextos.". O objetivo secundário é: "Identificar como as descrições auxiliam os deficientes visuais na compreensão de imagem.Coletar as expectativas de descrição dos deficientes visuais, bem como opiniões a respeito de descrições atuais.Identificar quando e por que uma descrição é satisfatória, ao considerar os diferentes contextos em que os deficientes visuais pode estar inserido."

Avaliação dos Riscos e Benefícios:

Os pesquisadores citam os seguintes riscos: “De acordo com o Conselho Nacional de Saúde (2002), qualquer pesquisa que envolva seres humanos pode ter um risco associado, na possibilidade de danos à dimensão física, psíquica, moral, intelectual, social, cultural ou espiritual do ser humano. Desta forma, conseguimos listar alguns destes riscos de acordo com o formato da dinâmica: 1. Dor de cabeça e cansaço ou aborrecimento durante a realização das tarefas ou durante as discussões; 2. Desconforto, constrangimento ou alterações de comportamento durante gravações de áudio e vídeo; 3. Divulgação de dados confidenciais ou quebra de sigilo. Para mitigar esses riscos, para o item 1 estamos elaborando um roteiro semiestruturado de questões e tarefas para orientar a dinâmica da entrevista, permitindo otimizar o tempo dos participantes e focar em questões relevantes ao nosso objeto de estudo. Para o item 2, as possíveis gravações de vídeo irão capturar somente o áudio e/ou tela do computador utilizado pelos participantes, assim não haverá a imagem física do mesmo associado ao seu desempenho durante a realização das atividades a serem propostas. Além disso, para a gravação de áudio deixaremos claro que o registro servirá apenas para o pesquisador revisar se conseguiu capturar todas as informações importantes durante a entrevista, posteriormente ao momento da conversa. Essa é uma ação que visa também reduzir o tempo da dinâmica das entrevistas que seria utilizado para transcrever detalhes das respostas. Em relação ao item 3, buscaremos desassociar o material coletado do participante respondente colocando etiquetas de controle para garantir o anonimato e ter um cuidado redobrado com o armazenamento de todos os dados coletados. Está prevista, também, a possibilidade de intervalos a cada 30 minutos, caso seja necessário.Por fim, conforme destacado no Termo de Consentimento Livre e Esclarecido, o qual será discutido com o participante antes de a atividade de coleta de dados (entrevistas) e tarefas iniciarem, o participante é livre para se retirar da atividade a qualquer momento, sem haver necessidade de explicitar razões para isso.”. Os pesquisadores citam os seguintes benefícios: “Não há benefícios a curto prazo para os participantes dessa pesquisa, contudo, ao término desse estudo é esperada a seguinte contribuição: identificação de quais informações são importantes para os deficientes visuais considerando os diferentes contextos; quando cada descrição é satisfatória e como elas ajudam os deficientes visuais na compreensão de imagens, podendo assim apoiar futuros pesquisadores de descritores automáticos e também pessoas com visão normal, para que as perspectivas e as opiniões dos deficientes visuais sejam consideradas.”.

Comentários e Considerações sobre a Pesquisa:

Este é um estudo nacional que fará uso de técnicas de entrevistas semiestruturada com 5 a 10

deficientes visuais com baixa visão ou cegueira parcial, maiores de idade que estejam inseridos no mercado de trabalho e utilizem leitor de tela. Tem caráter acadêmico, realizado para obtenção do título de mestre em Ciência da Computação. O cronograma de execução prevê que o recrutamento, as entrevistas e a análise de resultado serão feitas em março de 2021 após aprovação do Comitê de Ética em Pesquisa.

Considerações sobre os Termos de apresentação obrigatória:
Os termos de apresentação obrigatória estão adequados. No entanto, vide o campo "Recomendações".

Recomendações:
É preciso colocar no TCLE o procedimento que será adotado para assinatura e arquivamento do TCLE, esclarecendo como será feita a leitura e o aceite. Neste caso, é preciso alterar a frase “Se você concordar em participar deste estudo, você rubricará...”, para informar se será feita a leitura e o aceite oralmente e o áudio com a gravação será enviado por e-mail, ou se o TCLE será impresso em Braille e enviado para registro pelos correios (lembrando que, neste caso, haverá a necessidade de envio do endereço residencial, uma informação sensível a mais que deverá ser coletada).
Além disso, devido a um ofício circular recebido da CONEP com orientações para pesquisas realizadas em ambientes virtuais, é preciso incluir o seguinte parágrafo no TCLE: “Além dos desconfortos que você possa sentir em virtude das respostas a este questionário, é possível que, infelizmente, sua conexão falhe ou apresenta certa lentidão ou que você tenha dúvidas em como salvar suas respostas. O seu questionário será salvo automaticamente, não se preocupe OU você precisa clicar em XXXX para salvar adequadamente suas respostas. Nestes casos, não hesite em contatar o pesquisador (REPETIR CONTATOS).”.
Por fim, alterar o texto “...constam em nas respostas de...” para “...constam nas respostas de...”.

Conclusões ou Pendências e Lista de Inadequações:
Aprovação com recomendação.

Considerações Finais a critério do CEP:
Diante do exposto, o CEP-PUCRS, de acordo com suas atribuições definidas na Resolução CNS n° 466 de 2012, Resolução n° 510 de 2016 e a Norma Operacional n° 001 de 2013 do CNS, manifesta-se pela aprovação do projeto de pesquisa Um estudo qualitativo sobre a percepção de descrições de imagens por deficientes visuais proposto pelo pesquisador DUNCAN DUBUGRAS ALCOBA RUIZ com número de CAAE 42963621.0.0000.5336.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br