Describing Images to Visually Impaired Users: a Requirement Elicitation Approach

Master thesis

Yuxin Chu | May 2021

Master Thesis

Delft University of Technology
Industrial Design Engineering
MSc. Design for Interaction

Author

Yuxin Chu

Supervisory Team

Chair
Alessandro Bozzon
Department: SDE
Secton: KIND

Mentor
Himanshu Verma
Jeff Love
Department: SDE
Secton: KIND

Thanks to:
Ted van der Togt (KB)
Anne Bottenheft (Dedicon)
Anneke Wijtvliet (Dedicon)
Koen Krikhaar (Dedicon)

Abstract

Visually impaired people should enjoy the same rights to acquire information as people with normal sight.
Since visual contents become more and more pervasive in our daily life, image description becomes increasingly important to help visually impaired people to get equal access to the information contained in visual contents. However, how to produce image descriptions in a scalable and reliable way is still an unsolved problem. Therefore, researches on the requirements of image description from the perspective of visually impaired people are essential to approaching this problem.

Based on a review of existing study results on this topic, this thesis investigates the possibilities of utilizing interactive image description as an approach to collect visually impaired people's requirements on image description and the benefits of integrating interactive image description to the current image description production system.

The existing one-shot static description requires describers to evaluate the importance of the image, make choices on what should be described, and organize the content so that necessary information can be effectively conveyed. Through literature review, it is found that the requirements of image description are highly context-dependent and influenced by plentiful factors [1,2,3,6,9]. Therefore, existing guidelines are usually vague and require the describer to rely on experience and intuition while making a lot of subjective judgments, which increases the threshold for generating high-quality image descriptions.

On the other hand, through field research and literature research it is found that VIPs hope to have more control over the presentation of image description (both its presence and content). Early explorations of interactive image description showed the possibility of this affordance [9]. Since users are allowed to decide the description content actively, it is argued that the user's preference for image description can be collected through interactive image description. A design goal is proposed accordingly.

A prototype is developed to verify this proposal. Through a comparative experiment, the systems' function to collect user preferences and gradually improve the content of image description is confirmed. In addition, the qualitative research results also reveal the mental activities when users interacting with image description and the impact of interactive image description in this procedure, which is summarized as an image perception model. It is also argued that structured description and progressive description provide new perspectives to reduce the workload of describing images. A final design was developed as the demonstrator for the research findings and proposals.

Contents

0 Abstract 3

1 About this project 6

1.1 Introduction 7
1.2 Research Goal 8
1.3 Approach & Process 8

2 Background 12

2.1 Introduction of visually impaired people (VIPs) 14

2.1.1 Definition 14
2.1.2 Impact of visual impairment 15
2.1.3 Distribution of visual impairment 16

2.2 Image description for publications 18

2.2.1 Definition of image description 18
2.2.2 Demand for ID 20

2.3 Koninklijke Bibliotheek and Dedicon 21

2.4 Ways to generate IDs 23

2.4.1 The lack of first-party ID 23
2.4.2 AI image captioning 23
2.4.3 Human in the loop (HITL) 25

3 Requirements on image description 30

3.1 Introduction 32

3.2 What image should be described? 33

3.2.1 Definition 33
3.2.2 Aim and relevance 33
3.2.3 Findings 33
3.2.3.1 Decorative and informative 33
3.2.3.2 Surrounidng text is the key 33
3.2.3.3 Text-image relationships 34
3.2.3.4 Different approaches 34
3.2.3.5 Reuse of ID 36
3.2.3.6 Experts Inputs: users want control 36
3.2.4 Summary 36

3.3 What content of an image should be described? 37

3.3.1 Definition 37
3.3.2 Aim and relevance 38
3.3.3 Findings 38
3.3.3.1 Variability and variables 38
3.3.3.2 Amounts of details 40
3.3.3.3 Semantic factors 41
3.3.4 Summary 42

3.4 How to present the image description 43

3.4.1 Definition 43
3.4.2 Aim and relevance 43
3.4.3 How to organize textual ID 43
3.4.4 Spatial information 43
3.4.5 Rich presentations 44
3.4.6 Summary 45

3.5 Synthesis 46

3.5.1 Relevance to crowdworkers 46
3.5.2 A new research approach 47

4 Context research 50

4.1 What experts say 52

4.1.1 Aim of research 52
4.1.2 Research question 52
4.1.3 Research method 52
4.1.4 Key findings 52
4.1.4.1 Dutch accessibility publications
4.1.4.2 Dedicon's current products 54
4.1.4.3 Insight about users 55

4.2 What VIPs say 56

4.2.1 Aim of research 56
4.2.2 Research Questions 56
4.2.3 Research Method 56
4.2.4 Key findings 56
4.2.4.1 Reading habits and scenarios 56
4.2.4.2 Limitation of Choice 57

4.2.5 Difficulties in recruiting participants 57

5 Design Goal 60

5.1 Design Scope 62
5.2 Problem Statement 64
5.3 Design Opportunities 65
5.4 Design Goal 66

6 Research through Design 68

6.1 Introduction 69

6.1.1 Target 69
6.1.2 Research Question 69

6.2 Conceptualization 70

6.3 Test Setting 72

6.3.1 Design of test procedure 73
6.3.2 Articles and audios 73
6.3.3 Test Steps 74
6.3.4 Participants 75

6.4 Analysis 76
6.5 Results 77

6.5.1 Quantitive results 77
6.5.1.1 The effect of requirement collection 77
6.5.1.2 The impact of progressive image description (ID) 80

6.5.2 Qualitative results 81
6.5.2.1 The process of image perception 81
6.5.2.2 The influence of the structured ID 84

6.6 Transformation into design decisions 86

7 Demonstrator 89

7.1 Overview of the Workflow 91

7.3 Interfaces for VIP 92

7.3.1 Google Talkback System 92
7.3.2 Design Interfaces for VIPs 94

7.2 Interfaces for crowd workers 97

8 Evaluation 103

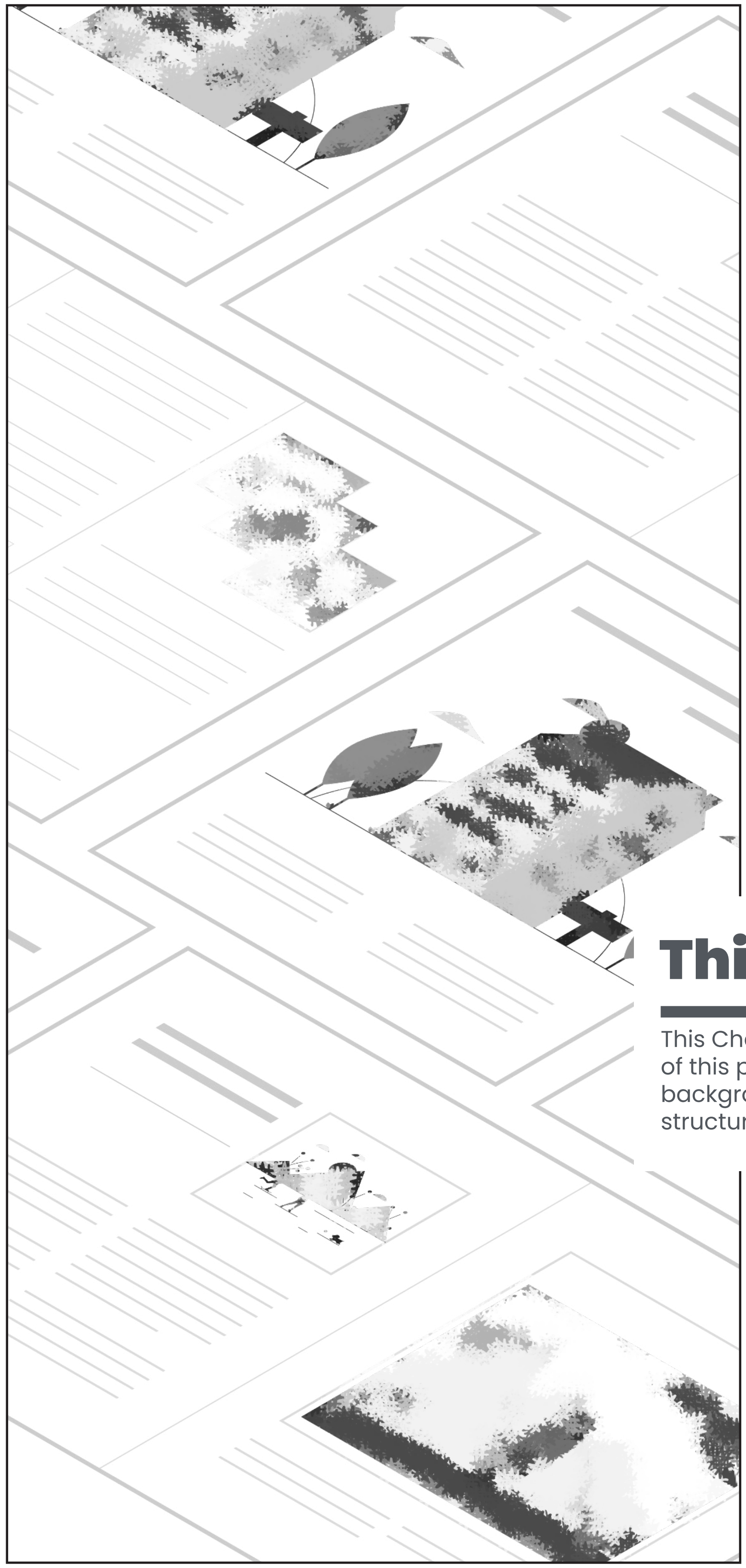
9 Conclusion 104

10 Reflections 108

10.1 Limitations 109
10.2 Recommendations 109
10.3 Personal reflections 110

11 Acknowledge 112

12 Reference 11



01

This Project

This Chapter provides an overview of this project, including its research background, target and research structure.

1.1Introduction

Designing image description system for Visually Impaired Users

This project is conducted in the context of the newly created Future Libraries Lab, which is a research and innovation collaboration between the Delft University of Technology and the Koninklijke Bibliotheek. The vision of this project is to make the information of images as accessible to visually impaired people as is to those with normal sights.

Figure-1.2 New EU rules will make key products and services accessible across the EU. A carer is instructing a visually impaired elderly person to use accessible services. Source: Social Europe, n.d., https://i.ytimg.com/vi/t5iW0TNQFP0/maxresdefault.jpg

Images are becoming a more prominent part of today's media. But have you ever wondered how visually impaired people use Instagram, Facebook ,or read books like people with normal sights?

They need image description(ID) to transform visual content into a way that they can assume. But this is not easy. Look at the image above. Can you quickly figure out how to describe it? Can you guarantee that your description is comprehensive enough? How to generate ID efficiently and effectively has gained increasing attention in various fields, including public policy, physiology, computer vision, and HCI. [5, 16]

Providing image descriptions will be a legal obligation in the near future. According to the European Accessibility Act, all new digital publications and services should be made accessible from 2025.[25] Thus, in the near future, should provide image descriptions for meaningful images, which means a huge demand for the production of the image description. But currently, we still lack a feasible solution to equip digital publications with image descriptions.

Fortunately, recent experiments with the human in the loop approaches (HITL) show possibilities to tackle this problem.[9] Traditionally, images are not widely accessible to visually impaired people (VIPs) due to issues of cost, scalability, timeliness, and quality. The development of crowdsourcing systems and AI captioning systems demonstrate their potential to produce image descriptions efficiently on a larger scale. Meanwhile, literature shows that the current AI captioning system is not reliable enough, and thus human participation is still necessary. To reduce the cost of time, money, and accuracy, more supports are crucial for crowd workers' production of ID and the scalability of this approach.

Therefore, focusing on the requirements of visually impaired people, the goal of this thesis is to elicit new knowledge and design an image description system, which can both facilitate the production of ID and satisfy the needs of visually impaired people.

[5] Morris, M. R. et al. (2016) '"With most of it being pictures now, I rarely use it": Understanding Twitter's evolving accessibility to blind users', Conference on Human Factors in Computing Systems - Proceedings, pp. 5506–5516. doi: 10.1145/2858036.2858116.

[16] Bhowmick, A. and Hazarika, S. M. (2017) 'An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends', Journal on Multimodal User Interfaces. Springer International Publishing, 11(2), pp. 149–172. doi: 10.1007/s12193-016-0235-6.

[25]Act, E. A. (2019) 'European Directive to Improve the Accessibility of Mainstream Ebooks', pp. 1–3, https://daisy.org/news-events/articles/european-directive-to-improve-the-accessibility-of-mainstream-ebooks/

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) '"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404.

[9]Salisbury, E., Kamar, E. and Morris, M. R. (2017) 'Conversational Crowdsourcing as a Tool', Aaai Hcomp 17, (Hcomp), pp. 147–156.

1.2 Research Goal

The target of this project is to investigate VIPs' needs and expectations of visual content and to develop new knowledge about their requirements on an image description system. The deliverable of this project is a demonstrator to communicate the learned visually impaired people's requirements and their preferable interaction mode for a technological image description system.

1.3 Approach & Process

The research activities are divided into 2 parts and 4 phases: Research (background research, requirements research, context research) and Design.

The original research goal of this project is to develop new knowledge about the requirements for image description from the perspective of people with visual impairments. In the background research phase, I conducted research on the project background and key concepts, including visually impaired people, image description, current problem, and solutions. Background research reveals: HITL approach is still the best way to generate an ID. It also helps me better understand the research goal: requirements collected should benefits crowd workers as well.

Based on this research question, researches on the requirements of VIPs was conducted, through a review of existing guidelines, academic researches, and field research. Three research questions are investigated:

- What images should be described?
- What content of an image should be described?
- How image should be presented?

The research results indicate it is found it's difficult to find comprehensive and clear answers which can directly benefit crowd worker's work and replace the role of subjective judgments and empirical knowledge, because of the complexity of the relationship image-text relationship and the various factors including the preference. On the other hand, literature review results also reveal the shortcomings of the existing one-shot description. In comparison, interactive image description has the potential to satisfy visually impaired users' need of having control, as well as their personalized preference of ID content. It inspired me that interactive ID can be utilized as an approach to collect requirements and guide crowd workers' work as well. According to these insights and the restrictions from the Corona situation, I adjusted the research direction and proposed my design goal:

> *"To develop a system which enables VIPs to have control on their ID and is able to collect VIPs' requirements that can be transformed into straightforward description tasks for crowd workers"*

The design phase of this project is aimed to verify the feasibility of this design goal and investigate the impact of an interactive image description system.

In the design phase, a prototype that simulates a progressive ID system with structured information was developed. Through a set of comparative experiments, the systems' function to collect user preferences and gradually improve the content of image description is confirmed. The qualitative research results also reveal the mental model when users acquiring information through such a system. These findings are transformed into design decisions as well.

Based on the evaluation results; a final design was proposed. Due to the limited access to visually impaired people under the COVID-19 situation and shortage of remaining project time, evaluation is planned after this graduation project. Finally, the overall conclusion was derived from the results of the design phases, as well as recommendations and reflections on this project.

Key context elements and stakeholders

■ Visually impaired people - VIPs
Visually impairment - VI

■ Image Describers
Transform image into text

■ Koninklijke Bibliotheek - KB
Dutch National Library

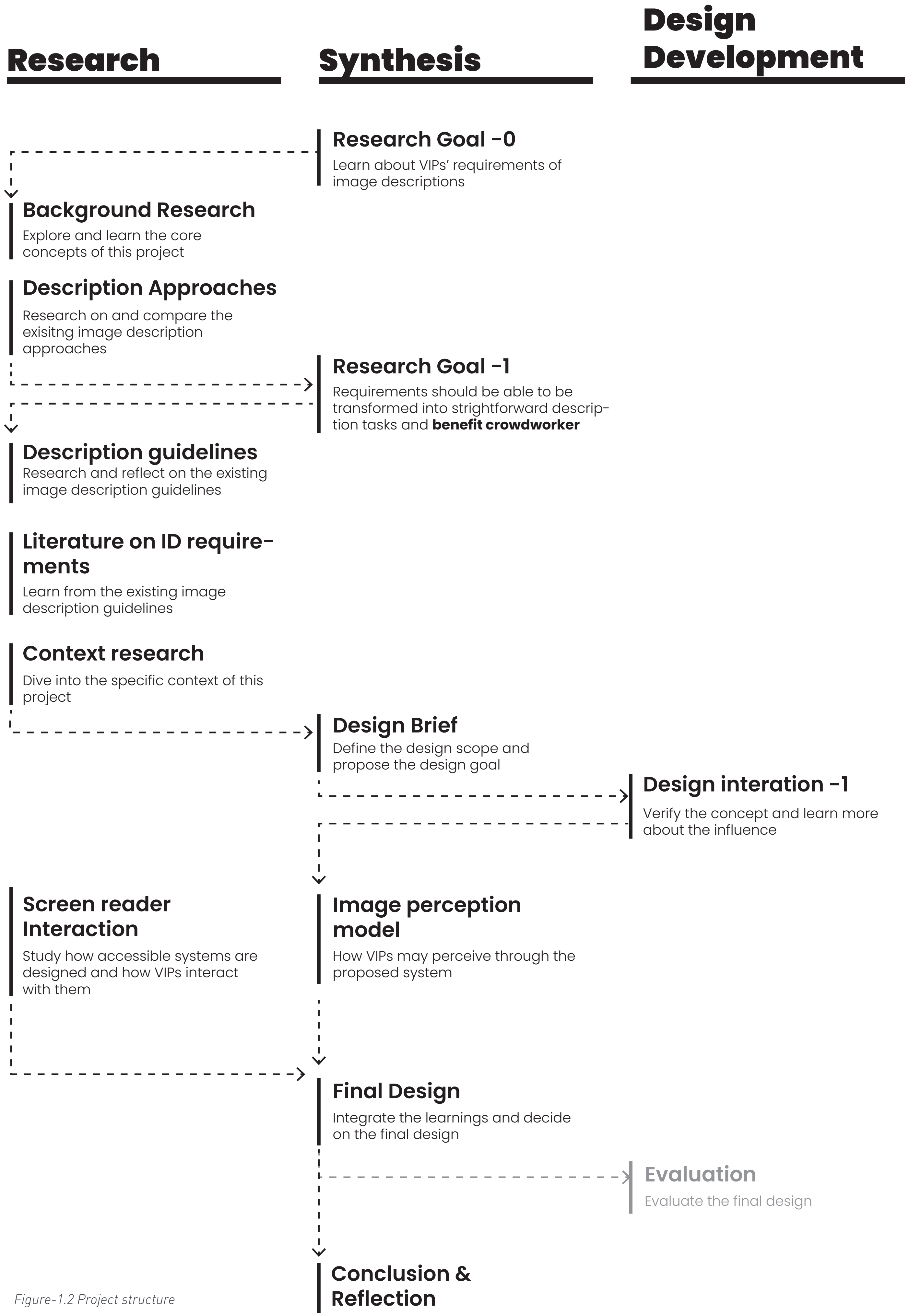
■ Dedicon
Work for accessible reading

■ Image Description - ID

■ Interactive ID - IID
ID provide interactive feature

■ Progressive ID
Interactive ID that allows you to acquire information in a progressive way

■ Structured ID
Interactive ID that arrange content in a structured way



Research

Synthesis

Design Development

Research Goal -0
Learn about VIPs' requirements of image descriptions

Background Research
Explore and learn the core concepts of this project

Description Approaches
Research on and compare the exisitng image description approaches

Research Goal -1
Requirements should be able to be transformed into strightforward description tasks and **benefit crowdworker**

Description guidelines
Research and reflect on the existing image description guidelines

Literature on ID requirements
Learn from the existing image description guidelines

Context research
Dive into the specific context of this project

Design Brief
Define the design scope and propose the design goal

Design interation -1
Verify the concept and learn more about the influence

Screen reader Interaction
Study how accessible systems are designed and how VIPs interact with them

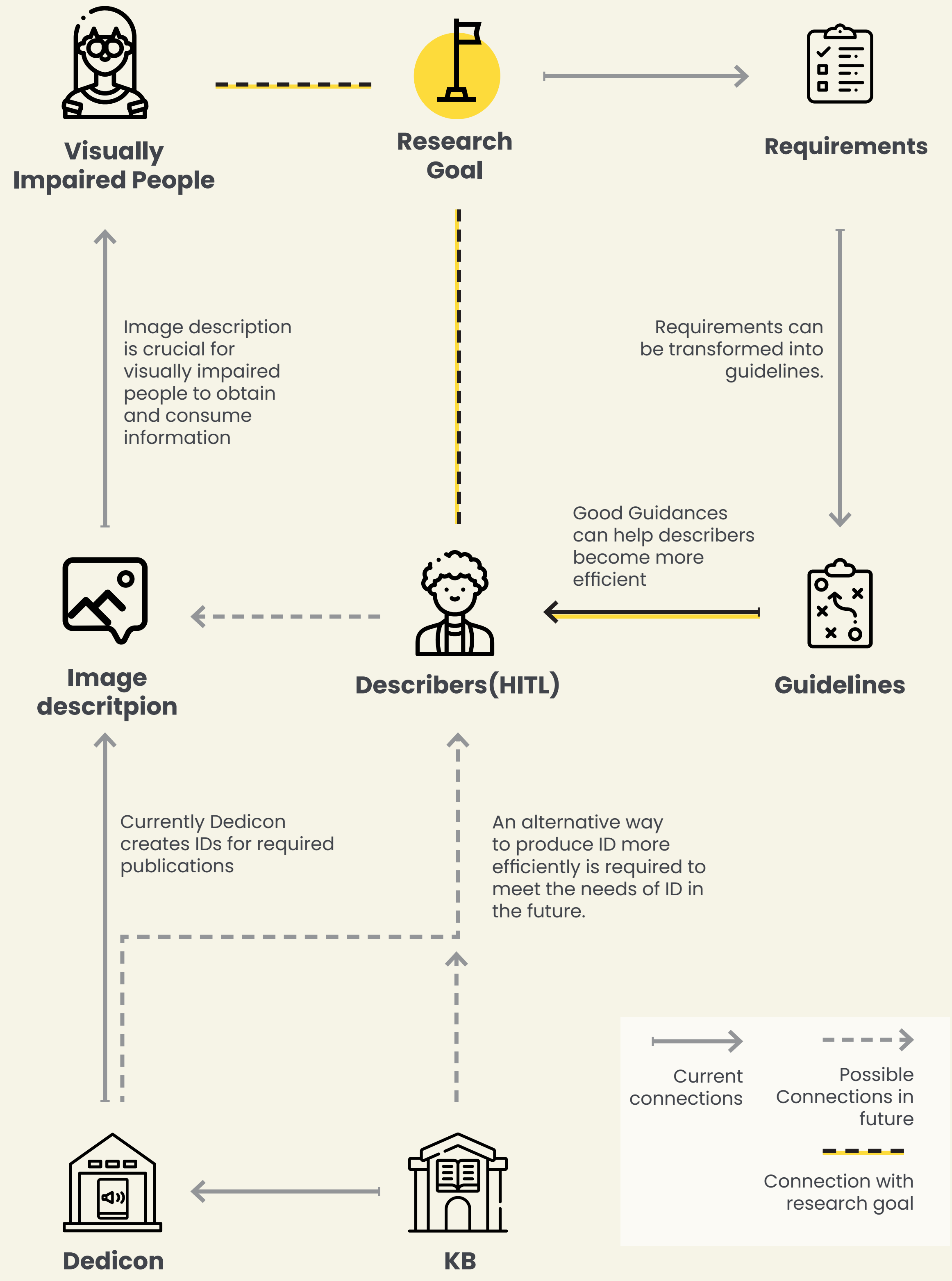
Image perception model
How VIPs may perceive through the proposed system

Final Design
Integrate the learnings and decide on the final design

Evaluation
Evaluate the final design

Conclusion & Reflection

Figure-1.2 Project structure



Visually Impaired People
Research Goal
Requirements
Image description is crucial for visually impaired people to obtain and consume information
Requirements can be transformed into guidelines.
Good Guidances can help describers become more efficient
Image descritpion
Describers(HITL)
Guidelines
Currently Dedicon creates IDs for required publications
An alternative way to produce ID more efficiently is required to meet the needs of ID in the future.
Current connections
Possible Connections in future
Connection with research goal
Dedicon
KB

Background

The research goal of this project is: to develop new knowledge of the requirements on image description from the perspective of visually impaired people. This chapter is aimed to set the background and refine this research goal.

Firstly, I will introduce the key concept of this project: visual impairment (VI) and image description (ID). What they are and why we need to create image descriptions for visually impaired people will be explained.

Next, two important stakeholders, Dedicon and KB will be introduced: I will explain their role in the production of Image description and as a potential client, their needs for this project, i.e. to find an alternative way to produce ID more efficiently.

Finally, based on the needs, a review of the image description methods will be presented, namely: First-party description, AI captioning and humanin--the-loop (HITL) approach, which explains why we cannot simply rely on automatically generated descriptions to solve the shortage of ID. And from this, a refined research goal will be proposed, i.e., the requirements collected should be able to benefit crowd workers as well.

VIPs

2.1 Introduction of visually impaired people (VIPs)

- What impact may visual impairment have on the lives of visually impaired people?

- What is the distribution of the VIPs?

2.1.1 Definition

Visual impairment (VI) is a condition of reduced visual performance that cannot be remedied by refractive correction (spectacles or contact lenses), surgery, or medical methods [42]. VI will cause functional limitations of the visual system, including irreversible vision loss, restricted visual field and decreased contrast sensitivity, increased sensitivity to glare [32].

Typically, VI is measured by exclusively visual acuity (VA, visual acuity of the better eye with the best possible refractive correction), with severity categorized as mild, moderate, or severe distance vision impairment or blindness, and near vision impairment [42]. In the clinical setting, other visual functions are also often assessed, such as a person's field of vision, contrast sensitivity, and color vision

[42] World report on vision (WHO), 2019, https://www.who.int/publications/i/item/9789241516570

[32] Vision, A. and Health, E. (2018) 'A review of visual impairment', pp. 1–4.

[42] World report on vision (WHO), 2019, https://www.who.int/publications/i/item/9789241516570

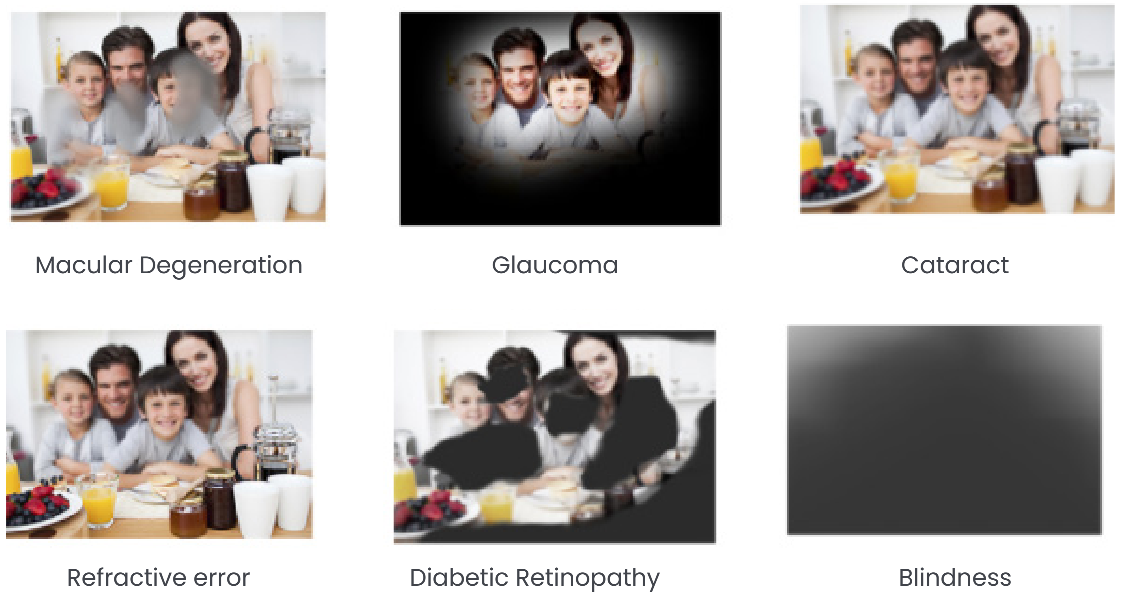
Category	**Visual acuity in the better eye**	
	Worse than:	**Equal to or better than:**
Mild vision impairment	6/12	6/18
Moderate vision impairment	6/18	6/60
Severe vision impairment	6/60	3/60
Blindness	3/60	
Near vision impairment	N6 or M 0.8 at 40cm	

Figure-2.1 Definition of different levels of visual impairments. The Visual impairment levels are decided by the visual acuity.

** Visual acuity is calculated by using two numbers. The first number indicates the distance between the chart and the person reading the chart. The second number is the distance that someone with normal vision is able to read at 20ft. distance from the chart. People with normal vision can read the 20 ft line at 20 ft., a 20/20 visual acuity.*

2.1.2 Impact of VI

VI decreases the ability of an individual to function independently and negatively impacts daily living and quality of life.[13-4] Different causes of visual impairments result in different symptoms, as is shown in Figure 2.2.


Macular Degeneration

Glaucoma

Cataract

Refractive error

Diabetic Retinopathy

Blindness

Figure-2.2 Symptoms of different causes of VI. These symptoms include blurred vision, spots in the vision, impaired central vision, impaired peripheral vision, etc. Source: Optelec International, 2018

The effect of VI contributes to deficits in performance on everyday tasks, including reading and writing. [44] The combination of social, functional, and psychological disabilities related to VI may result in an overall reduction in quality of life[45, measured by EQ-5D, On average VI has more negative impacts then diabetes typeII, coronary syndrome and hearing impairments]. Individuals with VI experience more symptoms of depression than those without VI. [46, 45] In addition, the range of problems caused by VI amongst VIPs is staggering – there is truly no "one size fits all" [28].

Barriers faced by blind and partially sighted people in life [28]:

- *Public attitudes*
- *Employment*
- *Navigating streets*
- *Claiming benefits*
- *Education and support for children and young people*
- *Social and leisure*
- *Transport*
- *Technology (inaccessible support, equipment and content)*
- ***Accessing information, products and services***
- *Coming to terms with sight loss and maintaining confidence*
- *Taking care of oneself and the home*
- *Diagnosis, treatment and ongoing care*

There are a number of design challenges relating to VIPs and one of them is access to information, products and services. Part of this is what this project is aimed to understand and improve.

[44] West, S. K. et al. (2002) 'How Does Visual Impairment Affect Performance on Tasks of Everyday Life?', 120(June).

[45] Langelaan, M. et al. (2009) 'Impact of Visual Impairment on Quality of Life : A Comparison With Quality of Life in the General Population and With Other Chronic Conditions Impact of Visual Impairment on Quality of Life : A Comparison With Quality of Life in the General', 6586. doi: 10.1080/09286580601139212.

[46] Binns, A. M. et al. (2012) 'How Effective is Low Vision Service Provision ? A Systematic Review', Survey of Ophthalmology. Elsevier Inc, 57(1), pp. 34–65. doi: 10.1016/j.survophthal.2011.06.006.

[28] Fisher, D. (no date) 'Barriers faced by blind and partially sighted people - RNIB Strategic prioritisation research Authors'.

Moreover, since congenitally blind people have no memory of visual contents, they are relatively better adapted to their situation. In contrast, acquired blindness will greatly undermine an individual's independence. People who become blind recently will need time to get used to the situation and retrieve their confidence gradually.[43]

[43] SSMR. (2009) 'Understanding the Needs of Blind and Partially Sighted People : their experiences , perspectives , and expectations'. https://www.rnib.org.uk/sites/default/files/Understanding_Needs_Lit_Review.doc

2.1.3 Distribution of visual impairment

According to the World report on vision 2020, globally at least 2.2 billion people have a vision impairment and have some degree of low vision.[42] A report based on the Dutch population model estimates that there were approximately 80,000 blind people and 290,000 people with low vision in the Netherlands in 2020. [30]

[42] World report on vision (WHO), 2019, https://www.who.int/publications/i/item/9789241516570

[30] Limburg, H. and Keunen, J. E. E. (2020) 'Blindness and low vision in The Netherlands from 2000 to 2020 — modeling as a tool for focused intervention Blindness and low vision in The Netherlands from 2000 to 2020 — modeling as a tool for focused intervention', 6586. doi: 10.3109/09286580903312251.

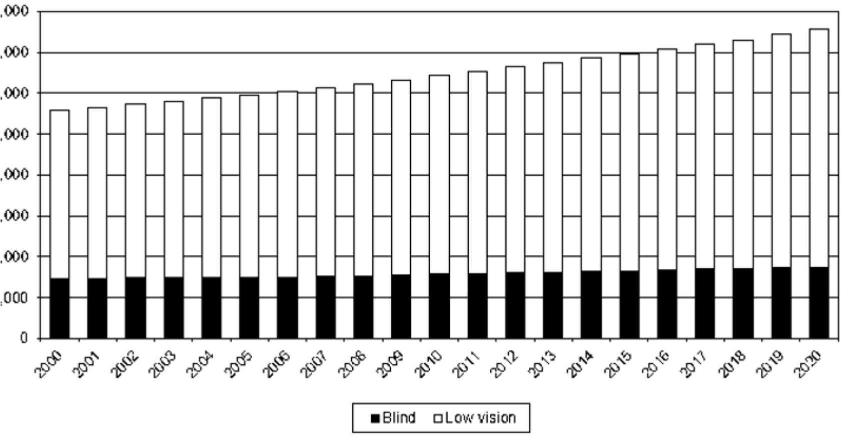


Figure-2.3 A bar graph showing a predictive model for showing the amount of visually impaired people in the Netherlands. In 2020, there were approximately 80,000 blind people and 290,000 people with low vision in the Netherlands. Source: Limburg, H. and Keunen, J. E. E. (2020) [30]

By age: visual impairment is unequally distributed across age groups. More than 80% of VIPs in the Netherlands are 50 years of age and older. (Figure) [30]

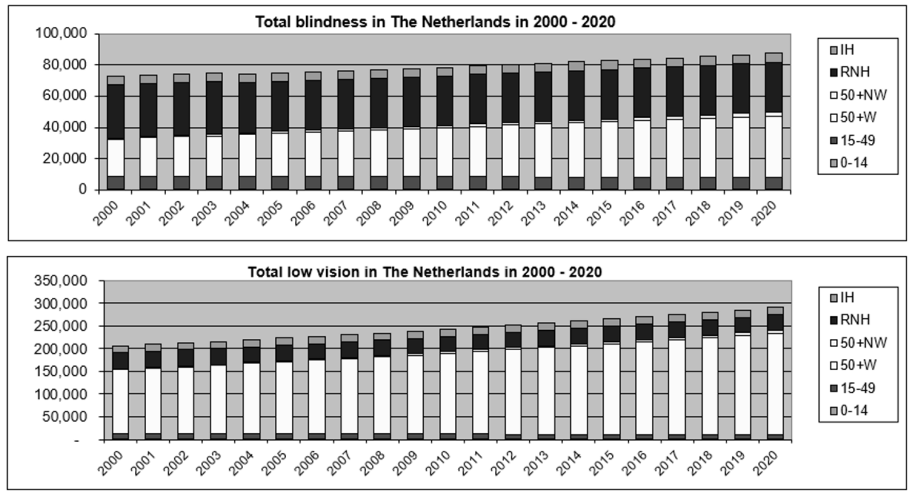


Figure-2.4 A bar graph showing a predictive model for visual impairment in the Netherlands. Information has been covered by the text. Source: Limburg, H. and Keunen, J. E. E. (2020) [30]

*[47]*Oogvereniging (2018). Slechtziendheid - Oogvereniging. [online] Oogvereniging. Available at: https://www.oogvereniging.nl/oogaandoeningen/oogaandoeningen-overzicht/blind-doofblind-of-slechtziend/slechtziendheid/

By classification(Blindness and low vision): Blindness means having less than 0.05 acuity, which accounts for about 24% of all VIPs. It is worth noting that legal blindness isn't equal to full blindness. People classified as blind might still be able to see contrast for instance [47].

By causes: There are various causes of vision impairments. Cataract is the most common reason for VI in the Netherlands, while AMD (Age-Related Macular Degeneration) is the leading cause for blindness.

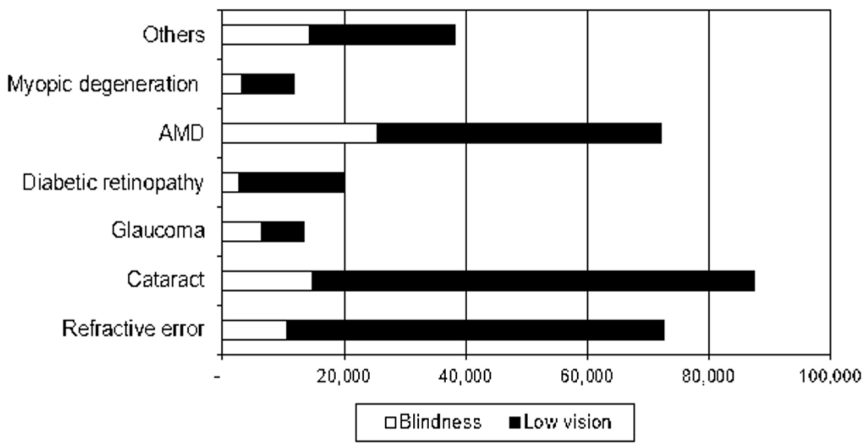


Figure-2.5 A bar graph showing a predictive model for visual impairment in the Netherlands. The number of reasons in descending order is: cataract, AMD, Refractive error, diabetic retinopathy, Glaucoma, Myopic degeneration and others. Source: Limburg, H. and Keunen, J. E. E. (2020) [30]

[47] Oogvereniging (2018). Slechtziendheid - Oogvereniging. [online] Oogvereniging. Available at: https://www.oogvereniging.nl/oogaandoeningen/oogaandoeningen-overzicht/blind-doofblind-of-slechtziend/slechtziendheid/

[48] Optelec International. (2018). Eye conditions. [online] Available at: https:// in.optelec.com/eyeconditions

Congenital and acquired blindness: The majority of the VIPs become blind due to an accident or disease, people with congenital blindness are a relatively smaller group.[47]. Diseases that cause visual impairment are progressive, people with a visual impairment often have a large chance of becoming blind later in life [48].

Summary

1. **The impact of visual impairment on the visually impaired is multi-faceted and diverse**. The impact of visual impairment is not limited to visual-related activities but also includes social and psychological impacts. The difficulty of obtaining information is also one of them.

2. The elderly in the visually impaired group constitute the majority. People with congenital visual impairment account for a small proportion. The proportion of blindness is also relatively small, especially for fully blind people.

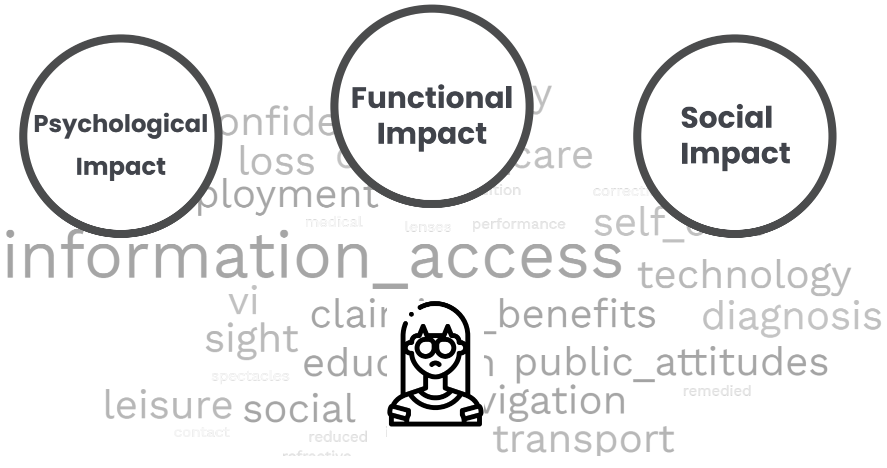
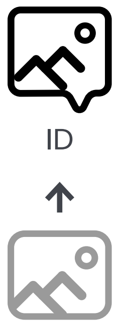


Figure-2.6 An illustration summarize the potential negative impacts for the visually impaired people, the content is explained in the main text.



2.2 Image description for publications

- What is the definition and function of image description?

- What is the current accessibility of images in publications? How about the trend in the future?

Visual content plays an important role in both analog and digital media. For a sighted person, colors, pictures and animations can help them better understand and navigate the information, and enhance the experience of digital services. However, for VIPs, they have difficulties consuming visual contents (including images, formats, layouts, etc.) and may thus miss key aspects of information. [3]

As proposed by CRPD(Convention on the Rights of Persons with Disabilities),

"persons with disabilities access, on an equal basis with others, … to information and communications, including information and communications technologies and systems, and to other facilities and services open or provided to the public, both in urban and in rural areas"

VIPs should enjoy the same rights as people with normal sight to obtain and consume information, including visual content. Image description (ID), as a textual alternative for the images, is produced for such a purpose.

In this section, we will introduce in detail the definition and scope of ID and provide a brief review of the current accessibility of visual content with the context of this project.

2.2.1 Definition of image description

ID is defined as "a textual description of images presented in the digital document" by WebAIM. It serves as an alternative to image and intends to provide an equivalent meaning. Although its definition seems self-explanatory, there may be nuances among what ID refers to under different contexts and different research fields. This section is intended to clarify this.

2.2.1.1 Levels of image descriptions

Image description is not a deterministic process, and there may be different ways to describe the same image. As is suggested by Hodosh, there are 3 kinds of image description [20], which are:

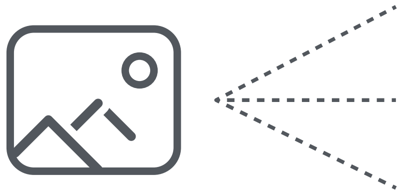
1. **Conceptual descriptions** that identify what is depicted in the image , and may be abstract (e.g., concerning the mood a picture may convey);

2. **Non-visual descriptions** provide additional background information that cannot be obtained from the image alone (e.g location in which the image was taken)

3. **Perceptual descriptions** capture low-level visual properties of images.(e.g. colors)

[3] Petrie, H. et al. (1999) 'Describing images on the Web : a survey of current practice and prospects for the future Centre for Human Computer Interaction Design City University London Northampton Square 2 The importance of describing images on the Web'.

[20] Hodosh, M., Young, P. and Hockenmaier, J. (2015) 'Framing image description as a ranking task: Data, models and evaluation metrics', IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua, pp. 4188–4192.



Conceptual description
This image contains mountains, sun.

Non-visual descriptions
This image is made by the author of this thesis

Perceptual descriptions
This image is an illustration, probably serves as an icon.

Figure-2.7 An illustration showing 3 layers of image description

Even though it is argued that conceptual description should be the focus of automatic image description system [10-0], recent studies have pointed out the short-comings of this proposal and concluded that mere conceptual image description cannot fully satisfy the needs of VIPs. [1, 9, 3] For clarification, ID in this report does not specifically refer to any of these 3. In fact, following research activities (which are introduced in Chapter 3) reveal that under different circumstances, the ID required may include any one or more of them.

2.2.1.2 Alt-text and image description

Image Description and alt-text have always been inextricably linked. On one hand, alt-text is the earliest and most common form for the dissemination of ID. On the other hand, with the development of web technology and accessibility, ID has developed new forms and contents in different media.

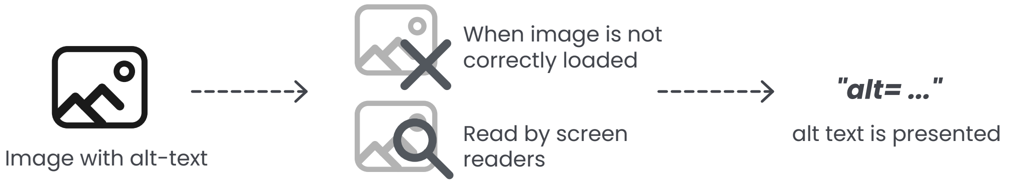
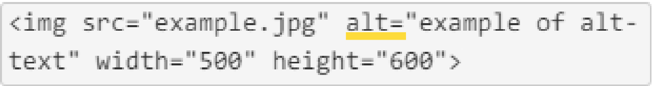


Figure-2.8 An illustration showing the function of Alt-text of websites

In the context of Web service, image description is widely referred to as alt-text. Alt-text was first introduced by the HTML 2.0 standard in 1995, wherein images allowed for an alt attribute.[49] This attribute contains text as an alternative to images in case images can not be loaded. For visually impaired users with screen readers, the content within the alt attribute will be translated into braille or audio as an alternative for visual elements.



```
<img src="example.jpg" alt="example of alt-
text" width="500" height="600">
```

Figure-2.9 A piece of code that shows how alt-text is used in HTML language

In most cases, descriptions of an image will be put in the "*alt*" attribute. However, even in web content, image descriptions are not necessarily equivalent to alt-text. For example, to meet the need for a progressive description, HTML allows the attribute *longdesc* to store more detailed descriptions of images.[50]

More importantly, for digital publications, the format of alt text is different from that on the web, even though some of them share a similar format to HTML, e.g. EPUB3. In particular, there are formats specifically produced for accessibility, such as DAISY, which require unique hardware or software to consume. [51, 52]

[20] Hodosh, M., Young, P. and Hockenmaier, J. (2015) 'Framing image description as a ranking task: Data, models and evaluation metrics', IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua, pp. 4188–4192.

[1] Miltenburg, E. Van (no date) Pragmatic factors in [automatic] image description.

[9] Salisbury, E., Kamar, E. and Morris, M. R. (2017) 'Conversational Crowdsourcing as a Tool', Aaai Hcomp 17, (Hcomp), pp. 147–156.

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) '"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404.

[49] Hypertext Markup Language - 2.0, https://tools.ietf.org/html/rfc1866]

[50] HTML longdesc attribute https://www.w3resource.com/html/attributes/html-longdesc-attribute.php]

[51] EPUB 3 http://diagramcenter.org/59-image-guidelines-for-epub-3.html]

[52] DAISY https://daisy.org/activities/standards/daisy/]

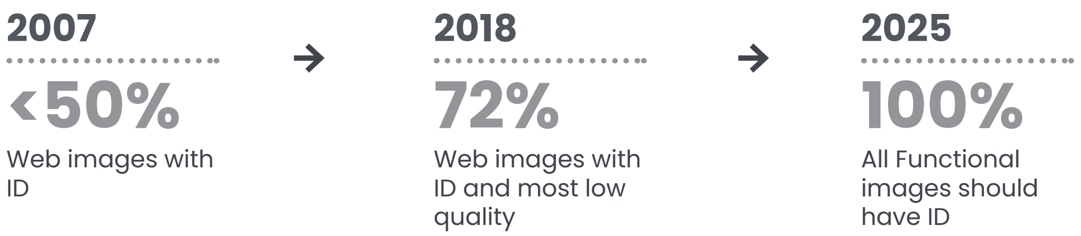
2.2.1.3 Captions

In the newspaper and other publications, there will be textual descriptions next to the picture to indicate the relation of the image with the article content and provide some additional information. The content of the caption may be similar to the non-visual description we mentioned before. The form of a caption is different from the ID discussed in this thesis as well. IDs for accessibility are usually invisible and are used to describe pictures to screen readers (or search engines), but captions are visible and for all users. [53]

2.2.2 Demand of ID for now and in the future

ID makes visual content available to all individuals. Studies have found that VIPs usually have only low-level engagement with images, although they do have intentions because of social, entertainment, and educational needs.[5, 54] And this low engagement mainly stemmed from inadequate descriptions of images. [2]

On one hand, in general, the pervasiveness of ID is constantly improving, but continuous efforts are still needed. In a 2007 study, Bigham pointed out that only about half of the pictures have alt-text [55]. In contrast, according to a more recent study in 2018, 72% of pictures from popular websites already have alt-text, which means web accessibility practices have become more widely adopted, at least on popular websites.[7] However, it was also pointed out that many of the existing image alt-text is not helpful enough and need to be replaced. And the knowledge and support for accessibility improvement are still lacking. In addition, the situation of websites is different from that of publications, since publications have more diverse formats, sources and more complicated issues of copyright and distribution (which will be discussed in the expert interview part).[19]



2007 → 2018 → 2025

<50%
Web images with ID

72%
Web images with ID and most low quality

100%
All Functional images should have ID

Figure-2.10 In 2025, all functional images should have image descriptions

On the other hand, on March 13 2019, the European Parliament approved the European Accessibility Act, which requires all e-book and digital services to be born accessible from 2025. This means that all the publications should include, for VIPs, all the features and functionality that those without VI can enjoy.[26] Thus, there will be considerable demand for ID in the near future.

Summary

ID is a "textual description of images", which is usually placed in the alt attribute and for screen readers to read. There are increasing regulations and practices dedicated to improving the accessibility of information, including the pervasiveness of ID. In the next chapter, 2 Dutch organizations related to this work will be introduced. They are also cooperators of this project.

[53] Beginner's Guide to Image SEO – Optimize Images for Search Engineshttps://www.wpbeginner.com/beginners-guide/image-seo-optimize-images-for-search-engines/#:~:text=What%20is%20the%20Difference%20Between%20Alt%20Text%20vs%20Caption,are%20visible%20below%20your%20images.]

[5] Morris, M. R. et al. (2016) '"With most of it being pictures now, I rarely use it": Understanding Twitter's evolving accessibility to blind users', Conference on Human Factors in Computing Systems - Proceedings, pp. 5506–5516. doi: 10.1145/2858036.2858116.

[54] Zhao, Y. et al. (2017) '[05-16] The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments', Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), pp. 1–22. doi: 10.1145/3134756.

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) '"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404.

[55] Bigham, J. P. et al. (no date) '[01-4]WebinSitu : A Comparative Analysis of Blind and Sighted Browsing Behavior', pp. 51–58

[7] Guinness, D., Cutrell, E. and Morris, M. R. (2018) 'Caption Crawler: Enabling reusable alternative text descriptions using reverse image search', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3174092.

[19] Power, C. et al. (2012) 'Guidelines are Only Half of the Story : Accessibility Problems Encountered by Blind Users on the Web', pp. 433–442.

[26] Kasdorf, B. B., The, C. and Union, E. (2019) 'Make E-books Accessible Now', pp. 1–4.

2.3 Koninklijke Bibliotheek and Dedicon

Visually impaired people (VIPs) rely on ID to consume visual content. Who is responsible for this work?
In this chapter, KB, Dedicon and their role in the image description production will be introduced. This report is conducted in cooperation with Koninklijke Bibliotheek (KB) and Dedicon.

2.3.1 Dedicon

Dedicon is a Dutch organization which aims at creating solutions for people with visual or reading disabilities. Dedicon receives subsidies from the Ministry of Education, Culture and Science and the Koninklijke Bibliotheek (KB). Thus, the objects of its products and services are mainly oriented to the education field and library system.

Figure-2.11 A worker from Dedicon is creating image descriptions in front of a recording devices.

2.3.1.1 Dedicon and accessible publications

A substantial part of Dedicon's activities consists of making existing (school) books, newspapers and magazines accessible. Publishers are main clients of these services.

If VIPs want a narrated or braille versions of a publication, he or she uses the services of Passend lezen and Dedicon. According to copyright law, organizations like Dedicon have the right to 'enhance', make content accessible and distribute the enhanced formats to users that meet certain 'impairment criteria'.

2.3.1.2 Dedicon and image description

Image description (ID) is an important part of making publications fully accessible. However, currently most products from Dedicon do not include ID for now. They might include a little reference to indicate the existence of the image (and maybe also the subtitle), but a full description of what can be seen on the images is currently not yet part of the regular production system in most cases, including the books Dedicon made for KB and leisure reading materials like magazines. Only for educated materials, all of the functional images will be described, which enables students to do exercises or learn the necessary information.

The problem facing Dedicon at this stage is that they "do not have a solution to make image description everywhere". Therefore, they need to find an alternative way to generate ID more efficiently.

2.3.2 Koninklijke Bibliotheek (KB)

Figure-2.12 A bird-view photo of the KONINKLIJKE BIBLIOTHEEK buidling, which has a huge volume

KONINKLIJKE BIBLIOTHEEK is the national library, located in The Hague. It collects everything that appears in and about the Netherlands, from medieval literature to contemporary publications, which is done together with partners in the field of heritage, science and with public libraries. “*As a national library, the job of KB is to make the library collection of the Netherlands visible, usable and sustainable.*” [56]

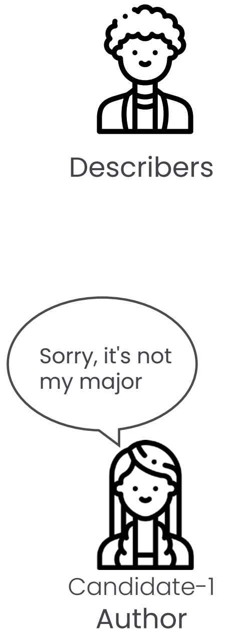
Relevance of KB with this project

Making archives accessible for VIPs is also within KB’s vision. However, existing public library resources may not provide enough accessibility support. For example, the services around the KB digital archives like Delpher are currently not accessible enough because of poor OCR quality. Although this project cannot directly offer KB a solution, describing images efficiently is undoubtedly a necessary step to realize KB’s vision.

Summary

In the Netherlands, Dedicon is responsible for making reading materials accessible and KB is one of Dedicon’s main clients. However, for Dedicon, ID is still a new technique and most of their products do not include descriptive image descriptions. An alternative way to generate ID more efficiently is required.

[56] KB, no date, Introduction of KB https://www.kb.nl/organisatie



2.4 Automatic or manual? Ways to generate IDs

Based on the requirements for a more efficient ID generation approach, this section will provide a brief introduction to ID methods, answering why we cannot simply rely on automatic generation, and explaining which method is the most feasible one at the moment.

2.4.1 The lack of first-party ID

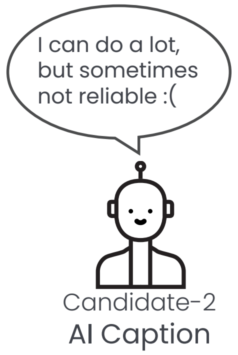
For ID production, the most ideal situation is that authors themselves can provide high-quality ID. But at the current stage, it is not enough to rely solely on the author's description, especially for publications.

Firstly, the popularity of existing IDs is still insufficient. Although there is no existing quantitative research on the popularity of image descriptions in publications, for web services and social media, missing or low-quality alt text is a pervasive problem [17]. In addition, in a qualitative study on VIP, it was also found that the participants who had experience using digital textbooks noted that the images presented within this source were not accessible to them. [2]

Secondly, except for publications that will be published in the future, images in archived historical publications also need to be described. And this demand can only be solved through the efforts of a third-party system. The potential demand in this area is also huge.

Thirdly, even if all authors have sufficient awareness and willingness to describe pictures, they still need relevant support and resources to effectively and efficiently complete this work. [34, 2] In fact, the majority of publishers are still relying on outsourcing (e.g. Dedicon) to improve the accessibility of publications [expert interview].

In conclusion, a system (either machine or manual) that can help produce IDs will still be an important supplement to first-party publishers in the foreseeable future.



2.4.2 Applications and limitations of AI image captioning

Due to the development of computer vision in recent years, it has become possible to use artificial intelligence to generate picture descriptions. Image understanding and automated image captioning [16] have become hot topics in the field of artificial intelligence in recent years, and have led to more and more related practices, including Microsoft's CaptionBot [57], Google's Show and Tell [58], and many more.

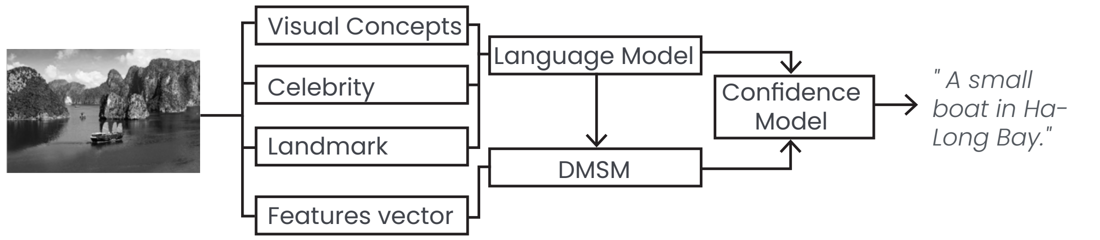


Figure-2.12 Illustration of an image pipeline example Source:[24]

[17] Morris, M. R. et al. (2018) ‘Rich representations of visual content for Screen reader users’, Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3173633.

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) ‘“Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions’, pp. 1–13. doi: 10.1145/3313831.3376404.

[34] Bigham, J. P. (2007) ‘Increasing Web Accessibility by Automatically Judging Alternative Text Quality’, pp. 349–352.

[16] Bhowmick, A. and Hazarika, S. M. (2017) ‘[06-01]An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends’, Journal on Multimodal User Interfaces. Springer International Publishing, 11(2), pp. 149–172. doi: 10.1007/s12193-016-0235-6.

[57] CaptionBot – For pictures worth the thousand words, 2017. https://www.captionbot.ai.

[58] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

After several years of development, current state of art AI captioning can not only perform object recognition, but also identify celebrities and landmarks based on the database, and utilize NLP technology to compose complete sentences.

The employment of AI is undoubtedly the most time and cost-efficient way to generate IDs. AI captioning system has also been implemented in social platforms (e.g. facebook) and accessibility system (e.g. ios talkback). In general, it is considered to be helpful in enhancing the experience of VIPs. [23]

However, recent studies have found that current AI image captioning systems still have a number of limitations, including the semantic gap, the deficiency of content, reliability problems, and possible ethical issues, which indicate that automatic image captioning systems still require more work before they are ready. [9]

[23] Wu, S. et al. (2017) 'Automatic alt-text: Computer-generated image descriptions for blind users on a social network service', Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, pp. 1180–1192. doi: 10.1145/2998181.2998364.

[9] Salisbury, E., Kamar, E. and Morris, M. R. (2017) 'Conversational Crowdsourcing as a Tool', Aaai Hcomp 17, (Hcomp), pp. 147–156

2.4.2.1 Semantic and pragmatic gap

First, for AI captioning systems, the lack of strict correlation between semantic concepts and visual features, referred to as the semantic gap, is a huge challenge. [24] When humans describe an image , appropriate inferences and associations will be made in conjunction with its context, so as to get expressions related to the context. (As the hints about the CEO, resignation given in the example). These vocabularies are essential for understanding the content of the picture, but correctly generating them is a challenging task for the machine so far.

[24] Tariq, A. and Foroosh, H. (2017) 'A Context-Driven Extractive Framework for Generating Realistic Image Descriptions'. IEEE, 26(2), pp. 619–632.

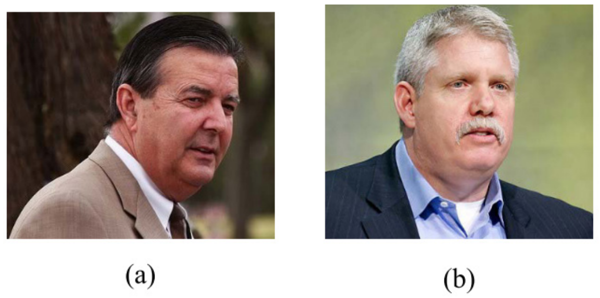


Figure 2.13 Sample image from MSCOCO; Caption: [Elderly man in brown suit and tie near tree in outdoor setting.]. The caption is an artificial image description and it provides no context for the image. (b) Sample news image; Caption: [BestBuy CEO Brian Dunn resigned amid investigation into ...]. Article: [BestBuy, low profits, CEO resignation]. The caption includes contextual hints. Source: [24]

2.4.2.2 Inability to offer details

Secondly, in certain contexts the AI captioning systems lack the ability to generate all necessary content required by the user. Normally the machine-generated description is a relatively short sentence, which is the result of the trade-off between descriptive quality and algorithmic accuracy. [23] However, the user's demand for more information is widely reflected.[3] The required content may include what the machine is not up to, such as the description of subjective issues and additional information beyond the content of the picture, for which world knowledge and reasoning is a pervasive need. [1]

[23] Wu, S. et al. (2017) 'Automatic alt-text: Computer-generated image descriptions for blind users on a social network service', Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, pp. 1180–1192. doi: 10.1145/2998181.2998364.

[3] Petrie, H. et al. (1999) 'Describing images on the Web : a survey of current practice and prospects for the future Centre for Human Computer Interaction Design City University London Northampton Square 2 The importance of describing images on the Web'.

[1] Miltenburg, E. Van (no date) Pragmatic factors in [automatic] image description.

2.4.2.3 Reliability and ethical concerns

Thirdly, reliability and implicit bias are additional concerns for AI captioning systems. A recent study found that VIPs tend to be over-trusting about AI captioning systems. [14] Therefore, when the ID is wrong, the AI caption will hinder VIP's ability to accurately understand the image. In addition, AI caption will inevitably expose systemic biases about gender and race that exist in the database.[18]

To sum up, the existing AI captioning systems are still not able to meet the needs of users for picture description, both in terms of accuracy and sufficiency. Considering the difference between usage scenarios of the pictures in the publication compared and other platforms (less quantity, higher quality), it is not the ideal choice to only focus on efficiency and fully rely on the AI image captioning system. Although there has been increasing research devoted to discussing and solving the proposed problems of AI image captioning, there has been no effective solutions so far. [18] Therefore, extra help in the production of ID for publications is still necessary, namely the help from the human.

[14] Salisbury, E., Kamar, E. and Morris, M. R. (2018) 'Evaluating and complementing vision-to-language technology for people who are blind with conversational crowdsourcing', IJCAI International Joint Conference on Artificial Intelligence, 2018-July, pp. 5349–5353. doi: 10.24963/ijcai.2018/751.

[18] Morris, M. R. (2020) 'AI and accessibility Discussion of ethical concern', Communications of the ACM, 63(6), pp. 35–37. doi: 10.1145/3356727.

2.4.3 Human in the loop (HITL) and crowdsourcing

Ai caption systems are not ready to generate a description that satisfies VIPs enough. And even if AI systems evolve, relying on human corrective techniques is still important. [9] Therefore, various directions have been explored to find ways to efficiently and accurately produce ID, chief among them is human-in-the-loop (HITL) approaches. HITL approaches usually deployed crowdsourcing platforms to produce IDs at a lower cost, which makes it possible to generate ID on a large scale. In this section, we will first briefly introduce HITL, crowdsourcing and existing crowdsourcing projects for ID. Based on this, we will discuss the focus of this project and its relevance with the generation of ID in the context of HITL.

[9] Salisbury, E., Kamar, E. and Morris, M. R. (2017) 'Conversational Crowdsourcing as a Tool', Aaai Hcomp 17, (Hcomp), pp. 147–156

2.4.3.1 Introduction of HITL

Crowdsourcing involves recruiting large groups of people online to contribute small amounts of effort towards a larger goal. Before the invention of electronic computers, organizations employed teams of "human computers" to perform various mathematical calculations. [36] Within the past decade, this notion of human computation has once again gained popularity. This is not only due to the increase of crowdsourcing platforms, but also because researchers have become better able to understand the limitations of machine computation. When the machine or computer system is unable to offer an answer to a problem, human intervention is needed. [37] This is the basic idea of human in the loop approach. Image description through crowdsourcing is definitely an example of this. See Figure 2.14 (on the next page) for an overview of the experiments

[36] Olson, J. S. and Editors, W. A. K. (no date) Ways of Knowing in HCI.

[37] Grier, D. A. (2013). When computers were human. Princeton University Press.

2.4.3.2 Early Explorations

In 2004, Von Ahn et al. first introduced the idea of human-powered captioning, using the ESP game to motivate online workers to create tags for images. Takagi et.al. have developed tools that allow readers online to improve accessibility, which demonstrated that readers online could be used to assess and improve accessibility barriers online [62]. And

[62] Brady, E. et al. (2013) 'Investigating the Appropriateness of Social Network Question Asking as a Resource for Blind Users'.

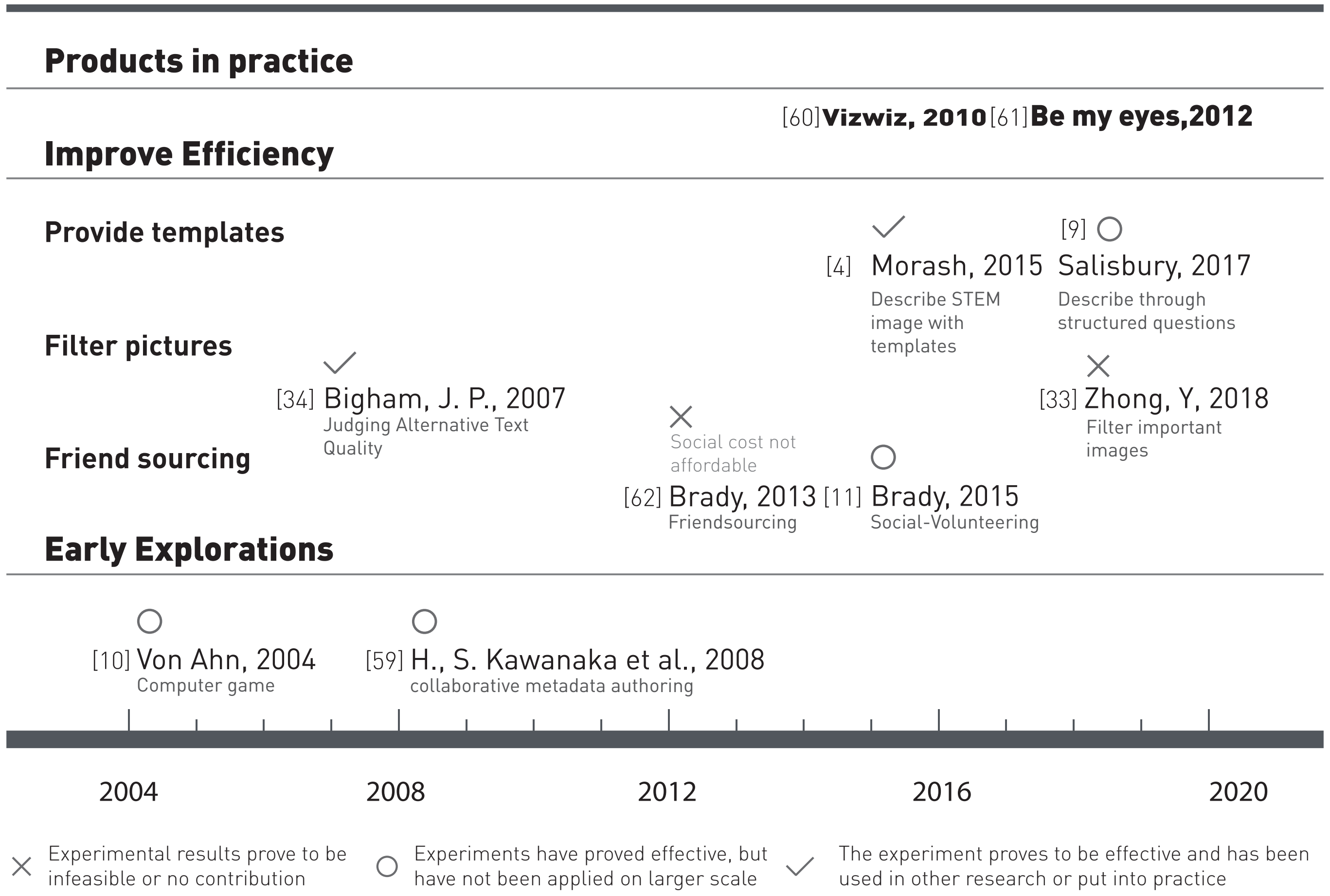

Products in practice
[60] Vizwiz, 2010
[61] Be my eyes,2012
Improve Efficiency
Provide templates
[4] Morash, 2015
Describe STEM image with templates
[9] Salisbury, 2017
Describe through structured questions
Filter pictures
[34] Bigham, J. P., 2007
Judging Alternative Text Quality
[33] Zhong, Y, 2018
Filter important images
Friend sourcing
Social cost not affordable
[62] Brady, 2013
Friendsourcing
[11] Brady, 2015
Social-Volunteering
Early Explorations
[10] Von Ahn, 2004
Computer game
[59] H., S. Kawanaka et al., 2008
collaborative metadata authoring
2004
2008
2012
2016
2020
Experimental results prove to be infeasible or no contribution
Experiments have proved effective, but have not been applied on larger scale
The experiment proves to be effective and has been used in other research or put into practice

Figure 2.14 An overview of the recent experiments of generating image description through HITL approach

Bigham showed that crowdsourced image labels could be created in near-real-time with their system Vizwiz. [60] Even though the potential value of crowdsourcing in ID was proved by early explorations, the utilization of HITL approaches will inevitably incur monetary (and accuracy) costs, which prevent these systems from being scalable.[7] Therefore, how to improve the efficiency of the system become the next focus.

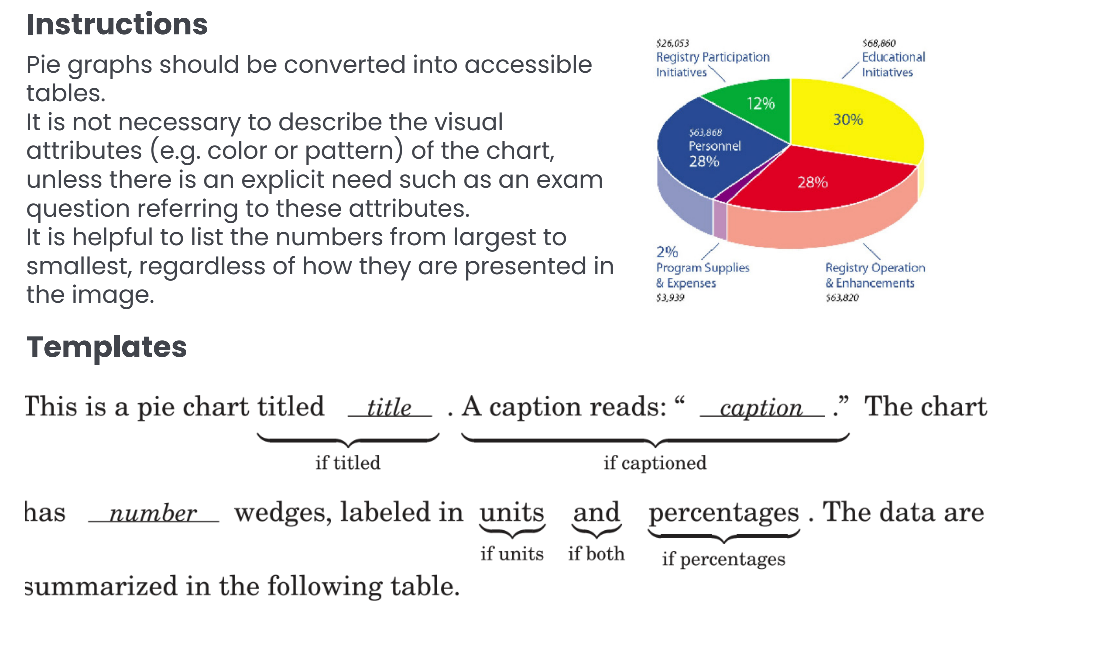
2.4.3.3 Review of recent experiments

To improve the cost-efficiency of the HITL approaches, different directions has been explored:

1. **Employ the "free" crowd workers.** The first direction is to find "free" volunteers to answer the request from VIPs and thus the cost was exempted. The success of Brady's micro-social volunteering system [11] and Be My Eyes [61] proved that the required people's altruism can be well benefits and benefits VIPs. But it is also pointed that the procedure should be cautiously designed in case of the social cost become unaffordable for VIPs. In addition, experiments in this direction are mostly used to solve simple visual questions (e.g., what is written on this jar?), whether it can be applied to describing images from publications is still questionable.

2. **Reduce the workload for the crowd worker**. Another direction is to filter the most necessary image to describe and thus reduce the workload for the crowd workers. Bigham has developed a system to judge the ID quality and thus filter those who need enhancement [34]. However, how to filter the informative images among them is not addressed. Judging the importance of an image requires evaluation of the image-text relationship, and could also be part of the crowd workers' description tasks. Zhong conducted an experiment to employ crowd workers to filter the informative images [33], but whether it can improve the cost-efficiency was not measured.

3. **Provide more supports.** The third direction is to provide more supports to the crowd workers so that their efficiency can be improved. Morash has conducted a set of controlled experiments and demonstrated it is better to generate ID using templates than using explicit instructions [4]. (Figure 2.15 comparison of instructions and templates) Likewise, Salisbury collected the requirements of VIPs and summarized them into a question list, which is proved to be able to improve both the efficiency



Instructions

Pie graphs should be converted into accessible tables.
It is not necessary to describe the visual attributes (e.g. color or pattern) of the chart, unless there is an explicit need such as an exam question referring to these attributes.
It is helpful to list the numbers from largest to smallest, regardless of how they are presented in the image.

$26,053
Registry Participation Initiatives
12%
$68,860
Educational Initiatives
30%
$63,868
Personnel
28%
28%
2%
Program Supplies & Expenses
$3,939
Registry Operation & Enhancements
$63,820

Templates

This is a pie chart titled _title_ . A caption reads: " _caption_ ." The chart has _number_ wedges, labeled in units and percentages . The data are summarized in the following table.

Figure 2.15 a comparison between the template and instruction which can be used to assist crowd workers

The third direction serves as the basis of this thesis. To meet the requirements of Dedicon and make the production of ID more scalable, the results of this project should not only satisfy VIPs' needs but also benefit crowd workers.

This project is thus aimed to **collect VIPs' requirements on ID for publications and transform the requirements into straightforward tasks for crowd workers**.

[60] Bigham, J. P. et al. (2010) 'VizWiz : Nearly Real-time Answers to Visual Questions'

[7] Guinness, D., Cutrell, E. and Morris, M. R. (2018) 'Caption Crawler: Enabling reusable alternative text descriptions using reverse image search', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3174092.

[61] Avila, M. et al. (2016) 'Remote assistance for blind users in daily life: A survey about be my eyes', ACM International Conference Proceeding Series, 29-June-20. doi: 10.1145/2910674.2935839.

[11] Brady, E., Morris, M. R. and Bigham, J. P. (2015) 'Gauging receptiveness to social micro-volunteering', Conference on Human Factors in Computing Systems - Proceedings, 2015-April, pp. 1055–1064. doi: 10.1145/2702123.2702329.

[34] Bigham, J. P. (2007) 'Increasing Web Accessibility by Automatically Judging Alternative Text Quality', pp. 349–352

[33] Zhong, Y., Matsubara, M. and Morishima, A. (2018) 'Identification of Important Images for Understanding Web Pages'. IEEE, pp. 3568–3574.

[4] Morash, V. S. et al. (2015) 'Describe STEM with Template --- Guiding novice web workers in making image descriptions using templates', ACM Transactions on Accessible Computing, 7(4). doi: 10.1145/2764916.

2.4.3.4 Description is a dynamic process

So why templates and question lists work better than the existing guidelines? Miltenburg et.al. 's findings may provide the answer. They used eye-tracking techniques to learn the procedure of ID generation. According to their research, people generate descriptions as they are interpreting the image and self-correcting their descriptions [1]. Building on this, it is reasonable to infer that a simple and step-by-step procedure to generate ID can effectively reduce crowd workers' cognitive load and simplify their work, thus improving their efficiency.

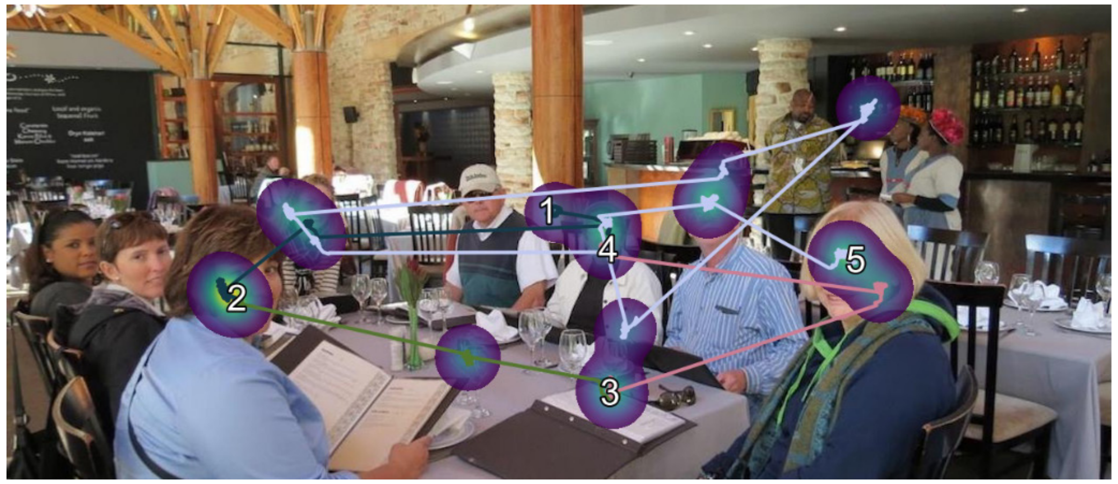
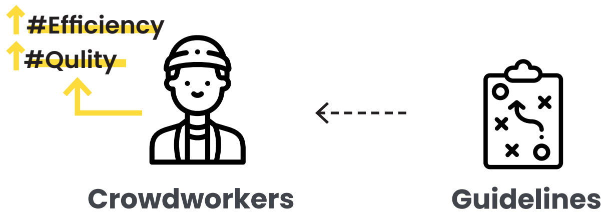


Figure 4.3 Eye-tracking experiment results from Miltenburg. Numbers indicate the following: 1. Start of experiment, 2. Speech onset, 3. Speaker realizes her mistake: the group hasn't ordered yet, 4. Start of corrected description, 5. End of description.

2.4.4 Product in practice

In addition to academic research, there are also products in life that use crowdsourcing to help visually impaired people. Typical examples are Vizwiz and Be my eyes. The first one allows users to take a picture with their phone, speak a question, and then receive multiple spoken answers [60]. Be My Eyes connects VIPs with untrained volunteers through a free-of-charge service. Volunteers can communicate and solve problems through real-time video. [61]

On the one hand, these applications are designed to answer simple questions in real life and may not be adaptable to ID in publications. On the other hand, the popularity of these applications proves the value of altruistic gains in such a crowdsourcing setting.



Summary

1. HITL approaches have proven its value to generate ID for VIPs. However, due to the limitations of monetary (and accuracy) costs, follow-up work is still needed to make it scalable.

2. Simplification of the image description tasks can significantly improve the efficiency and quality of the description.

[1] Miltenburg, E. Van (no date) Pragmatic factors in [automatic] image description

[60] Bigham, J. P. et al. (2010) 'VizWiz : Nearly Real-time Answers to Visual Questions

[61] Avila, M. et al. (2016) 'Remote assistance for blind users in daily life: A survey about be my eyes', ACM International Conference Proceeding Series, 29-June-20. doi: 10.1145/2910674.2935839.

MOVING-ON

Conclusions for this chapter:

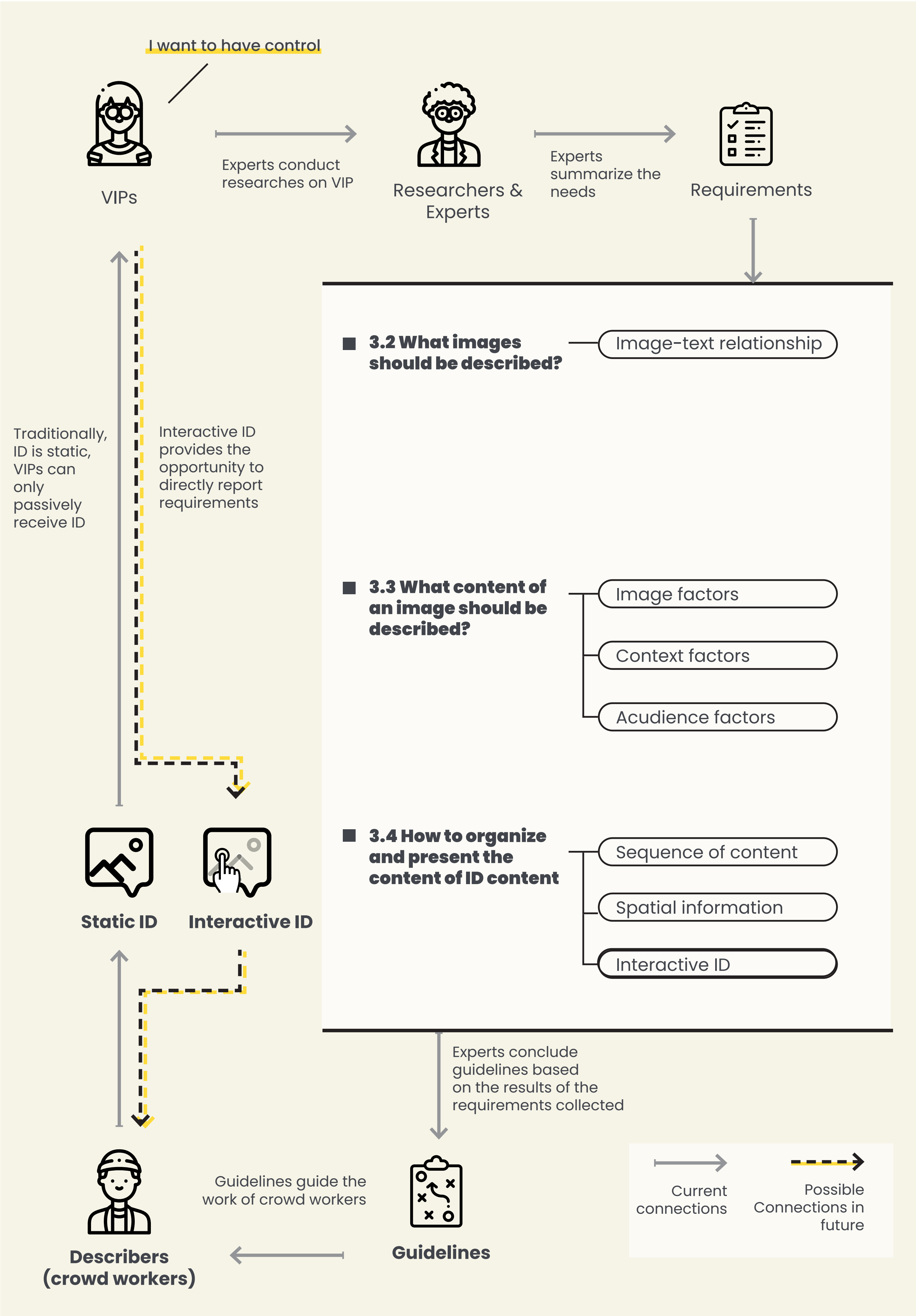
1. The impact of visual impairment is multi-faced. VIPs consume images not only for accessing necessary information, but also for entertainment and social needs.
2. ID is essential for VIPs accessing information within the visual content. There are different layers of image description. Current AI captioning is not enough for generating ID for VIPs. Generating IDs through HITL approaches is the best trade-off between efficiency and effectiveness in the near future.
3. Better supports can significantly help crowd workers improve their work efficiency. Thus, a refined researched goal can be drawn from the content of this chapter:

***Refined research goal**: To develop new knowledge about requirements on image description from the perspective of VIPs, which can be transformed into straightforward and simple tasks that can benefit crowd workers and improve their efficiency.*

Questions for the next chapters

What are the requirements on ID from the perspective of the VIPs?

1. What are the existing researches?
2. What is still lacking and what can be improved?
3. What are the suggestions from the experts and feedbacks from the VIPs?



I want to have control

VIPs

Experts conduct researches on VIP

Researchers & Experts

Experts summarize the needs

Requirements

- 3.2 What images should be described?
  - Image-text relationship
- 3.3 What content of an image should be described?
  - Image factors
  - Context factors
  - Acudience factors
- 3.4 How to organize and present the content of ID content
  - Sequence of content
  - Spatial information
  - Interactive ID

Traditionally, ID is static, VIPs can only passively receive ID

Interactive ID provides the opportunity to directly report requirements

Static ID

Interactive ID

Experts conclude guidelines based on the results of the requirements collected

Guidelines guide the work of crowd workers

Describers (crowd workers)

Guidelines

Current connections

Possible Connections in future

03

Requirements on image description

This chapter presents a comprehensive review of the existing research results on VIPs' requirements on image description. Combined with inputs from experts and visually impaired participants, the direction of future work is discussed, serving as a basis for the design goal.

3.1 Introduction

In the last chapter, the background of the project is introduced in detail and a refined research goal is derived. As is noted in the introduction chapter, how to generate image description (ID) has become an increasingly popular topic in the past decades. There has been a set of guidelines and tools aimed at training describers as well.

This chapter is aimed to provide a preliminary answer to this question, based on existing guidelines, existing research results and my own field research results. To illustrate this topic in a detailed and logical way, the original question of "requirements" is divided into 3 sub-question, which corresponds to 3 stages of ID generation and can benefit the work of crowd workers from different perspectives:

- ***(RQ-3.1) Before description: What images should be described? (Before the production of ID)***
- ***(RQ-3.2) Generating description: What contents of an image should be described?***
- ***(RQ-3.3) After generating ID content: How to present the image description (ID)?***

The content of this chapter is organized by these 3 questions. Based on the results, the target of this chapter is to find the niche for future research and potential design opportunities.

3.1.1 Methods

The methods used for the research of this chapter include desk research, literature research, and field research.

The contents of desk research mainly refer to the content of an image description training tool named POET (Link: Poet Image Description - How to Describe (diagramcenter.org)), which is developed by the Diagram Center for training ID experts. It provides a detailed and complete tutorial to help novices gradually learn how to describe images in educational publications. POET consists of two modules of training. The 2 modules are "When to describe" (RQ-3.1) and "How to describe" (RQ-3.2 & 3.3). Other guidelines are also reviewed as supplements.

The literature research consists of a review of recent researches on relevant topics.

The contents of field research come from an interview with experts (from Dedicon and KB) and target users. The interview is conducted through online meetings and phone calls.

***Terminology: Context and surrounding text**

According to Cambridge University, "context" has 2 meanings relevant to this project.:

- the situation within which something exists or happens, and that can help explain it
- the text or speech that comes immediately before and after a particular phrase or piece of text and helps to explain its meaning

In fact, both of the meaning is frequently used in this project, especially in this chapter. To clarify, we will use context referring to the first meaning, surrounding text for the second one.

3.2 What images should be described?

3.2.1 Definition

This section discusses the question" What images should be described?". Or more specifically, in what situation an image should be described. As was found by Petrie et.al, all of the interviewed VIPs agreed that not all of the images need description. [3] If there are clear indicators that help judge the necessity of describing a certain image, it will no wonder undoubtedly improve the efficiency of the describer and reduce the total cost of ID.

[3] Petrie, H. et al. (1999) 'Describing images on the Web : a survey of current practice and prospects for the future Centre for Human Computer Interaction Design City University London Northampton Square 2 The importance of describing images on the Web'.

3.2.2 Aim and relevance

The aim of researching this question is to find clear **INDICATORS** which can help crowd workers (or even machine) decide whether to describe an image, so that the number of the images to be described can be reduced without missing crucial information.

3.2.3 Findings

3.2.3.1 Decorative and informative

The earliest answer to this question (RQ3.1) comes from Petrie's research in 2005, which summarizes the types of web images that most VIPs think don't need to describe on the websites [3]: Decorative images/ Bullets or spacers/ Logos/ Images that are described in the text.

The images that require descriptions are mainly informative images [3]. However, there are barely any follow-up researches clearly defining informative images. But this is obviously related to the relationship between the image content and the surrounding text, as is pointed by most existing guidelines as well.

[3] Petrie, H. et al. (1999) 'Describing images on the Web : a survey of current practice and prospects for the future Centre for Human Computer Interaction Design City University London Northampton Square 2 The importance of describing images on the Web'.

3.2.3.2 Surrounidng text is the key

According to the POET guidelines, the surrounding text is also the key. It is suggested that the following conditions can be used to judge whether an image needs ID:

- **The purpose of an image.** According to POET guidelines, there are 3 the purposes of an image (visual interests, functional image (i.e., icon, button, link, etc.) and provide information for understanding subjects) [65]. Only the third one needs ID.

- **Unique knowledge.** The image should provide information that is essential and not available in the surrounding text.

In summary, both of them are implicit and requires describers to make subjective judgements according to the surrounding text. Therefore, even with the help of guidelines, empirical knowledge is still vital to get proficient and generate qualified ID efficiently, which is not enough to supports crowd workers' work.

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

3.2.3.3 Taxonomy of text-image relationships

In the field of information retrieval, Marsh et.al. has developed a set of taxonomy to describe the function of images [64]. The 46 possible functions of the picture can be divided into three major categories and 11 sub-categories, as is shown in Table 3.2.1.

A Functions expressing little relation to the text	B Functions expressing close relation to the text	C Functions that go beyond the text
A1 Decorate A1.1 Change pace A1.2 Match style A2 Elicit emotion A2.1 Alienate A2.2 Express poetically A3 Control A3.1 Engage A3.2 Motivate	B1 Reiterate B1.1 Concretize B1.1.1 Sample B1.1.1.1 Author/Source B1.2 Humanize B1.3 Common referent B1.4 Describe B1.5 Graph B1.6 Exemplify B1.7 Translate B2 Organize B2.1 Isolate B2.2 Contain B2.3 Locate B2.4 Induce perspective B3 Relate B3.1 Compare B3.2 Contrast B3.3 Parallel B4 Condense B4.1 Concentrate B4.2 Compact B5 Explain B5.1 Define	C1 Interpret C1.1 Emphasize C1.2 Document C2 Develop C2.1 Compare C2.2 Contrast C3 Transform C3.1 Alternate progress C3.2 Model C3.2.1 Model cognitive process C3.2.2 Model physical process C3.3 Inspire

Table 3.2.1 Taxonomy of functions of images to the text Source: [64]

In addition, it is pointed out that:

1. **The same image may have different functions in a paragraph of text**
2. **The function of the same image in different text fragments will change**

According to Marsh's work, the determination of the image function is definitely a complicated task, since there are a number of image functions. Images and functions are not one-to-one correspondence as well. To precisely judge whether an image should be described or not, it requires an evaluation of image-text relationship. This evaluation cannot be done by machine, nor can it be judged by a clear "indicator" (at least it is far out of my expertise).

3.2.3.4 Different image type, different approaches

Although it's difficult to precisely access the importance of an image, it is possible to filter some images based on their types – certain types of images do not require to be described by HITL approaches. Table 3.2.2 summarized the widely used approaches to providing alternative text in websites:

[64] PMarsh, E. E. and White, M. D. (2003) 'A taxonomy of relationships between images and text', 59(6), pp. 647–672. doi: 10.1108/00220410310506303

Image function	Examples	Alt text needed	Accessible via semiauto-mated systems	Accessible via fully auto-mated systems
Informative images	–Simple pictures –Icons –Succinct images –Photographs –Images of text	○	○	○
Navigational images	–Logos –Image links	○	✕	✕
Functional images	–Button images –Icons	○	✕	○
Decorative image	–Bullet points –Borders –Fillers	✕	✕	✕

Table 3.2.2 Image types and alternative text approaches Source [8]

For images with sufficiently simple contents, AI has been able to describe them with reliable accuracy: text, color, etc. For example, for the tool WebInSight developed by Bigham, OCR technology was used to help describe certain images [55]. And Salisbury's experiment also utilized AI captions as references for crowd workers [9].

Gregorio Pellegrino's experiments employed the similar approach for images in the publications. Images were divided into twelve categories, as shown in the Figure 3.2.1. Before generating the description, the algorithm will first classify the type of the image and then adopt different description methods. (Figure 3.2.2)

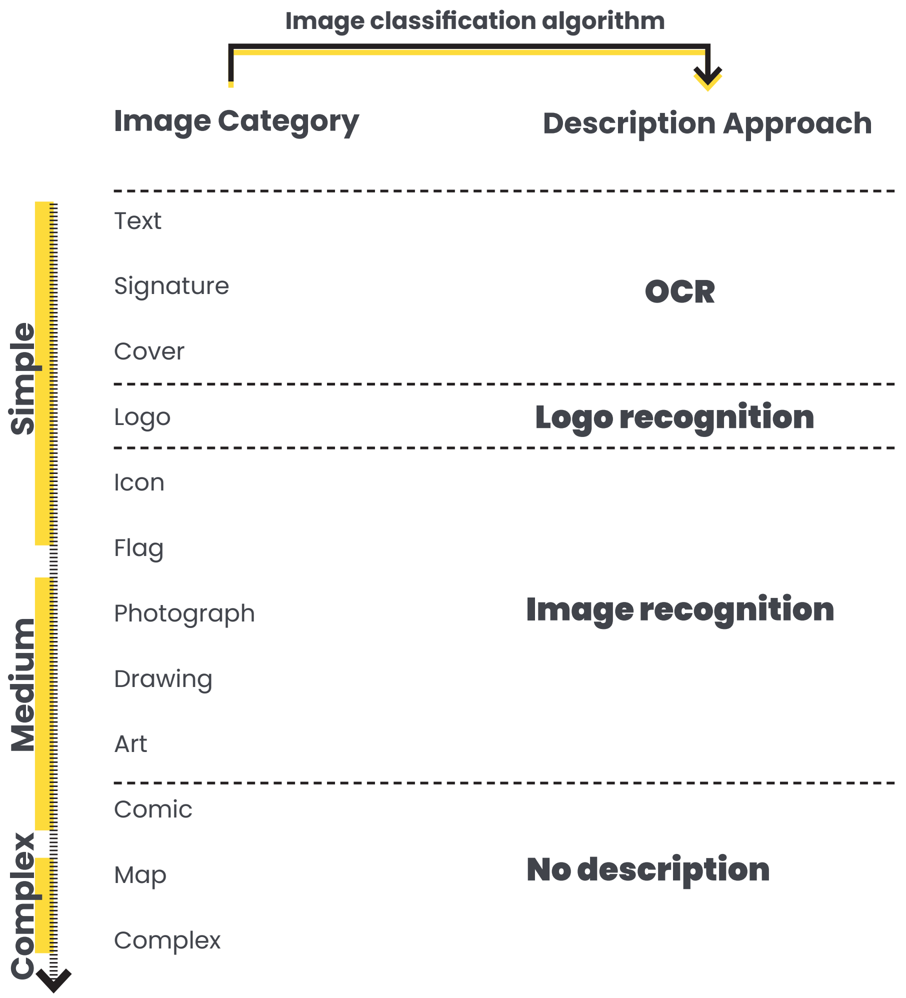


Image Category	Description Approach
Text	OCR
Signature	
Cover	
Logo	Logo recognition
Icon	Image recognition
Flag	
Photograph	
Drawing	
Art	
Comic	No description
Map	
Complex	

[55] Bigham, J. P. et al. (no date) '[01-4]WebinSitu : A Comparative Analysis of Blind and Sighted Browsing Behavior', pp. 51–58

[9] Salisbury, E., Kamar, E. and Morris, M. R. (2017) 'Conversational Crowdsourcing as a Tool', Aaai Hcomp 17, (Hcomp), pp. 147–156.

[73] Gregorio Pellegrino,2019, DPUB SUMMIT 2019 - 6 - Improving automatic image description in EPUB using Artificial Intelligence, https://www.youtube.com/watch?v=XZpgGNoBQoo&t=672s

3.2.3.5 Reuse of ID

Regarding what images should be described, another strategy to reduce the workload of image description while generating human-quality ID is the reuse of ID.

Bigham's WebinSight [13] system integrates a Web Context Labeling module, which would retrieve the title and header elements from a page linked to by an image to act as an alternative text description. Inspired by this, Darren et.al. developed a Caption Crawler [7] which reverse image search to find existing captions on the web and make them accessible to a user's screen reader. In addition, Kuppusamy et.al proposed a model AIMS [11], which utilizes metadata of images to build self-describing images for assisting screen reader users and thus eliminate the redundant IDs.

Even though the literature has pointed out that the requirements on ID are highly context-dependent, as we discussed in 3.4, the same description content can still be presented in various ways. Therefore, the reuse of ID is still a meaningful topic.

3.2.3.6 Inputs from Experts: users want control

According to inputs from Dedicon's experts, their existing workflow also determines whether to describe a picture by assessing the image-text relationship. At the same time, it was also mentioned that ID might have negative impacts on users' reading in some cases: "sometimes it takes away their attention from the flow of the story or the flow of the news". Therefore, it is suggested that users should have control over what images should be described, or at least an option to skip the description.

Summary

Findings:

- Whether the image should be described should be judged from its **relationship with surrounding text**: is its function to provide information; whether the information it provides is not available in the surrounding text.
- There are **no clear indicators** that can be applied to all images to directly determine whether they should be described or. However, we can filter some images according to the image's classification.

Design opportunity:

- To **give user control and collect metadata for further researches**.
- Establish a database and reuse ID when an image is repeatedly cited

[13] Bigham, J. P. et al. (2006) 'WebInSight: Making web images accessible', Eighth International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2006, 2006, pp. 181–188. doi: 10.1145/1168987.1169018

[7] Guinness, D., Cutrell, E. and Morris, M. R. (2018) 'Caption Crawler: Enabling reusable alternative text descriptions using reverse image search', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3174092.

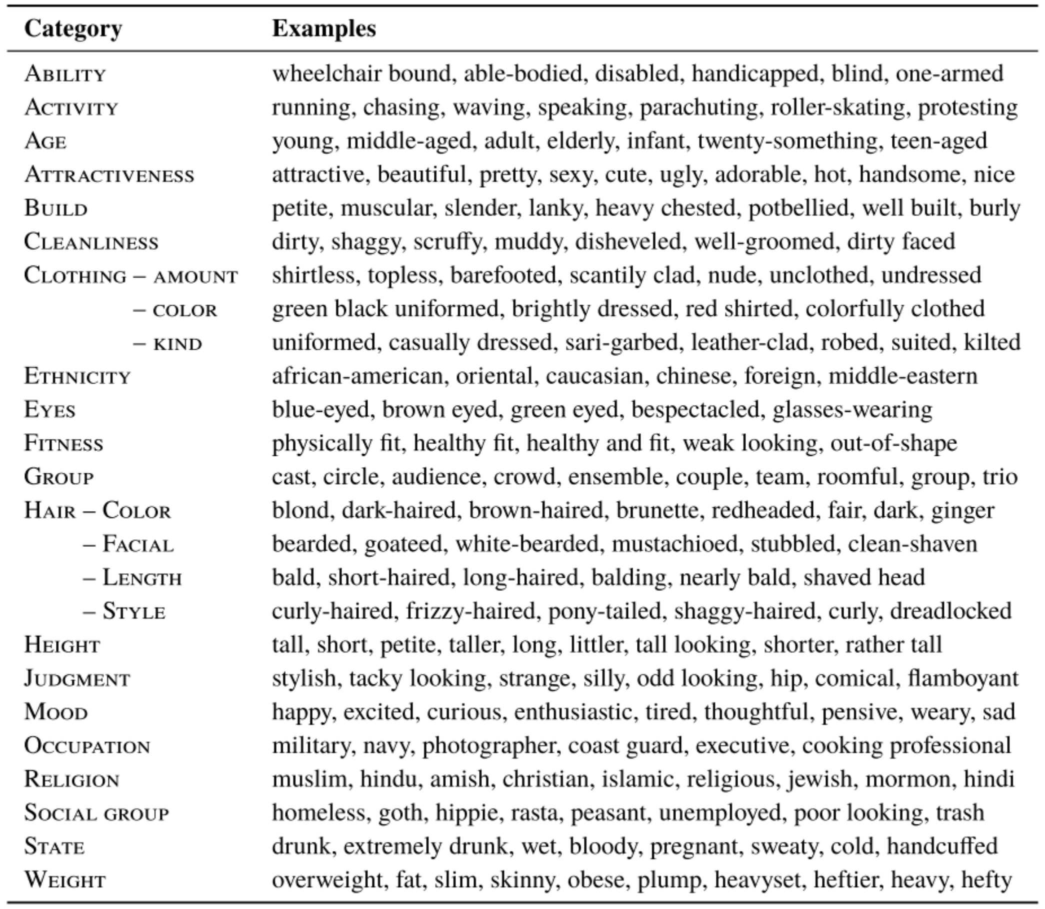
[11] Guinness, D., Cutrell, E. and Morris, M. R. (2018) 'Caption Crawler: Enabling reusable alternative text descriptions using reverse image search', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3174092.

3.3 What content of an image should be described?

3.3.1 Definition

An image can convey 1000 words.

When a describer encounters an image and decides to describe it, the most essential problem is: how to describe the image? There are different objects in a imagge, and each object may contain different details. As is shown in Figure 3.3.1, there are at least 25 categories of adjectives that can be used to describe humans [1]. There will also be corresponding interactions and spatial relationships between objects and objects. In addition, as is stated in 2.2.1, there are three different levels of picture description. In addition to the specific content in the image, the image description (ID) may also include Non-visual and perceptual content. Obviously, ID cannot completely cover all the contents. Therefore, instructions should be provided to the description to make them fully aware of the answer to this question: What content of an image should be described?



Category	Examples
Ability	wheelchair bound, able-bodied, disabled, handicapped, blind, one-armed
Activity	running, chasing, waving, speaking, parachuting, roller-skating, protesting
Age	young, middle-aged, adult, elderly, infant, twenty-something, teen-aged
Attractiveness	attractive, beautiful, pretty, sexy, cute, ugly, adorable, hot, handsome, nice
Build	petite, muscular, slender, lanky, heavy chested, potbellied, well built, burly
Cleanliness	dirty, shaggy, scruffy, muddy, disheveled, well-groomed, dirty faced
Clothing – amount	shirtless, topless, barefooted, scantily clad, nude, unclothed, undressed
– color	green black uniformed, brightly dressed, red shirted, colorfully clothed
– kind	uniformed, casually dressed, sari-garbed, leather-clad, robed, suited, kilted
Ethnicity	african-american, oriental, caucasian, chinese, foreign, middle-eastern
Eyes	blue-eyed, brown eyed, green eyed, bespectacled, glasses-wearing
Fitness	physically fit, healthy fit, healthy and fit, weak looking, out-of-shape
Group	cast, circle, audience, crowd, ensemble, couple, team, roomful, group, trio
Hair – Color	blond, dark-haired, brown-haired, brunette, redheaded, fair, dark, ginger
– Facial	bearded, goateed, white-bearded, mustachioed, stubbled, clean-shaven
– Length	bald, short-haired, long-haired, balding, nearly bald, shaved head
– Style	curly-haired, frizzy-haired, pony-tailed, shaggy-haired, curly, dreadlocked
Height	tall, short, petite, taller, long, littler, tall looking, shorter, rather tall
Judgment	stylish, tacky looking, strange, silly, odd looking, hip, comical, flamboyant
Mood	happy, excited, curious, enthusiastic, tired, thoughtful, pensive, weary, sad
Occupation	military, navy, photographer, coast guard, executive, cooking professional
Religion	muslim, hindu, amish, christian, islamic, religious, jewish, mormon, hindi
Social group	homeless, goth, hippie, rasta, peasant, unemployed, poor looking, trash
State	drunk, extremely drunk, wet, bloody, pregnant, sweaty, cold, handcuffed
Weight	overweight, fat, slim, skinny, obese, plump, heavyset, heftier, heavy, hefty

Figure 3.3.1 Taxonomy of labels referring to other people, with selected examples for each category. [1]

In this chapter, based on the results of research activities, rules are concluded, aimed to help crowd workers choose what content they should describe. The results should help crowd workers work more efficiently, while fully satisfying VIPs' information demand on image contents.

[1] Miltenburg, E. Van (no date) Pragmatic factors in [automatic] image description.

3.3.2 Aim and relevance

The aim of researching this question is to generate rules that can help crowd workers select the content to describe, instead of relying on their intuition or experience. At the same time, under the guidance of clear rules, the generated ID is expected to better satisfy VIPs' needs.

3.3.3 Findings

3.3.3.1Variability and variables

The most necessary information to be included in the description is context-dependent.[3] In 2005, Petrie has summarized a list of elements that require ID in the majority of cases:

- Objects, buildings, people in the image
- What is happening in the image?
- Purpose of the image
- Colors in the image
- Emotion, the atmosphere of the image
- The location depicted in the image

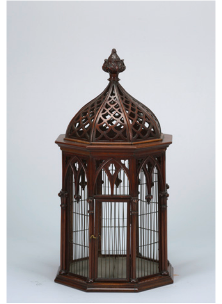


Figure 3.3.2 Octagonal wood and wire cage with carved decorative features [66]

This is far from comprehensive: in certain cases, some aspects are not necessary while some aspects need a richer description. Here is an example from the official sample from Cooper Hewitt Guidelines. Regarding the elements Petrie concluded, the example ID only includes information about the object itself (only one aspect), but information about the object's material and appearance are well elaborated [66]. The content of multiple works of literature has pointed out that the current one-approach-fits-all approaches cannot meet the diverse needs under different situations well.

Through a review of guidelines and literature, multiple factors affecting the demand for ID are found. According to the different subjects they correspond to, I divide them into three categories:

- **Image factors**
- **Context factors**
- **Audience factors**

3.3.3.1.1.1 Image factors

Image factors refer to the factors related to the image itself, including the type of image and the visual focus* of the image.

Different types of images need different kinds of descriptiona [65]. The description required for different types of images is also different. For example, for a photo, its setting and subject are the most important, but for drawings, the overall color and texture (perceptual level) of the image may also be critical. While the image is a comic with text, the text in the image may become the focus of the ID.

The visual focus of the image may influence the content required to be described as well. Through interviews with 28 VIPs, Stangl et.al. concluded a table (Table 3.3.1) to illustrate which specific content is needed in various situations [2]. According to their results, visual focus and source are used as the two main variables that affect the description. Visual focus determines most of the basic information that needs to be described, and establishes a clear framework for how to describe an image.

*Visual focus refers to the central focus of the image's visual content. In most cases, it represents the content people pay the most attention to when they see an image

[3] Petrie, H. et al. (1999) 'Describing images on the Web : a survey of current practice and prospects for the future Centre for Human Computer Interaction Design City University London Northampton Square 2 The importance of describing images on the Web'

[66] COOPER HEWITT GUIDELINES FOR IMAGE DESCRIPTION, no date, https://www.cooperhewitt.org/cooper-hewitt-guidelines-for-image-description

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404.

	News	SNS	ecommerce	Employment	E-Publicatio
Event/Scene					
People Present	○	○	○	○	○
Text	○	○	○	○	○
Activity	○	○		○	○
Interaction	○	○		○	○
Landmarks	○	○			○
Building Features	○	○		○	○
Weather	○	○			○
Lighting				○	
People					
Text	○	○	○	○	○
Salient Objects	○	○		○	○
Activity	○	○		○	○
Gender	○	○	○	○	
Race/Diversity	○	○		○	○
Name of Person	○	○			○
Celebrity Name	○	○	○		○
Expression	○	○		○	
Attire/Clean		○		○	
Body Shape/Size			○		
Pets		○			
Object					
Text	○	○	○	○	○
Name	○	○	○	○	○
Form	○	○	○		
Fit		○	○		
Color		○	○		
Overall Style		○	○		
Material	○	○	○		
Logos/Symbols		○	○		
Damage			○		
Unique Features			○		

Table 3.3.1 Cross-source description content requirements [2]

3.3.3.1.1.2 Context factors

Context factors refer to environmental factors of the image in addition to the image itself. Specifically, the surrounding text and the source of the image may affect the user's demand for ID content.

As mentioned in the previous section, Stangl's research proves the influence of **the source** on the demand for ID, which is clearly presented in Table 3.3.1. For example, for images with people as a visual focus, when in the context of dating websites, audience's requirements for appearance details have obviously increased. Dedicon's experts also expressed similar observations. Even for publications, the identity of a certain publication may also affect the description requirements.

The surrounding text may also affect the content that needs to be described. Just as it affects whether an image should be described, the surrounding text also determines what information should be included in the description (e.g. the concept included has been explained in the surrounding text), and how detailed the description should be (e.g.: the image is a critical part of the learning) [65].

3.3.3.1.1.3 Audience factors

Audience's personal factors may also affect their demand for ID.

The type of VI may affect the reliance on th description. Low-vision VIPs are reported to prefer to use their residual vision. Therefore, if some information can be obtained through vision, then there is no need to rely on the content of the ID.

The time of being visually impaired is also influential. For example, people who have ever had vision will try to paint the image in their minds and would probably rather "be on sensory overload" [2] *.

In addition to VI-related factors, audience's own experience and taste will also have an impact. When a person does not have prior experience with the content area or a similar cultural reference point (familiarity), a higher degree of detail (and/or additional modes of representation) may be needed. In addition, the amount of time available can also influence the length of the description VIPs to prefer [2]. Salisbury's studies also highlight the influence of users' interest levels [9].

3.3.3.1.1.4 Summary

in this section, a set of factors are found to be influential for the demand for ID. Essentially, Stangl's table provides a good foundation for establishing a mapping between these factors and description content, which is possibly of great help for crowd workers when it is converted into templates or guidance.

3.3.3.2 Amount s of details

Regarding how to generate ID, in addition to knowing what content to describe, how detailed the content should be described seems to be another important issue. As is in 3.3.3.1, situations are frequently mentioned when users have requirements that require more details. However, for crowd workers, how to describe a picture in "more detail"? In this section, we discuss this issue from two perspectives.

On one hand, how detailed the description needs to be can **be transformed** into "what content may need to be described". "To describe in detail" is an implicit requirement, and there are always huge differences among how different describers implement it. In fact, adding the details is equivalent to adding the content of the description. Take the following picture as an example [66]. A short description might be: "*A man stands across from us in a wallpapered room*". And a detailed version might be: "*A light-skinned man with dark hair and a beard wears all black with a light grey overcoat. He is standing in a room with light-blue wallpaper*". The detailed description additionally describes the person's skin color, clothes, and the color of the wallpaper. Therefore, questions about details can be transformed into questions about content. Moreover, compared to "a more detailed description is needed", it is a clearer and more helpful way to directly indicate what content needs to be described in the instructions.

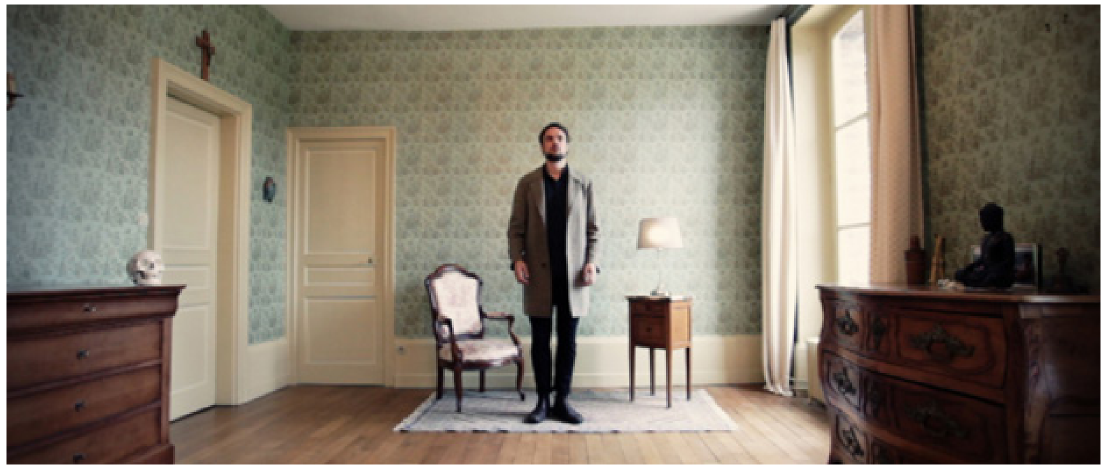


Figure 3.3.3 A man stands across from us in a wallpapered room [66]

On the other hand, **the total number of ID characters** is another condition that limits the detailed description of the picture, but existing research has not reached an agreement on this matter. Descriptions that are too long can be tedious to read. [65] But as to how long a description will cause fatigue, there is no strong indication of the optimum length. Different views vary from 2 or 3 words [68], to 150 characters [67] or even more. For Dedicon, the question about how much information the ID should contain is also a question that Dedicon is still looking for answers. After all, this may be highly related to the audience's interest and the quality of the ID. Therefore, it is difficult to have a fixed answer. The solution suggested here is to hand over control to the audience.

3.3.3.3 Semantic factors

Bringing interpretive knowledge to a description is not always preferred. Generally, it is suggested that the content of ID should be limited to what can be seen when looking at the image [65]. However, in some cases, interpretive knowledge definitely helps the visual understanding of the images and thus this rule should be incorporated.

For instance, POET mentioned the following rules for being objective:

- Describe physical appearances rather than emotions and intentions
- Do not interpret the material and allow readers to form their opinions
- Do not omit uncomfortable content like politics or sex.

But according to the literature, at least emotion and intention are the information VIPs would want to know.[2]

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) '"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404.

[9] Salisbury, E., Kamar, E. and Morris, M. R. (2017) 'Conversational Crowdsourcing as a Tool', Aaai Hcomp 17, (Hcomp), pp. 147–156.

[66] COOPER HEWITT GUIDELINES FOR IMAGE DESCRIPTION, no date, https://www.cooperhewitt.org/cooper-hewitt-guidelines-for-image-description

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

[68] Hudson, R (2003). Text Alternatives for Images. Retrieved 27 February 2005, from http://www.usability.com.au/resources/image-text.cfm

[67] Slatin, J., & Rush, S. (2002). Maximum accessibility: Making the web more usable for everyone

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) '"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404.

In addition to the risk of incorrect information, subjective inferences can also be controversial (e.g., gender, race). Different institutions have different approaches to this problem. For example, Google chose to avoid using inferences regarding races. While most image description guidelines suggest, when gender is clearly performed and/or verifiable, it should be described, which is also approved by experts from Dedicon. [65, 66]

To conclude, a fixed rule may not be applicable to all situations. After all, describers still need to make judgments based on the situation by themselves. However, the following instructions can be offered as assistance:

1. Provide instructions to **raise the describers' awareness** of possible subjective inference, and ensure that their description can be fully verified

2. Regarding ethnicity, guide describers to use **non-ethnic terms** such as "light-skinned" or "dark-skinned" when clearly visible [66].

3.3.3.3.2 Jargon

Avoid jargon or other kinds of privileged knowledge except where it is essential for describing an object. If used, the jargon term should be explained [66].

3.3.4 Summary: One-fits-all approach is not enough

Findings:

- Affected by various factors, the content that users need to describe is varied. The description required by the user not only changes with the content of the image, but is also **context-dependent and personalized.**
- A mapping between image & context factors and ID content can be established, which can be transformed into clear guidance for crowd workers. However, regarding the impact of the user's personal factor, there is no such set of rules.
- The requirement for details can be transformed into the need for more content to be described.
- Crowd workers may need guidance to help them keep aware of their subjective inference and improve their presentation

Design opportunities:

- Provide guidance for crowd workers through existing research on image and context factor
- Let VIPs choose what they want and collect their demand data
- Provide crowd workers with tips on subjective inference

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

[66] COOPER HEWITT GUIDELINES FOR IMAGE DESCRIPTION, no date, https://www.cooperhewitt.org/cooper-hewitt-guidelines-for-image-description

3.4 How to organize and present the ID content

3.4.1 Definition

In the previous chapter, we discussed what content of an image needs to be described, i.e., in a given situation, what information needs to be included in the image description (ID). There is another important part left to be addressed: how to present the image description? It includes the sequence of how information is delivered, so that the content of ID more logical and can be effectively conveyed (e.g., from general to specific). The medium and form through which ID is communicated to visually impaired people (VIPs) enables ID to exert its greatest value. Therefore, how to organize and present the ID content is discussed in this section, aimed to conclude a proposal which can best support VIPs consume ID.

3.4.2 Aim and relevance

The aim of discussing this question is to find rules and solutions which can help VIPs consume the content of ID most efficiently and thus crowd worker the generated ID can best embody its value.

3.4.3 How to organize the content

The order of description contents is emphasized in almost all guidelines. ID should start with high-level context and then drill down to details that enhance understanding [65]. This provides the reader with options about how much information to read and helps them to form a structure of knowledge.

In addition, when contains the enormous necessary information, ID segment content should be organized in logical and digestible ways. [65] To be more specific, the elements of ID can be organized according to their orientation and relationships. When a person is the subject of an image, it can be used as a surrogate to describe relationships. (e.g., to the person's right is...)

3.4.4Spatial information

Although we can use language to describe the spatial relationship of the elements in the image (A is to the left of B), this does not mean that textual ID can always convey the spatial information in the image well, which in fact is an inefficient way. When there are multiple elements in the image, only using text to describe the spatial information makes the ID excessively verbose and leads to cognitive fatigue. For example, when a flowchart similar to (Figure 3.4.1) is described, each element may need at least two sentences to indicate its relative relationship with other elements, such as *Middle: back to "former"; forward to "later"*. This not only makes the communication of information inefficient but also completely loses images' function of organizing and simplifying information. Therefore, POET guidelines suggest a complex image should be converted

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

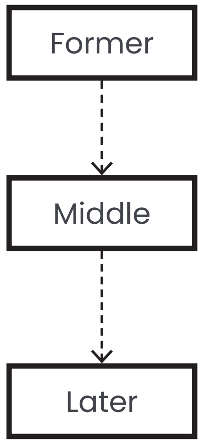
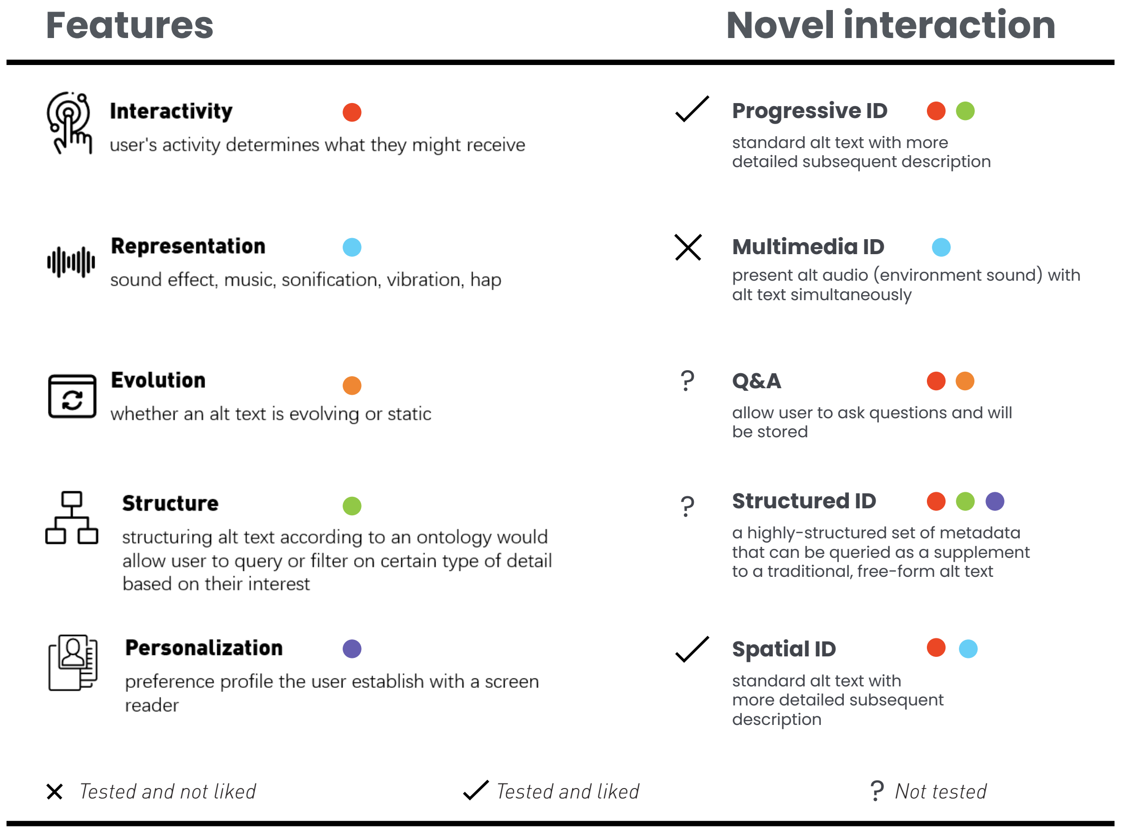


Figure 3.4.1 A flowchart used to illustrate the trounble to desribe flowchart

into a table or a tactile version [65], which allows VIPs to use alternative sensations to perceive spatial information. In addition, the researcher has also carried out various explorations in this field, including using a tactile model to explore artworks [69], labeling objects to help VIPs explore the spatial layout within the image [74], which all receive good feedback. Currently, these experiments are limited to specific usage scenarios (museum) or require specific equipment. If a more universal method (such as using the vibration feedback of a mobile phone) can be used, it will significantly improve the experience of receiving complex picture information in daily life

3.4.4 Rich presentations

The efforts of rich presentation of ID actually are not limited to the experiments about conveying spatial information. Because of the VIPs' diverse requirements on ID and the limitation of one-shot textual content, researchers have proposed a set of ideas to enrich the affordance of the ID. Morris et.al proposed a taxonomy of possible design space of ID presentation, as is shown in Figure 3.4.2 [17].



Features			Novel interaction		
Interactivity	user's activity determines what they might receive	🔴	✓	**Progressive ID** standard alt text with more detailed subsequent description	🔴 🟢
Representation	sound effect, music, sonification, vibration, hap	🔵	✕	**Multimedia ID** present alt audio (environment sound) with alt text simultaneously	🔵
Evolution	whether an alt text is evolving or static	🟠	?	**Q&A** allow user to ask questions and will be stored	🔴 🟠
Structure	structuring alt text according to an ontology would allow user to query or filter on certain type of detail based on their interest	🟢	?	**Structured ID** a highly-structured set of metadata that can be queried as a supplement to a traditional, free-form alt text	🔴 🟢 🟣
Personalization	preference profile the user establish with a screen reader	🟣	✓	**Spatial ID** standard alt text with more detailed subsequent description	🔴 🔵

✕ *Tested and not liked* ✓ *Tested and liked* ? *Not tested*

Figure 3.4.2 Design space and noveal interactions of image description, see demos on https://www.youtube.com/watch?v=gE7OToBouPg&ab_channel=MeredithMorrisMeredithMorris

Morris et.al also integrated the design directions into novel interactions, and tested 3 of them with VIPs, as is shown in Figure 3.4.2

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

[69] Bartolome, J. I. et al. (2019) 'Exploring aRt with a voice controlled multimodal guide for blind people', TEI 2019 - Proceedings of the 13th International Conference on Tangible, Embedded, and Embodied Interaction, pp. 383–390. doi: 10.1145/3294109.3300994

[74] Zhong, Y. et al. (2015) 'Regionspeak: Quick comprehensive spatial descriptionsof complex images for blind users', Conference on Human Factors in Computing Systems - Proceedings, 2015-April, pp. 2353–2362. doi: 10.1145/2702123.2702437

[17] Morris, M. R. et al. (2018) 'Rich representations of visual content for Screen reader users', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3173633.

The results indicated that *progressive ID* receives the best feedback because VIPs **enjoyed the ability to choose how many and which levels of details to listen to**. *Spatial ID* was also like by the participants because of understanding the location of objects within the images, which is aligned with our discussions in the last section. However, *Multimedia ID* got negative reviews, since the integration of extra auditory information (e.g., music) is thought to interfere with the ability to comprehend the primary alt text [17]. The concepts Morris did not test are inspiring as well, such as *structured ID* which is promising to help for users navigation among information.

To conclude, Morris' experiment proved the potential of *interactive ID* to enhance the VIP experience and how capabilities of modern technology can be leveraged to provide a rich and evocative experience [17]. They are also inspiring since t**hey provide opportunities for VIPs to actively choose and decide on the content of ID.**

At the same time, as is noted by Morris, their experiments were just "the first study of this space" and only focus on subjective metrics. Their research was not presenting the image together with its context, nor did they consider time, cost, long-term use value and other factors

3.4.6 Summary

Findings:

- ID should usually follow the order from general to specific. When there is more content, the size and position relationship of the elements can be used in the image to help VIP "paint the image" in the mind.
- For spatial information, alternative techniques (e.g., table/tactile material) rather than only textual information could be used to convey the spatial information in the image.
- Utilization of modern technology can enable ID to be interactive and improve the experience of reading images. Further research is needed to verify their value within the reading context.

Design opportunities:

- Compared to one-shot textual description, interactive ID can significantly improve VIPs' experience.

[17] Morris, M. R. et al. (2018) 'Rich representations of visual content for Screen reader users', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3173633.

3.5 Synthesis of current guidelines and researches

In this chapter, we comprehensively discuss the requirements of VIPs for ID through three separate questions.

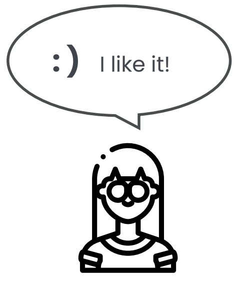
Regarding What images should be described, the surrounding text is widely regarded as the key factor to decide whether an image should be described, that is, whether the image is an informative or a decorative image [3, 65]. However, due to the complexity of the image-context relationship[64], there are no indicators that can directly evaluate the importance and subjective judgment of the crowd workers still plays an important role. In addition, the appearance of ID may interrupt the rhythm of the narrative (regardless of the image is informative or decorative). Therefore, ideally, users should have control over this and further researches are needed from the perspective of the VIPs.

Regarding what content of the picture should be described, it is affected by various factors from the image itself, the context of the image, and personal factors of the audience, and thus the current one-fits-all approach is not enough[2, 65, 9]. So it is definitely context-dependent and personalized. Based on the current research results, it's possible to establish a preliminary mapping between the image & context factors and ID content needed to be described (Table 3.3.1). In addition, the content of ID should be delivered in a personalized way to meet the nuances among VIPs' requirements (public factors).

The third question is how to organize and present ID. On one hand, the content of textual ID should be delivered in a logical and digestible way. For example, ID should always be presented from general to specific levels [65]. And the description for multiple elements should be organized according to their locations [66]. On the other hand, the studies of interactive ID provide a new perspective for enhancing the user's experience of reading images [17]. For example, progressive ID is used to provide users with the autonomy to control the length of the ID or haptic feedbacks to help users perceive the spatial information of the image.

3.5.1 Relevance to stakeholders

The research goal proposed in the last chapter was to collect new knowledge of the visually impaired people (VIPs) so as to benefits the crowd workers and simplify their tasks. There are definitely a set of opportunities found for future improvements, but currently, the most pressing problem is the insufficient distribution of ID. To address this problem, we need to **reduce the complexity (by providing guidance) and work amount (by filtering the unnecessary images)** of crowd workers' tasks. From the perspective of crowd workers, the existing research results can help crowd workers decide which content of a certain image needs to be described based on contextual and image factors (at least Stangl et.al. research provides a framework for this). Meanwhile, structured ID provides an opportunity to deconstruct ID and directly link ID contents to a certain category (and thus the variables).



On the other hand, from the user's point of view, **giving them control** is consistently preferred, because the existing research results indicate that there are no fixed and comprehensive answers to clarify VIPs' requirements. Interactive ID provides the possibilities for this vision: progressive ID can help them choose whether to listen to longer content, structured ID can help them **acquire what they want at their own will and pace**.

3.5.2 A new data collection approach is needed

In addition, through the review of the literature, it is found that most of the existing research on the demand for VIPs is done in a qualitative way [2, 3, 5, 6, 71, 9]. Considering the complexity of the image-context relationship, the variety of image types, and the diversity of the VIPs group, it seems that it is difficult to derive complete and specific conclusions within a limited number of samples. In order to study how to answer visual questions (describe photos taken by VIPs), the well-known dataset Vizwiz was established to serve as the basis for subsequent research [60,72], which becomes a very rich source of information about the domains that VIPs are interested in [1]. In order to better understand how to describe the pictures in the text (i.e., the three questions discussed in this chapter), a plentiful collection of ID samples, which includes attributes of context, ID, audience and ID itself, is needed. From this perspective, interactive ID enables users to actively choose length and content, which sets the conditions for researchers to collect their preference data on a larger scale.

In summary, interactive ID, i.e., progressive and structured ID, can not only improve users' experience but also benefits crowd workers' and researchers' work

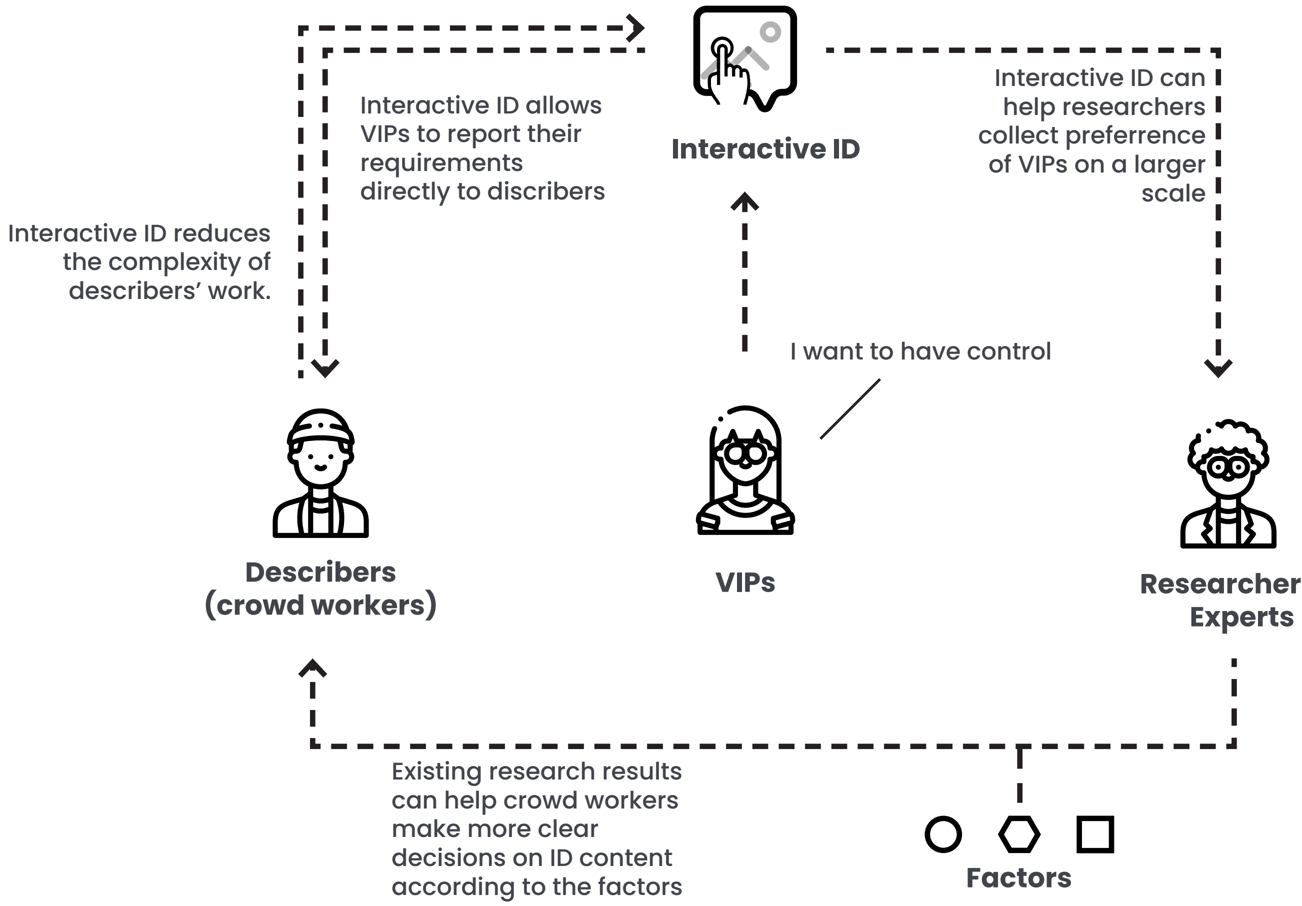

Interactive ID
Interactive ID allows VIPs to report their requirements directly to discribers
Interactive ID can help researchers collect preferrence of VIPs on a larger scale
Interactive ID reduces the complexity of describers' work.
I want to have control
Describers (crowd workers)
VIPs
Researchers & Experts
Existing research results can help crowd workers make more clear decisions on ID content according to the factors
Factors

Figure 3.4.3 A illustration showing how Interactive ID can benefit multiple stakeholders

[3] Petrie, H. et al. (1999) 'Describing images on the Web : a survey of current practice and prospects for the future Centre for Human Computer Interaction Design City University London Northampton Square 2 The importance of describing images on the Web'

[16] Marsh, E. E. and White, M. D. (2003) 'A taxonomy of relationships between images and text', 59(6), pp. 647–672. doi: 10.1108/00220410310506303

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404

[65] POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

[9] Salisbury, E., Kamar, E. and Morris, M. R. (2017) 'Conversational Crowdsourcing as a Tool', Aaai Hcomp 17, (Hcomp), pp. 147–156.

[17] Morris, M. R. et al. (2018) 'Rich representations of visual content for Screen reader users', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3173633.

[60] Bigham, J. P. et al. (2010) 'VizWiz : Nearly Real-time Answers to Visual Questions'

[72]Gurari, D. et al. (2018) 'VizWiz Grand Challenge: Answering Visual Questions from Blind People', Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3608–3617. doi: 10.1109/CVPR.2018.00380.

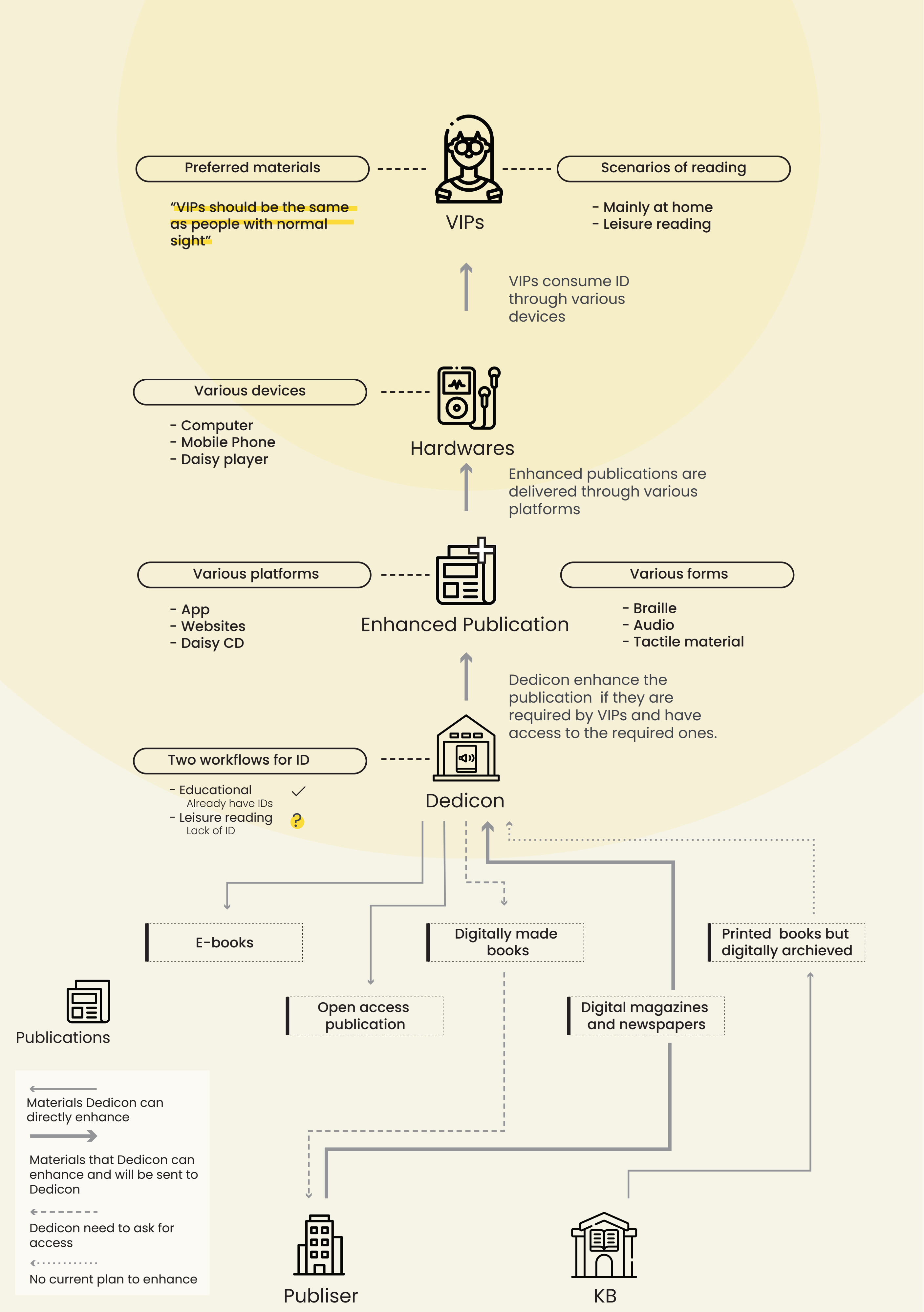
MOVING - ON

Conclusions for this chapter:

1. Due to the complexity of the image-context relationship, there are no indicators that can directly evaluate the importance of image and subjective judgment still plays an essential role.
2. Based on the current research results, it's possible to establish a preliminary mapping between the image & context factors and ID content needed to be described, which is beneficial to crowd workers' work.
3. Interactive ID reveals a new perspective for enhancing the user's experience of reading images.
4. How to utilize interactive ID to improve users' experience as well as collect data for further research to support crowd workers' work, would be a valuable direction to explore.

Questions for the next chapters

1. How does Dedicon produce IDs now?
2. What are the most urgent needs within this problem (sources, themes of reading materials)?
3. How to narrow down the scope of this project?



Preferred materials

"VIPs should be the same as people with normal sight"

VIPs

Scenarios of reading

- Mainly at home
- Leisure reading

VIPs consume ID through various devices

Various devices

- Computer
- Mobile Phone
- Daisy player

Hardwares

Enhanced publications are delivered through various platforms

Various platforms

- App
- Websites
- Daisy CD

Enhanced Publication

Various forms

- Braille
- Audio
- Tactile material

Dedicon enhance the publication if they are required by VIPs and have access to the required ones.

Two workflows for ID

- Educational ✓
  Already have IDs
- Leisure reading ?
  Lack of ID

Dedicon

E-books

Digitally made books

Printed books but digitally archieved

Open access publication

Digital magazines and newspapers

Publications

Materials Dedicon can directly enhance

Materials that Dedicon can enhance and will be sent to Dedicon

Dedicon need to ask for access

No current plan to enhance

Publiser

KB

04

Context

Following the research on VIPs' requirements, the topic of this chapter is focused on the specific context of this project, such as the current status of the Dutch accessible publication industry, the job and the responsibility of my clients, their needs, and suggestions. In this chapter, the findings from the context research, which comes from interviews with experts from Dedicon, KB, and their visually impaired users, are presented to help set the scope of future research activities.

4.1What experts say

This section introduces the key findings from experts.

4.1.1 Aim of research

Learning of the current accessible publication industry and the role of image description (ID); Dedicon's business scope; suggestions for ID generation

4.1.2 Research questions

What's the context of ID regarding this project?

- the sources of the images/ target groups

How professional describers from Dedicon describe images?

- Workflow/ Content of ID/ Personalization

From Dedicon's perspective, what is the reading experience of existing readers?

- Scenario/ Deficiencies/ Suggestions

4.1.3 Research method: Online interviews

To obtain knowledge about accessible publications, staff from Dedicon and KB are invited to the interview. Expert 1 is responsible for the research department and the accessibility projects from KB. Expert 2 is a product manager from Dedicon. Her work focuses on the production of audiobooks, including ID within them. Expert 3 is from Dedicon too. She is an editor with more than 10 years experience, and her scope of work includes materials on the art and visual content. Expert 4 is responsible for Dedicon's relationship with KB.

4.1.4 Key findings

4.1.4.1 Dutch accessibility publication Industry

Access to accessible contents

If VIPs want a narrated or braille version of a novel or a thriller, he or she uses the services of Passend lezen and Dedicon. According to the copyright law, organizations like Dedicon have the right to 'enhance' and make content accessible, but they can only distribute the enhanced formats to users that meet certain 'impairment criteria'.

Existing public library resources may not provide enough accessibility support. For example, the services around the KB digital archives like Delpher are currently not accessible enough because of poor OCR quality

Various sources and formats

Different publications may come in different formats. Based on the sources and formats of publications, the ways and possibilities that they are translated into accessible versions are also different, either for now and for the future.

E-books

- Format: epub3

- Access: cannot be simply modified because of copyright law, but Dedicon has the right to do so.

- Challenge: Publishers currently need assistance to learn how to make accessible publications or adapt their workflows to outsource that part

Digital magazines and newspapers

- Format: mostly only available in app

- Access: Dedicon will receive the archival version of the publishers (in NewsML / NITF), but may not be a complete version

- Challenge: KB currently has no idea about how publishers will enhance them. KB is currently only storing them.

Books that are made with digital technology but only available in print

- Format: mostly PDF/X

- Access: Dedicon can ask for access, but publishers don't need to cooperate.

- Challenge: If VIPs need an accessible version, Dedicon sometimes needs to scan the physical book and do OCR.

Printed publication that has been digitized

- Format: DAISY(Dedicon), PDF (KB)

- Access: already archived

- Challenge: older publications (and books that are published exclusively in print) need to be converted to an accessible format. Doing this (cost-) effectively is still a challenge. Images need to be recognized, classified, and then described. But the requirements for VIPs or for search are also different.

Open-access publication

- Format: PDF

- Access: Open access

- Challenge: Researches have little knowledge of image description. But the number of requests may be quite low.

Figure 4.1.1 Different sources of publications, including their format and access to Dedicon

Various Devices

Access to publications may also be done via various 'reading systems'. Here are software and hardware applications that are used a lot by the target groups includes:

- A computer with a 'screen reader' and/or braille display
- A computer with a very large screen
- A mobile device using the assistive technology already in the OS (like Voice Over in IOS)
- Daisy Reader

Summary

The accessible publication is a complex industry. Due to the complexity of the stakeholders and the restriction of the policy, the procedure of making accessible versions can be quite different, especially in the future. Therefore, it is particularly important to choose a suitable domain for this project.

4.1.4.2 Dedicon's current products

Dedicon's job is to make information accessible for people who have reading disabilities. They transformed the text into accessible forms (braille and audio). Production of ID is also part of this work, but currently, **it is still a new technique to explore.**

"For educated materials, we do have image descriptions"

"Describe image is still in a pilot phase"

Two separate workflows

Workflow Dedicon has 2 separate workflows to produce ID for educational and leisure reading materials.

For educational materials, contents of ID are made before reading. While for leisure content, IDs are made on the spot.

For leisure reading materials, Dedicon will not read all of the content, they will pick the most important article every week and provide a bit more background information. Describers will describe the image on the spot. When there are too many images, they are sometimes skipped.

Different forms and platforms

Dedicon will normally provide two versions (braille and audio) for the processed material. In addition, for artworks, Dedicon will also make textile images with the museum if possible. The content produced by Dedicon will be provided to VIPs through different platforms, including App (Daisy reader), websites and CD.

Summary

Although Dedicon plans to provide more ID in the future, apparently it is not possible for them to manually provide ID for all content. At this stage, Dedicon's choice is to give priority to providing high-quality descriptions for educational materials, and the content for leisure reading needs a more efficient method.

4.1.4.3 Insights about users

"Our customers are like people with normal sight"

"Our customers are like people with normal sight"

This is the most important insight I gained from the interview. Although Expert4 pointed out that viewers of different genders may have different reading preferences (nonfiction and fiction). However, Expert 3, as an experienced description, emphasizes this belief. VIP's preference should be the same as the rest of the people who can see. **So, there is not a specific subsection.**

This view is also reflected in the two observations mentioned in the interview. One is that the work of Expert 3 includes making visual artworks accessible, and this work has been welcomed by many VIPs. The second one is that Dedicon has recently experimentally added descriptions to the images in fashion magazines, and Dedicon users have shown great enthusiasm for this. Although VIP may not have the concept of some visual elements (e.g., color), this does not eliminate their interest in visual content.

Scenarios: mostly reading at home

Most of Dedicon's users read at home, especially for leisure reading. (Of course, students will consume educational materials in school). Meanwhile, it was pointed out that users may also read in public transportation, vacation or other possible scenarios, and this may become more and more common in the future. In addition, although Dedicon's products can support users to read independently. In some cases, VIPs will still depend on their family members.

Summary

It is not appropriate to define a preferred subsection for VIPs. As ordinary people, they may be interested in the content from all themes and should enjoy the same right to consume various information like everyone else. Reading at home is the main scenario for VIPs to consume Dedicon's products.

4.2 What VIPs say

This section introduces the key findings from visually impaired people (VIPs)

4.2.1 Aim of research

Further understanding of VIPs reading habits and scenes; and their process of consuming ID

4.2.2 Research Questions

What are VIPs' usual reading habits and scenarios?

- Main reading platform/ equipment/ preferred materials/ theme/ location

Their experience with ID

- How important and how accessible

4.2.3 Research Method: Phone calls

In order to get more insights from the view of VIPs, I try to in get contact with VIPs in real life through Dedicon and KB's channels. This research includes two parts:

1. the first is interviews with participants about their reading habits;

2. 3 materials (image and context) with different levels of complexity to simulate the real reading scenario, and then collect users' feedback. (The results from this part is integrated into the last chapter)

Unfortunately, due to the COVID-19 pandemic situation and the language barrier (most participants prefer to speak in Dutch), not enough subjects choose to participate in my test. In the end, I only got responses from 2 subjects. (P1 is congenital Low-vision and P2 is congenital blindness, both of them are among 20-40 years old) Because the number of subjects is so small, I choose to only take the insights as a supplement to experts' interviews.

" I'm reading like magazines in braille or just for hearing."

4.2.4 Key findings

4.2.4.1 Reading habits and scenarios

Both of the participants spend less than 7 hours on E-book every week. Instead, they use news websites/ applications much more frequently (every day). Both of them use social media a lot, including Facebook and Twitter. P2 also has a LinkedIn to find a job. In addition, P1 mentioned she will use both braille and audio for reading (switch between them), which is consistent with the expert interview results.

" I read the social media area Facebook, and also Twitter."

Regarding the degree of accessibility of images, participants gave relatively **low scores** to both of the 2 media. (Avg.=2 for news websites/ apps; Avg.=1.5 for E-books, Total: 0~4)The lack of Image description is also consistent with the views of the literature. In contrast, regarding the importance of picture description, the scores for 2 media **are quite different**. (Avg=4 for news websites/app; Avg =2 for E-books, Total: 0~4)

4.2.4.2 Limitation of choices

Although both participants have used social media, their choices are very limited because of the level of accessibility of the content, because "*Normally there are no platforms that tell me about images*"(P1). Both of them will use Facebook, because it provides some simple picture descriptions to help them imagine images. On the contrary, although they are interested in Instagram, most of its content is not accessible, thus restricting their use.

The limited accessible content/platforms to VIPs may affect their daily usage habits. Stangl's research also found that users' low levels of engagement may be related to their familiarity with the medium, which stemmed from inadequate descriptions of images [2]. Customers' enthusiasm for Dedicon's fashion magazine with ID description reflects this as well. Therefore, it may not be rigorous enough to conclude their preferences from the existing reading habits of VIPs, because their preferences are shaped by limited access.

4.2.5 Difficulties in recruiting participants

In the process of field research, another important finding was the difficulty of recruiting subjects. It didn't just happen to me as a graduated student. For organizations like Dedicon, it is also difficult to conduct large-scale tests. It is mainly caused by the following reasons:

1. **The process of recruiting participants.** To protect customers' privacy, the recruitment of participants can only be carried out through the newsletter, which increases the time cost and difficulty of communication.

2. **Uncertainty of the test results.** P2 mentioned that for her, what she cares most about the test is whether she can receive follow-up updates. She hopes that her time can contribute to the development of new projects, but sometimes it's not the case. Expert 2 also mentioned that not every study can give participants follow-up updates, which affects their motivation to participate in the test.

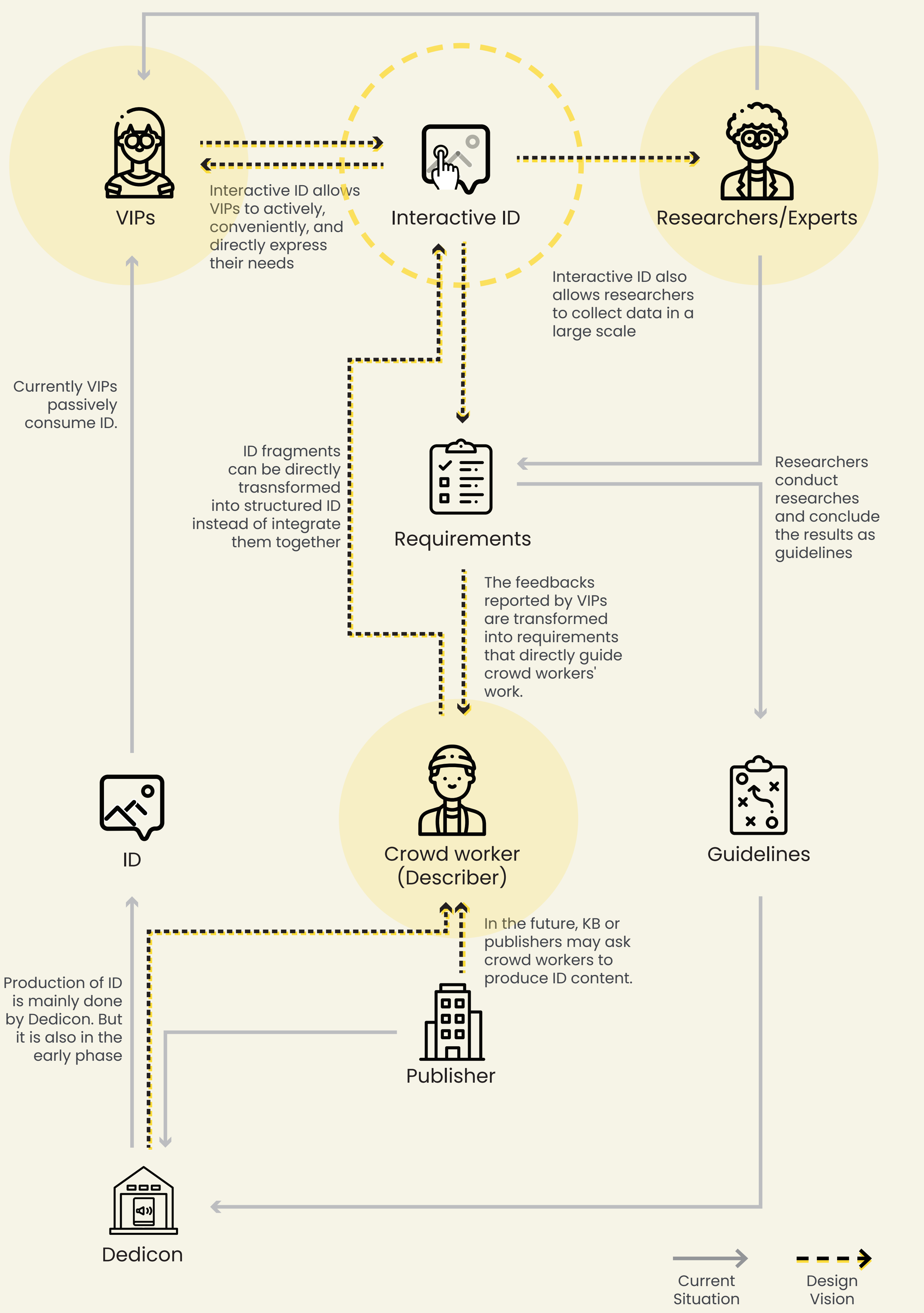
3. Language. Language is an additional obstacle, especially for me as a non-Dutch speaker. Although Dutch VIPs have good English skills, according to P1's feedback, they are still more willing to communicate in Dutch.

The difficulty of recruiting participants makes it inappropriate to continue to study the needs of VIPs directly as a research goal. As mentioned in the previous chapter, learning how to collect data may be an equally valuable issue in the long run.

MOVING -ON

Conclusions for this chapter:

1. The raw materials of Dedicon' products may have various sources and formats, and they are processed into various forms and distributed on various platforms as well.
2. Among the reading materials that Dedicon is mainly responsible for, there is already a relatively reliable solution to make ID for images in educational materials. In contrast, leisure reading materials still need additional ways to help produce ID. Meanwhile, leisure reading at home is one of the main reading scenarios for users, so it can be taken as the research focus of this project.
3. Because of the difficulty of recruiting participants, the focus of this project no longer focuses on what are the needs of VIPs, but how to support VIPs to express their needs and how to collect their needs.



VIPs

Interactive ID

Researchers/Experts

Interactive ID allows VIPs to actively, conveniently, and directly express their needs

Interactive ID also allows researchers to collect data in a large scale

Currently VIPs passively consume ID.

ID fragments can be directly trasnsformed into structured ID instead of integrate them together

Requirements

Researchers conduct researches and conclude the results as guidelines

The feedbacks reported by VIPs are transformed into requirements that directly guide crowd workers' work.

ID

Crowd worker (Describer)

Guidelines

In the future, KB or publishers may ask crowd workers to produce ID content.

Production of ID is mainly done by Dedicon. But it is also in the early phase

Publisher

Dedicon

Current Situation

Design Vision

Design Goal

In this chapter, findings from the previous chapters are synthesized and transformed into Design Brief. Design Brief clarifies the direction of future research and design activities. In addition, the knowledge of how to describe images (i.e., the contents from Chapter 3) will also serve as the basis for subsequent research and be integrated into the prototypes.

5.1Design Scope

Based on the results from the context and literature research, the design scope is clarified here to set the focus of future research activities.

5.1.1 Target group

All visually impaired people (VIPs). Dedicon's target users are all people officially regarded as visually impaired. Although the number of groups in VIP is not the same (Chapter 2.1.3), their dependence on image description (ID) may also be different. (e.g., Mild low-vision may prefer to use their remaining sights.) Due to the lack of sufficient evidence to support the selection of a specific group, all VIPs are included as the target group.

5.1.2 Source and theme

News and magazine for leisure reading. Dedicon has different workflows (4.1.4.2) for educational materials and leisure reading materials. They have already been able to provide complete ID for educational materials. Now they need a solution to provide descriptions for leisure reading materials more efficiently.

5.1.3 Form: Audio

Data shows that VIPs using Braille are on the decline, and **Audio has become an increasingly popular way of reading** [31]. Some views argue that Braille is irreplaceable, and VIPs have different requirements for reading through Braille and voice (Braille is for accuracy, Audio is for speed). But Audio is a more dominant way. In addition, due to the restrictions of the epidemic, there is no opportunity for me to approach Braille reading equipment.

5.1.4 Hardware: mobile phone

Reading hardware is set to mobile phones. Since the reading scenarios of VIPs are mainly at home or on public transportation, mobile phones fit the theme of leisure reading more. In contrast, brand-new special hardware is difficult to be popular.

Alternatives like Daisy reader are not available to me. After all, the theme of this thesis mainly focuses on the content of ID, and the content can always be easily adapted to other devices.

5.1.5 Image classification: Excluding complex and simple images

Complex and simple image types, as is shown in Figure 5.1.1, are temporarily excluded from the research scope of this project. For simple image types like text and signatures, as is discussed in 3.2.3.4, AI captioning system has been able to meet the needs of users. For complex image types like flowcharts, as is stated in 2.4.3.3. and 3.4.4, there are already templates that can guide crowd workers to describe them and it may be more appropriate to describe them in other than narrative ways.

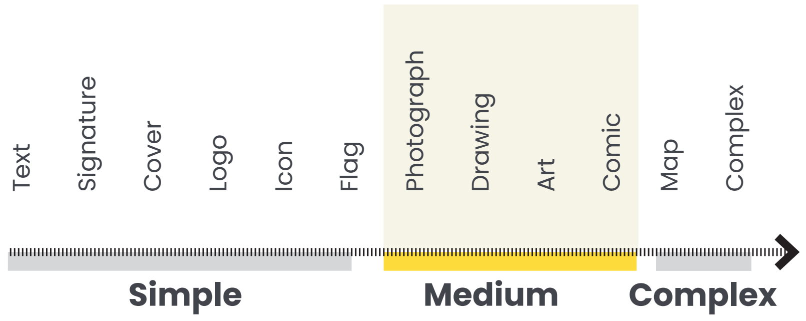


Figure 5.1.1 Images of medium complexity (i.e Drawing, Art, Comic, Photos) is the focus of this study

5.1.6 Out of scope

Design of the crowdsourcing workflow

- How to allocate resources
- How to improve their motivation

Connections between different stakeholders

- How can KB or publishers employ crowd workers to produce IDs
- How Dedicon cooperates with publishers
- How VIPs get publication resources from Dedicon or publishers

5.2 Problem Statement

This section presents the main problems derived from the previous chapters. Problem definition serves as a synthesis of key findings of the major problems that may occur for any stakeholder within the context.

5.2.1 For VIPs

5.2.1.1 Lack of ID

For VIPs, the biggest problem they face is still the lack of ID. Because of the lack of ID, they cannot effectively obtain complete access to the information in part of the publications, although they may hold a high interest.

5.2.1.2 Lack of autonomy

The existing one-shot, static ID is out of VIPs' control. Due to the lack of options for deciding whether and how to describe the ID, it may:

1. interrupt the flow of VIP reading articles;
2. cause cognitive fatigue if the ID is too long;
3. Be not able to support users to further obtain the desired details

5.2.1.3 The Demand of personalized descriptions

In different situations, users may have different needs for ID, which are affected by the following factors:

1. Context factors: context, source of the image
2. Image factors: the type of image, the content of the image (especially its subject)
3. Audience factors: The user's familiarity with the content, the user's interest in the content, the user's VI type, the duration of VI, and even the user's available time may affect their needs for ID content.

However, the existing one-approach-fits-all, static ID can't meet such diverse needs well.

5.2.2 For Dedicon, KB and other publishers

An alternative efficient way to describe images

In the future, there will be an increasing demand for producing ID. However, they have limited resources to describe images (Dedicon & KB) and lack of experience (publishers), so they need a more efficient way to describe images, which achieves the best trade-off among cost and quality

In addition, Dedicon's experts have the realized personalized needs of various subgroups with VIPs. They have made some preliminary experiments on this, but more detailed research results are needed..

5.2.3 For crowd workers

Reduce workload and complexity

For crowd workers (and possibly future first-party descriptors), they need a guided procedure to help them make decisions and streamline the description procedure. Existing guidelines are not sufficient because they still require empirical knowledge to provide high-quality Description. Especially for the application of HITL approach in describing images from publications, how to guide crowd workers to describe images effectively and efficiently is the key to making HITL approaches scalable.

5.2.4 Other barriers

Due to the epidemic situation, it is difficult to recruit participants.

5.3 Design opportunities

- Interactive image description (IID)

By presenting the ID content in an interactive form, users can actively decide what and how the ID is presented (Chapter 3.4.5). It is also able to allow VIPs report their requirements on ID content actively, which sets the condition for other opportunities.

- Decompose one-shot ID into fragments that belong to different categories

Compared to let the crowd worker generate the entire image description, decomposing the description task into several seperate parts (i.e. sub-questions about certain aspects) can reduce the complexity of the description (Chapter 2.4.3.3)

- Utilize existing research results to decide potential ID content (categories)

Based on the current research results, a guiding program can help crowd workers decide what categories of information need to be describe, based on the image and context factors (Chapter 3.3.3.1)

- Reduce the efforts to organize ID content

The generated ID fragments can be directly trasnsformed into structured ID (one feature of interactive ID) instead of integrate them together.

5.4 Design Goal

Based on the review of problems and design opportunities, the design goal is proposed in the form of a statement. It serves as a helping hand for design decisions in the left part of this project.

> ***"To develop a system which enables VIPs to have control on the ID content and is able to collect VIPs' requirements that can be transformed into straightforward description tasks for crowd workers"***

5.4.1 Design Requirements

A list of the design requirements is created to supplement the design goal and guide the design details for the design activities. The design requirements will also serve as measurements for the concepts. The requirements are created according to the problem statements and design goals. For some that cannot be implemented in this project and may not be detected, they are placed in Wishes.

1. Autonomy

- Can provide VIPs with the details about the image they want, if not, can provide users with acceptable feedback
- Allow VIPs to skip the ID
- Allow VIPs to give feedback on ID easily
- Provide intuitive and smooth interaction without interrupting their flow of reading articles

2. Navigation

- Can help VIPs navigate through the information of ID
- Enable VIPs to understand the content of ID more effectively

3. Requirement collection

- Users are willing to trust the system and provide behavior data
- The collected data is beneficial for the improvement of ID and can infer VIPs' preference

4. Personalization

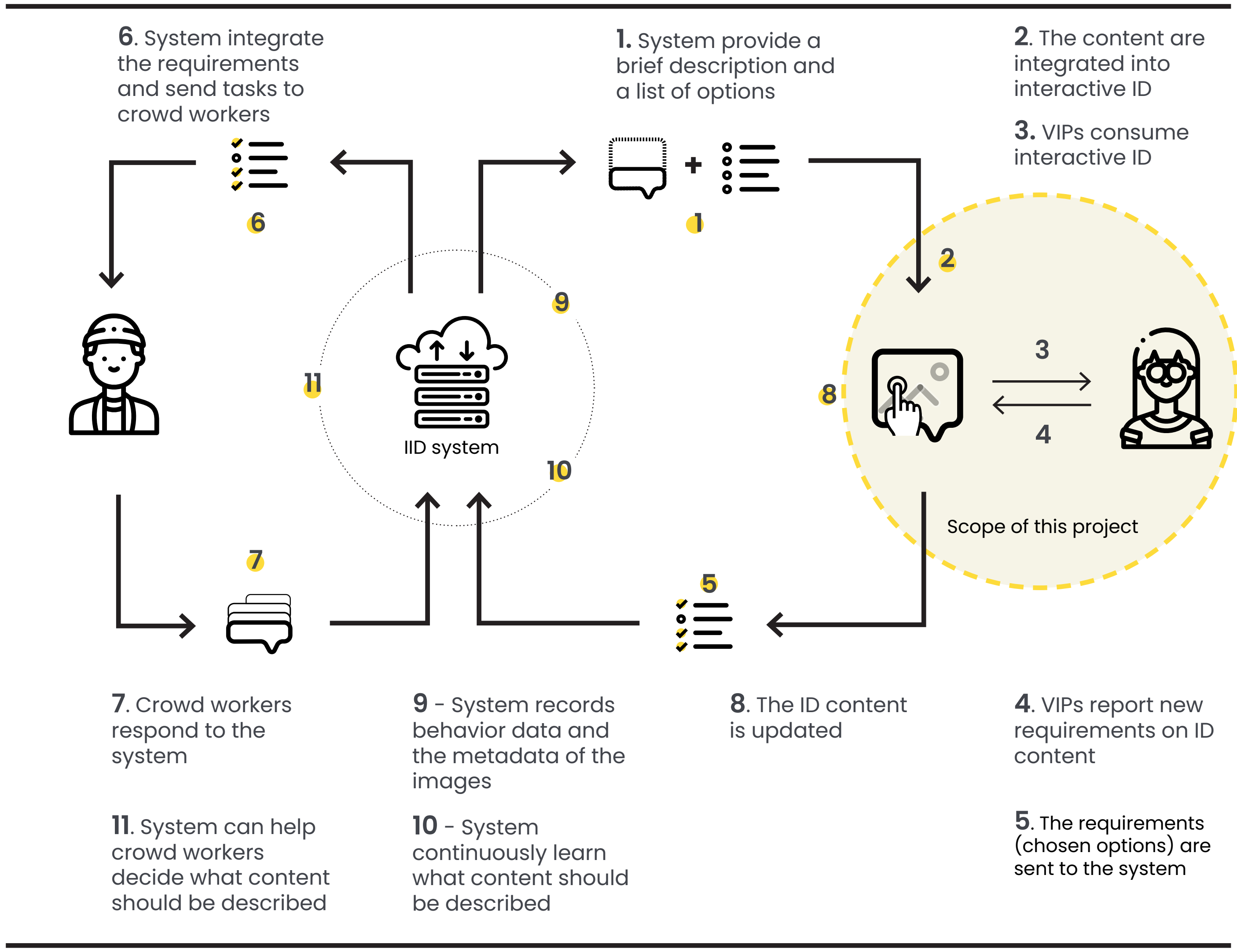
- Users can easily and actively feedback on the described results
- The user can feel that his feedback is valuable

5. Benefits for crowd workers

- The collected information can be transformed into tasks
- Can help the crowd worker describe the image more efficiently

5.4.2 Intended Situation: IID system

Figure 5.4.1 shows how the interactive image description(IID) system will eventually run, and how VIPs and crowd workers will interact with it. In this project, focus will be put on the VIPs' side.



6. System integrate the requirements and send tasks to crowd workers

1. System provide a brief description and a list of options

2. The content are integrated into interactive ID

3. VIPs consume interactive ID

7. Crowd workers respond to the system

9 - System records behavior data and the metadata of the images

8. The ID content is updated

4. VIPs report new requirements on ID content

11. System can help crowd workers decide what content should be described

10 - System continuously learn what content should be described

5. The requirements (chosen options) are sent to the system

Figure 5.4.1 An illustration illustrates how the intended interactive image system could work.

5.4.3 Relevance to the original research goal

The research goal of this project is to develop new knowledge about the requirements of VIPs. However, through field research, I realized the difficulty to recruit participants and research the requirements directly. Therefore, the design goal is proposed to solve the research question in an alternative way. If the proposed system can effectively collect the preferences of VIPs, and VIPs are willing to use it, then it may provide answers to research questions in the future.

Correspondingly, in the next research activities, the feasibility of this requirement collection system is tested. Based on that, the following iterations on improving the user experience and usability of this system are implemented.

06

Research through design

6.1 Introduction

6.1.1 Target

The design goal proposed is to develop a system that can enable VIPs to have control of the ID content and collect their requirements. Therefore, as is suggested in 3.5.1, interactive ID is required to support crowd workers to do so. However, currently, almost all IDs are static. There are no existing researches on the specific impacts of an interactive ID system within a reading context and the function of requirement collection is not verified as well. Therefore, research through design method is employed in this chapter and the target of this phase is to:

1. Verify the function of requirements collection
2. Evaluate the impact of a progressive description system with structured information
3. Acquire more knowledge about the procedure of ID perception, so that more design opportunities could be inspired.

6.1.2 Research through Design

In this phase, the method of "Research through Design(RtD)" has been adopted in the design phases. RtD stands for **the design activities that play a role in the generation of knowledge.** One of the most common design activities for research through design is the development of prototypes [40]. Prototypes serve as a role to simulate the proposed IID system. So that I can collect further knowledge about the procedure of visual content consumption and the possible influence of such a system.

[40] Stappers, P. J., & Giaccardi, E. (2017). The Encyclopedia of Human-Computer Interaction.

6.1.3 Research Question

The results of test data should answer the following research questions:

- (RQ 6.1) How effectively a description system with structured information can collect VIPs' requirements
- (RQ 6.2) What's the procedure of perceiving an image through this system?
- (RQ 6.3) What's the impact of such a description system?

6.2 Conceptualization

Rather than ideation through creative techniques (e.g., a creative session), the conceptualization of this phase is directly based on Morris's research results. The concept of this phase is a combination of *progressive ID* and *structured ID*. Firstly, the idea of a progressive description is adopted since it (is supposed to) allow users to control the length of the description. Secondly, the idea of structured information is also adopted since it (is supposed to) help VIPs acquire the information they want and facilitate the collection of VIPs' requirements on the ID content.

+
Structured ID
Progressive ID
Prototype
Interactive ID

Figure 6.2.1 The prototype concept has the feature of Structured ID and Progressive ID

To realize a combination of a progressive ID and structured ID, the potential ID content is divided into 15 categories (structured ID) and put in 3 layers (progressive ID). For structured ID, the categories are gathered from the conclusions of multiple sources discussed in section 3.3 [2,3,65,9].

Options for information of different categories:

● Celebrity	● Weather	● Color	● Text
● People's characteristics	● Lighting	● Style	● Activity
●● Salient Object	● Landmarks	● Material & Texture	● Environment
●● Overall Emotion	● Building feature		● Connect with context
Relevant image focus			
● *PEOPLE*	● *Scene*	● *Object*	● *General*

Figure 6.2.2 The 15 categories of information

Users can select certain categories to get its corresponding description. The contents of the 3 levels are:

1. Image type and name of the images' visual focus (usually what an AI caption will contain)
2. the ID content categories that are regarded as consistently required among most images (text, activities ,and environment)
3. the remaining 12 categories.

*It is worth noting that due to the difference in images, usually part of the category content is empty. The reason why all the categories are provided is that it is supposed to be beneficial for the collection of VIPs' requirements

[17] Morris, M. R. et al. (2018) 'Rich representations of visual content for Screen reader users', Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3173633.

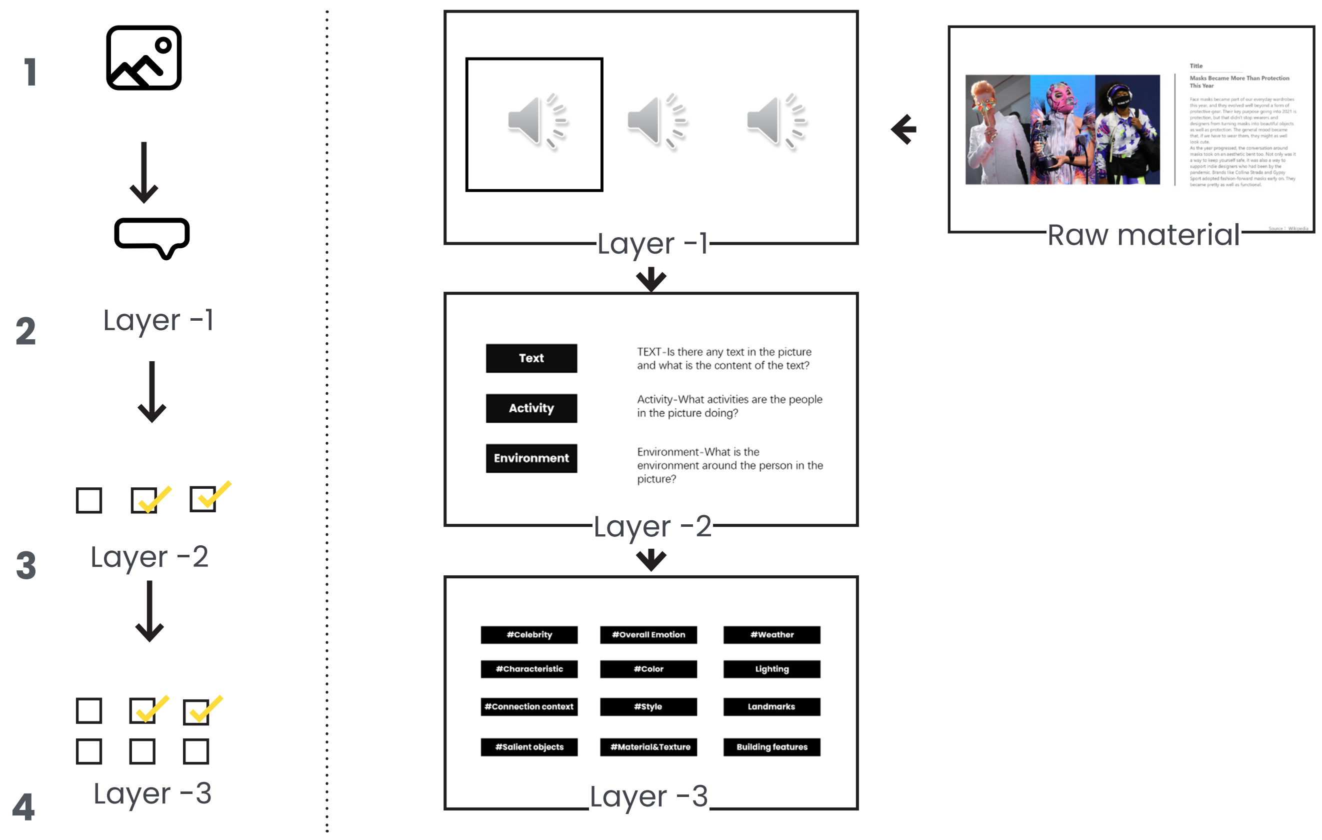
The procedure of using the prototype is:

1. Reading the surrounding text and encountering the image

2. Read the description from layer 1 (brief description)

3. View options for categories within layer 2 and choose additional descriptions

4. View options for categories within layer 3 and choose additional descriptions



TEXT-Is there any text in the picture and what is the content of the text?

Activity-What activities are the people in the picture doing?

Environment-What is the environment around the person in the picture?

Figure 6.2.3 Interfaces to simulate the IID system

The progressive ID is designed in this way to

1. Understand the dynamic process of users through interactive ID perceiving the image

2. Understand the user's satisfaction with IDs with different levels of content, and infer the probability of users requesting a detailed ID

3. Understand the impact of this system on users

6.3 Test Setting

6.3.1 Design of test procedure

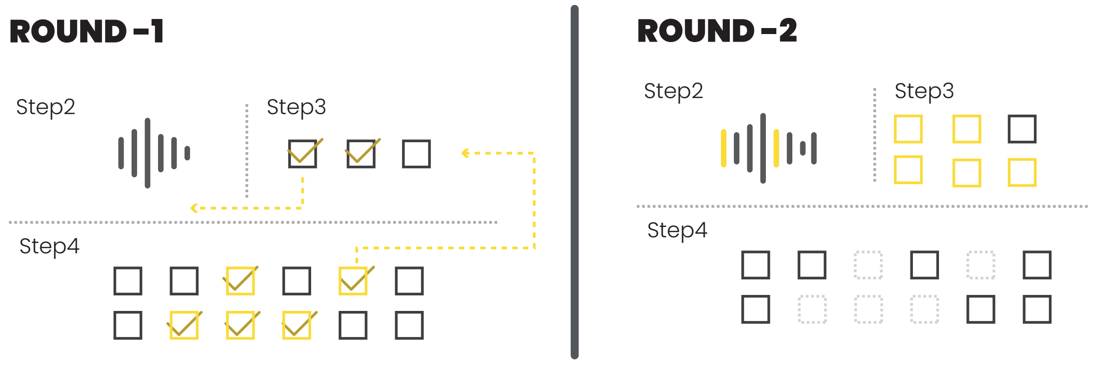


Figure 6.3.1 Optimization of Descriptions between 2 rounds

The research questions are surveyed through questionnaires(See Appendix -2 for questionnaire details) and prototypes, as is shown in the following content:

(RQ 6.1) How effectively a description system with structured information can collect VIPs' requirements

1. Two rounds of testing are set for this study, which simulates the evolvement of ID in the future system. In the first round of testing, users' feedback on ID will be collected. In the second round, the default ID and the options provided will be optimized accordingly. The participants for 2 rounds were different.
2. Participants are invited to evaluate the ID. The results from 2 rounds are compared. (Likert scale)

(RQ 6.2) What's the procedure of perceiving an image through this system?

- Participants are invited to evaluate the ID of different layers
- Intentions to know more. Why? (Yes/No)
- Satisfaction of the current ID (Likert scale)
- What do you currently want to know? Why? (Interview)
- User's interest and familiarity of the current topic (Likert scale)

(RQ 6.3) What's the impact of such a description system?

Participants are invited to evaluate the influence of the progressive content and the options for categories provided

- Is additional description helpful? (Likert scale and interview)
- Rate the coverage of the options provided? (Likert scale)
- How do these options help you think information you want? (Likert scale & Interview)

6.3.2 Articles and audios

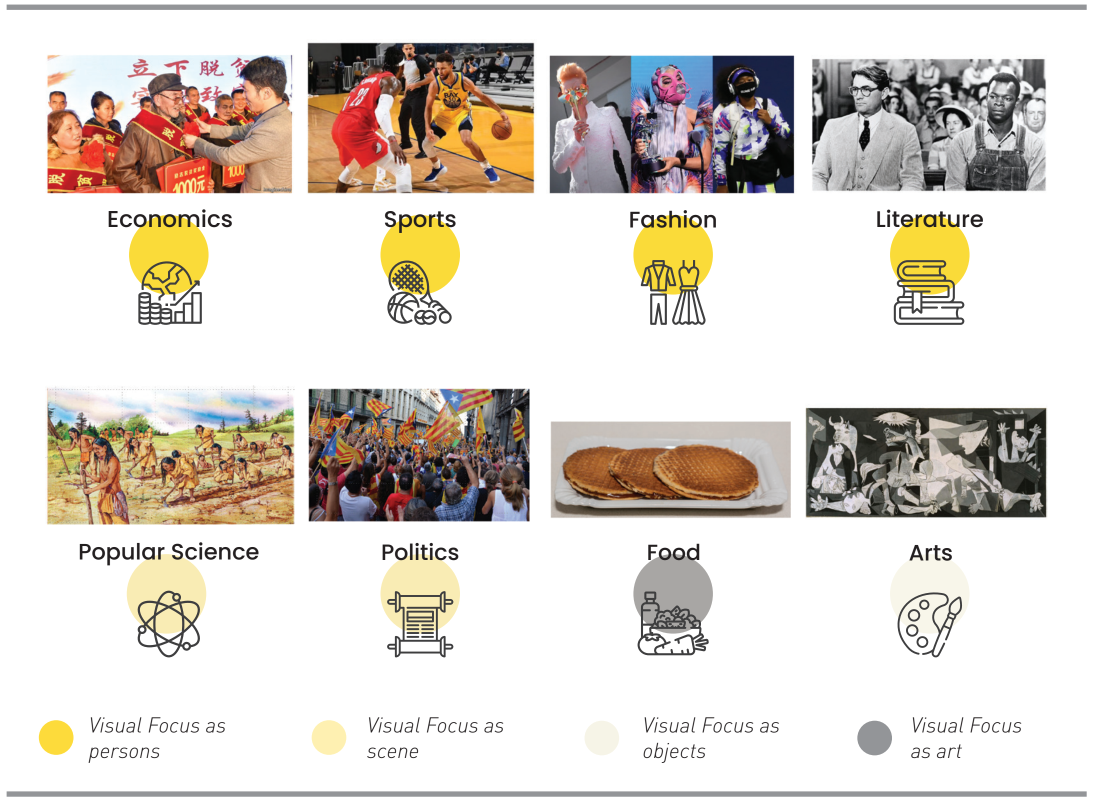


Figure 6.3.2 Overview of the tested images

To achieve enough diversity of the leisure reading materials, eight of the most popular topics in the Netherlands are selected for this test. These topics are *business, politics, arts, sports, food & recipe, fashion, literature, popular science*.

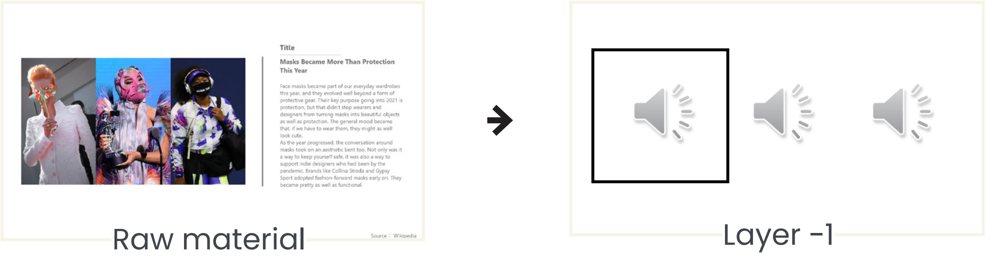


Figure 6.3.3 Raw materials are transformed into audios

Each article originally included several paragraphs of text and an image, which were converted into audio (text audio and image description) for the Participants to listen to. The default image description (layer-1, round 1) only includes the type of image and a brief description of its subject, for example: "*Informative image. This image may include multiple persons.*"

6.3.3 Test Steps

This section introduces the specific test steps

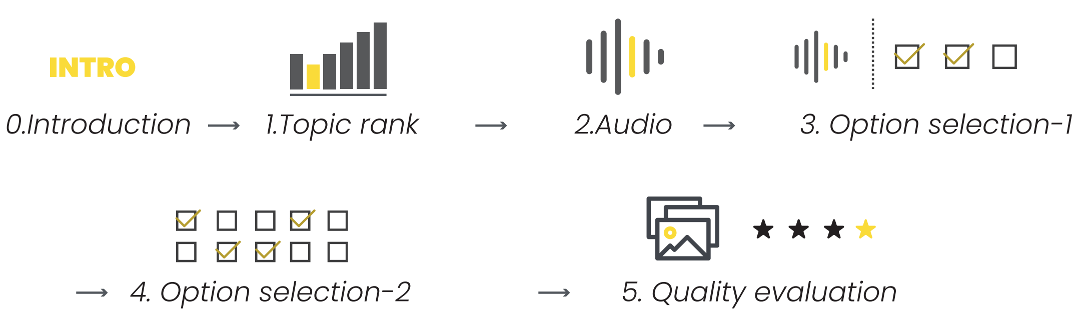


Figure 6.3.4 An overview of the test steps

0 Introduction & Consent

First, background information and procedures about the testing were introduced to the participants. Participants will be asked whether they agree to the video recording, which might be used for the analysis of the project.

1 Basic information and interest selection

Participants were asked for basic information and asked to sort the articles based on their degree of interest in the topics. The test materials in the next step will be the most interesting two articles and the least interesting two articles.

2 Listening to audios (Description layer-1)

According to the result of the previous step, the audios for the content of selected articles were played for the participants (including the title, text). Then, an audio for layer-1 description is played. After that, participants were asked to evaluate the familiarity and interest of the article topic and the sufficiency of the current description.

3 Additional description selection - 1 (Description layer-2)

Participants were displayed with several additional description options (text, activity, environment). Participants could select multiple ones of them, and additional descriptions about the selected attribute would be provided accordingly. Participants were then asked to evaluate the value of the additional description, as the evaluation of a progressive description

4 Additional description selection - 1 (Description layer-3)

In this step, more options were shown to the user, and the user needs to select the ones thought as relevant or important for understanding the picture. After that, participants were asked to evaluate the influence of these options.

5 Evaluation of sufficiency and accuracy

Since the participants were simulated as visually impaired, it provides the opportunity for participants to evaluate the description quality after actually viewing the picture. In this step, participants were shown the described picture and asked to evaluate the adequacy and accuracy of the picture description based on this. The results of this part can be a good reference to learn the relationship between the satisfaction of image description and the actual sufficiency and accuracy of image description.

6.3.4 Participants

Nine participants were recruited for this study (5 for the first round and 4 for the second round). In order to reduce the possible impact of fatigue on the test results, different users are arranged to read articles of interest and disinterest in different orders. Details are shown in the list below.

Number	Age	Education	First topic	Second Topic	Third topic	Fourth topic
				Round 1		
1	22	Master	Art (I)	Fashion(I)	Politics(U)	Economics(U)
2	23	Master	Sports(U)	Politics(U)	Fashion (I)	Art(I)
3	26	Master	Art (I)	Economics(U)	Sports(I)	Politics(U)
4	50	College	Economics(U)	Popular Science(I)	Politics(U)	Food & Recipe (I)
5	25	Master	Economics(I)	Popular science(U)	Politics(I)	Fashion(U)
				Round 2		
6	24	Master	Food & Recipe (I)	Economics(U)	Popular Science(I)	Politics(U)
7	25	Master	Economics(U)	Art(I)	Popular Science(I)	Politics(U)
8	49	College	Food & Recipe (I)	Fashion(U)	Popular Science(I)	Arts(U)
9	25	Master	Economics(U)	Art(I)	Food & Recipe (I)	Sports(I)

Table 6.3.1Participants for the test

6.4 Analysis

Firstly, explorative analysis of the collected quantitative data was conducted to get some preliminary conclusions, which mainly include the comparison of average scores and correlation analysis. The processed data were visualized and the outcome provided a good overview of users' feedback at different phases, which is helpful to verify the effect of requirement collection (RQ 6.1) and provide preliminary results for RQ 6.2 and RQ 6.3.

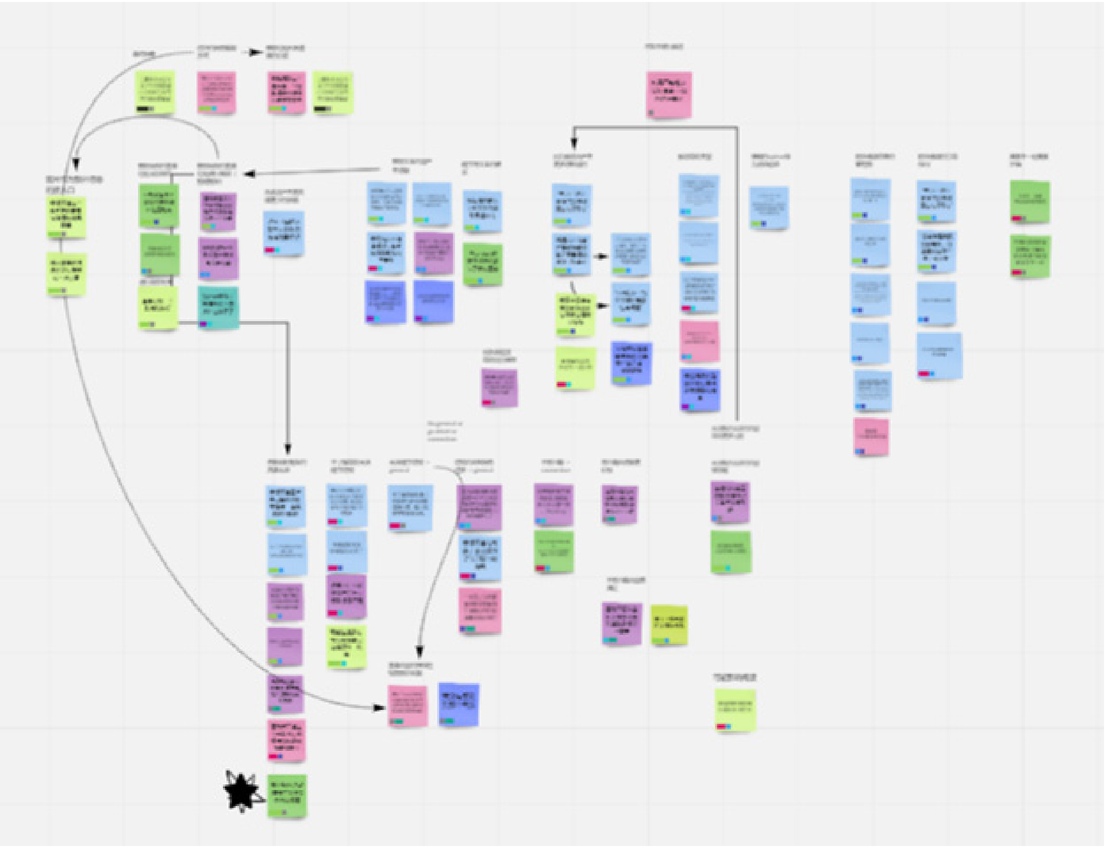


Figure 6.4.1 An screenshot of the qualititive analysis results

Secondly, to gain deeper insights into users' procedure of perception (RQ 6.2) and the influence of the system (RQ 6.3), all of the recorded video footage were reviewed for qualitative analysis. Insightful quotes and observations are recorded from transcripts and clustered to acquire universal knowledge. The results are combined with quantitative data to achieve a more reliable conclusion if possible.

6.5 Results

In this section, conclusions derived from analysis results will be presented, combined with quantitative results as supporting material. The assumptions corresponding to the research question will be proposed, accompanied by a discussion of possible design opportunities and adjustments.

6.5.1 Quantitive results

This section provides the results from the quantities results. The results of this section prove that the simulated IID system can significantly improve the ID quality based on users' feedback. And it also reveals other positive impacts brought by interactive ID.

6.5.1.1 The effect of requirement collection

This section provides the results from the quantities results. The results of this section prove that the simulated IID system can significantly improve the ID quality based on users' feedback. And it also reveals other positive impacts brought by interactive ID.

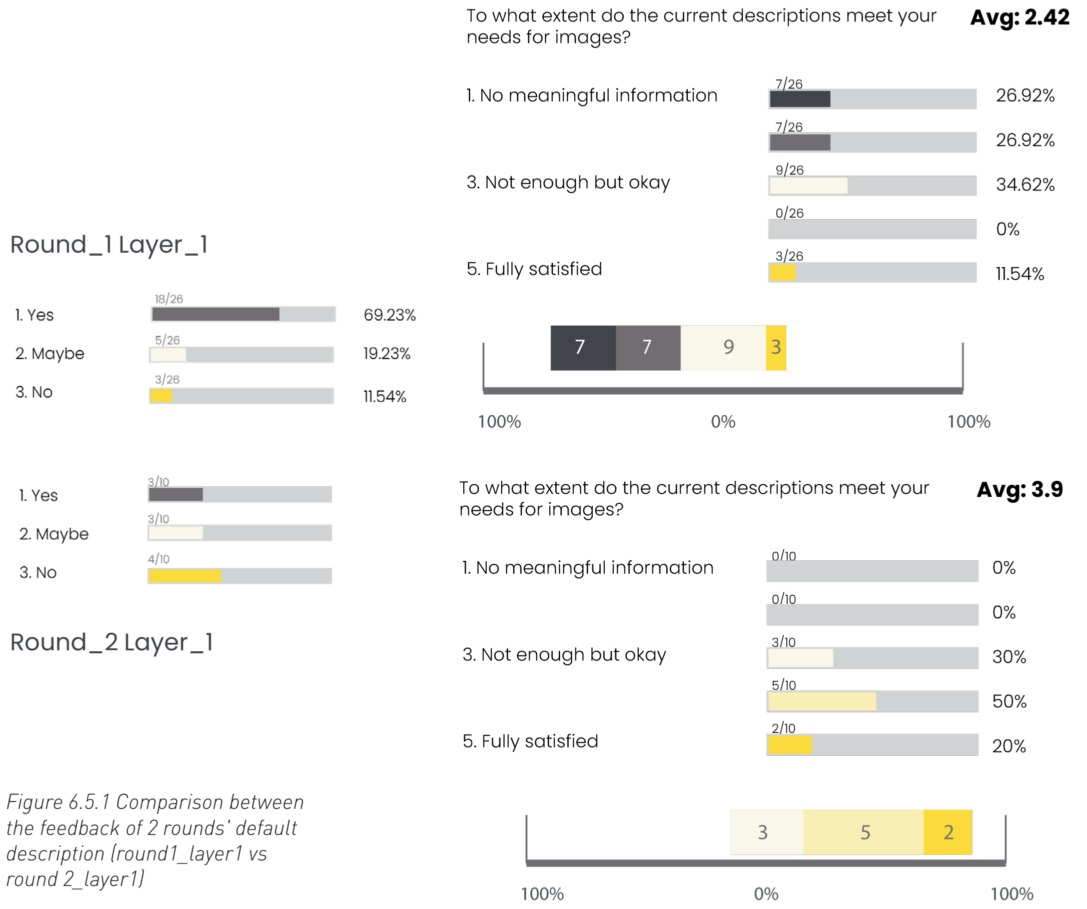
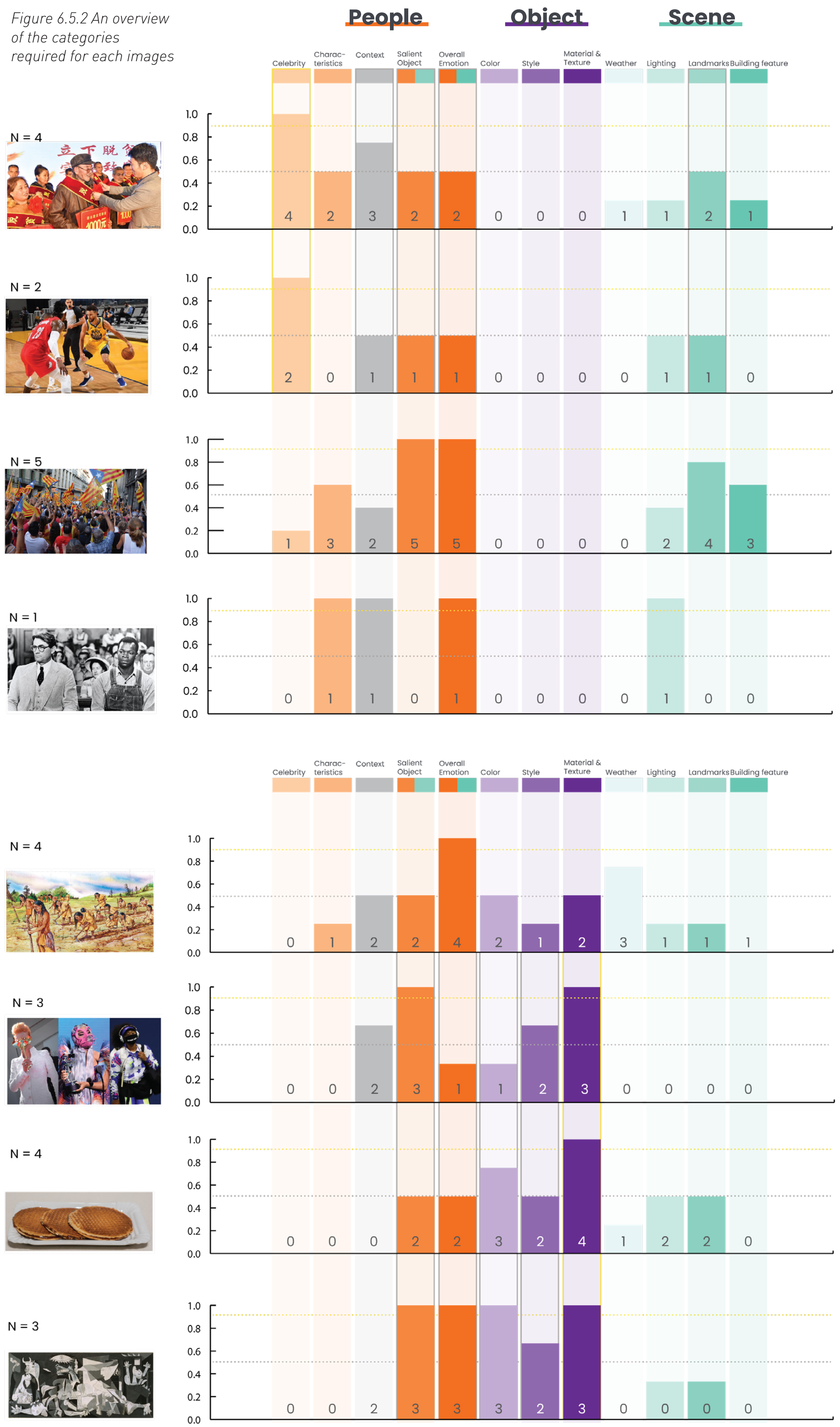


Figure 6.5.1 Comparison between the feedback of 2 rounds' default description (round1_layer1 vs round 2_layer1)

The original default description (round-1, layer-1) is consistently considered to be inadequate (Avg. 2.42). Most users (77.78% + 5.56%) with medium or high interest in the topic want to know more. In contrast, the improved default description (round-2, layer-1) achieves significantly better feedback (Avg 3.69), which proves that the content of the requirements description collected by the system can be significantly optimized.

Figure 6.5.2 An overview of the categories required for each images



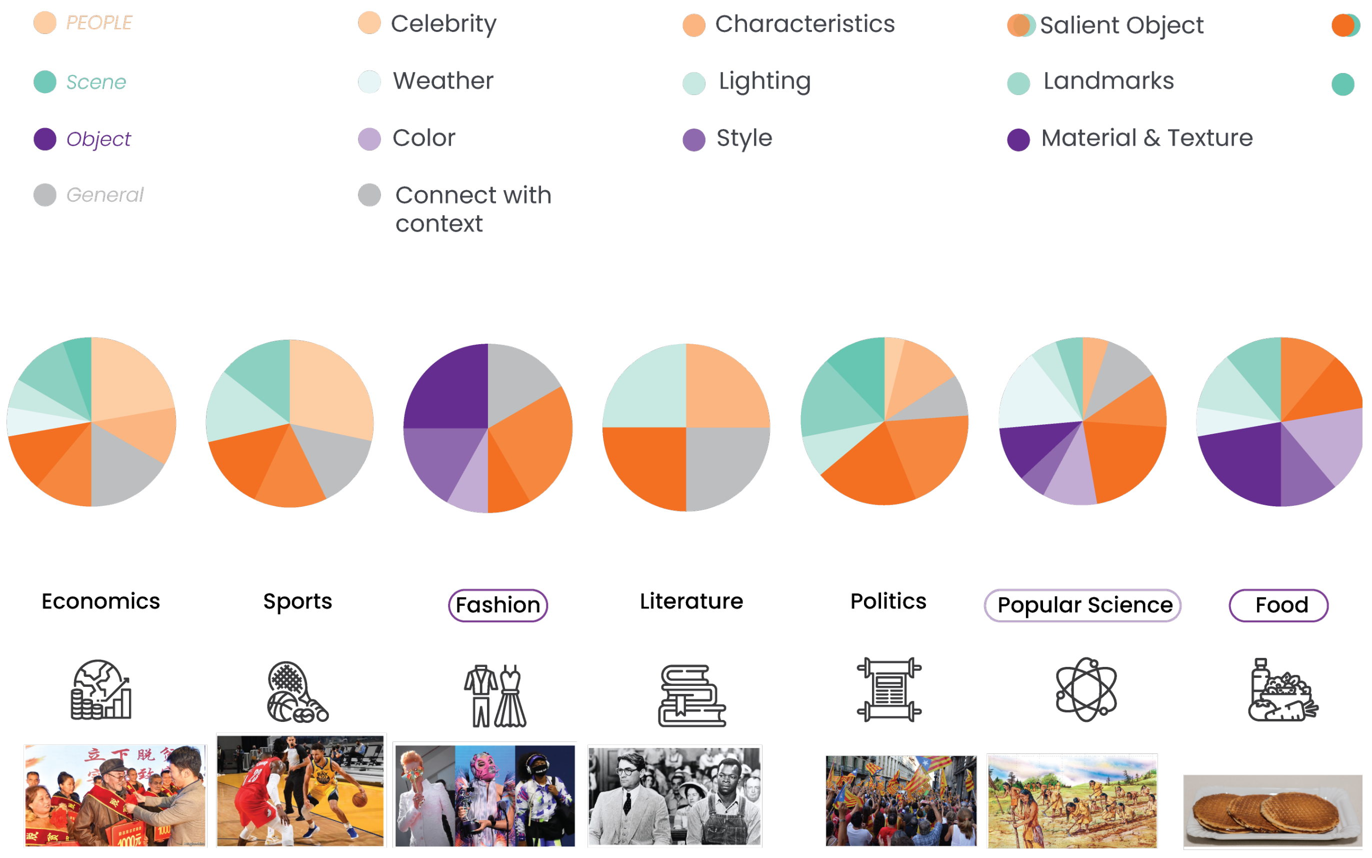


Figure 6.5.3 An overview of the categories required for each images, piechart

In addition, as is shown in Figure 6.5.2 & Figure 6.5.3, the collection results of the requirements also reveal a mapping between the image visual focus type and the content should be described, which is consistent with the research results of Stangl et. al [2].

The two exceptions are image *#science* (image of a scene but requirements also include contents belonged to objects) and image *#fashion* (visual focus is 3 persons but contents required are for objects). According to qualitative research results, the exception image *#science* is because of its image type (drawing). And image *#fashion* is different because its **contextual focus type** is "object", which is different from its visual focus. These findings, especially the influence of contextual focus, supplement existing research results.
It is found that Stangl's taxonomy is not comprehensive enough as well. First of all, for describing artworks, a set of new taxonomy is required to represent the unique content and attributes of works of art (such as dominant shape, strokes).

Secondly, the content that needs to be described between different visual focus is not distinct, especially for the categories of "scene".

[2] Stangl, A., Morris, M. R. and Gurari, D. (2020) "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions', pp. 1–13. doi: 10.1145/3313831.3376404.

6.5.1.2 The impact of progressive ID

First of all, compared to the original default description (round-1, layer-1) the progressive description (round-1, layer-2) significantly receives users' better feedback.

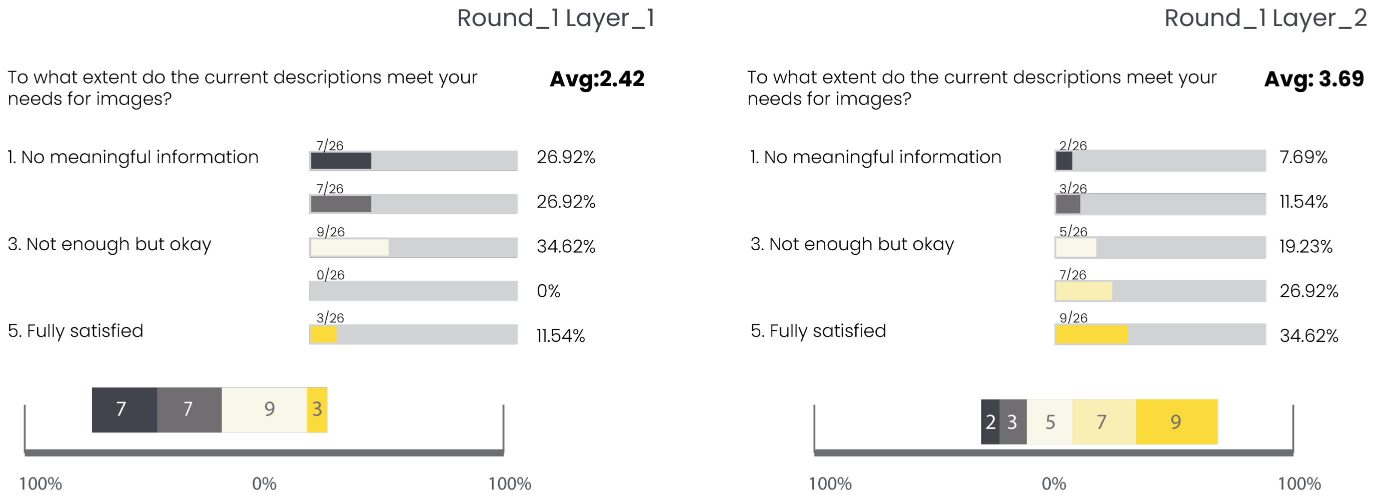


Figure 6.5.4 Comparison between the feedback of 2 layers' description (round1_layer1 vs round 1_layer2)

Wilcoxon analysis results of paired samples

p= 0.000 <0.01

And this is not only due to accessible information. Figure 6.5.5 shows a comparison of the original progressive description (round -1, layer-2) and the improved default description (round-2, layer-1). Although the latter (round-2, layer-1) contains more information, but the former(round-1, layer-2) has more fully satisfied users. Combined with the qualitative results, it indicates the extra positive impacts of progressive ID (ie, having control).

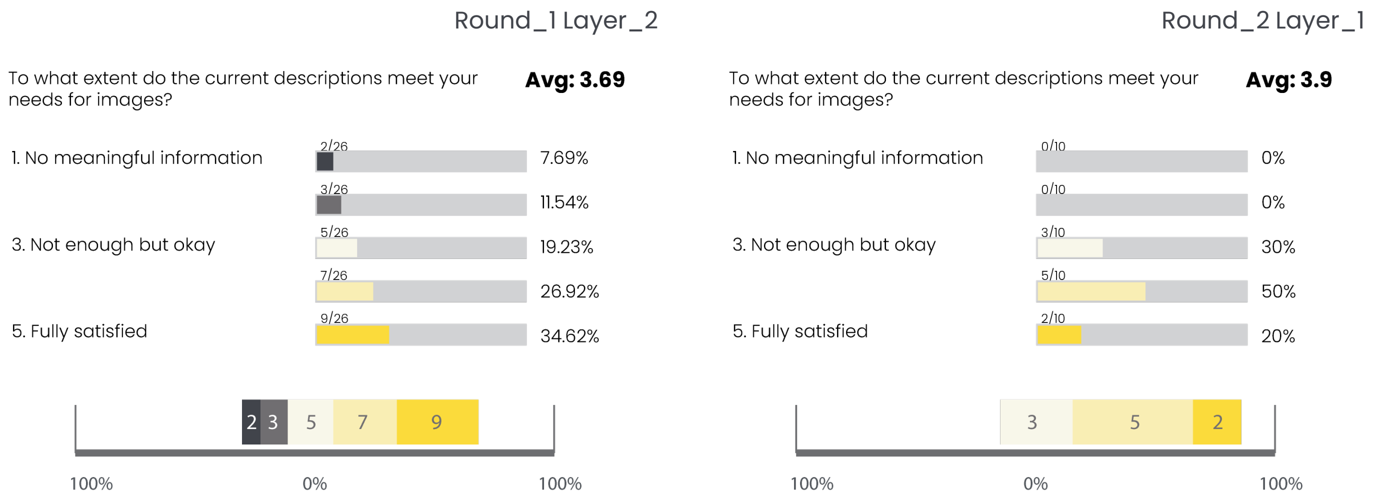


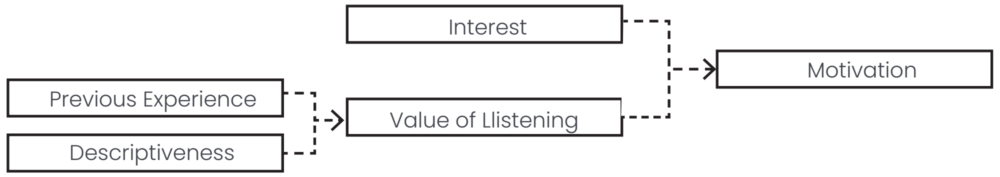
Figure 6.5.5 Comparison between the feedback of 2 descriptions (round1_layer2 vs round 2_layer1)

6.5.2 Qualitative results

6.5.2.1 The process of image perception

Through the interview results, it is found that the users' **perceiving the image description (ID) is a gradual and dynamic process, which contains a set of judgments, imagination, evaluation, and adjustment to expectations.** Image description is not necessarily a one-shot linear process, a progressive image description is more in line with users' cognitive process. We will elaborate on this point in detail below.

6.5.2.1.1 Motivation of listening ID



When a user encounters an image, a rough judgment will first be made on the value of listening to the picture description. It decides how active (the motivation) the subject will be when acquiring the image description. This initial judgment usually affects how much time and energy users are willing to invest. Except for the context, judgment may also be based on:

1.**Past experience** of the image functions under a certain topic. For articles on specific topics, users will consider images as merely a supplement to the rendering atmosphere and have no additional value beyond the information contained in the text. This judgment is usually based on their past experience of reading articles. For example, P3 mentioned that the pictures in the article on economics usually do not contain much additional information.

2. **Inference of the descriptiveness** of the image content. The user will judge whether the picture description can effectively convey the content of the picture according to the topic of the context. For example, for works of art, most subjects reported that the image description would never really replace viewing the picture.

3. **Subject's own interests.** For topics that are not of interest, subjects' expectations are generally low. In this case, they usually only pay attention to the connection between text and the image and will be easily satisfied.

Meanwhile, in most cases, users choose to learn about the basic information of the image, which is not only for curiosity but also for a judgment of the image value. As is revealed by Figure 6.5.6, even people

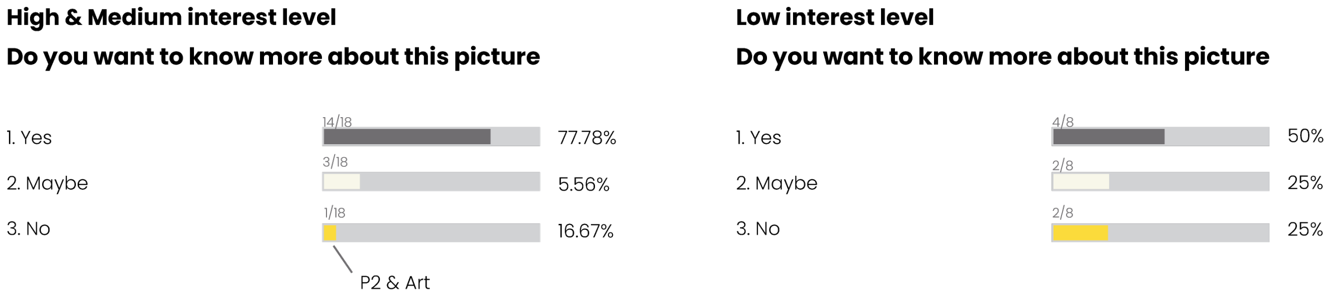
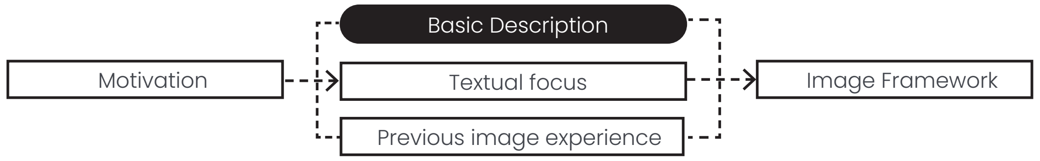


Figure 6.5.6 Participants interest to know more after listen to description (round-1, layer-1)

with low interest still want to learn more after listening to the description (round-1, layer-1).

Due to the existing one-shot description approach, normally the research focus is whether an image should be described. However, in the context of progressive ID, whether an image needs a detailed description is also a valuable question to explore.

6.5.2.1.2 Framework generation



When the user has enough motivation to perceive the image description, the user will speculate on the framework of the image, sort out the possible main content in the image and its focus, and thus determine the information they need to acquire in detail. Common frameworks include

- Separate and salient person(s) or object(s)
- Multiple people or things engaged in certain activities

Most importantly, this process usually relies on the following information combined with ID:

- **Text content.** The text content is the main source for users to imagine possible scenarios in the picture, which is also well reflected in the statistical results. For example, for the image “fashion”, everyone puts their focus on the mask based on the text content, even if the image description mentioned the main content of the image was 3 persons. For image “food”, most subjects focused on the waffle, but P5, P8 and P9 thought that the picture might contain the waffle production process because the text mentioned it.
- **Previous image experience.** Participants will use their own experience to guess what the picture depicts. For example, P3 mentioned that as a design student she was sure what image “fashion” will look like.

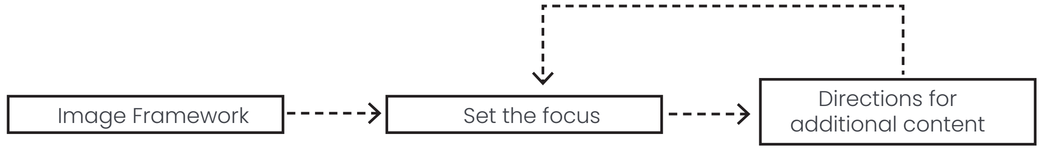
Quotes:
Pictures of salient objects or people are easier to understand completely and accurately (P2, Sports)
But specific information about masks is missing. The text content prompted me to think about information about the face masks. (P1, Fashion)
I will link the content of the article to the picture. The article mentioned ice cream so I would also imagine that there is ice cream in the picture. (P9, Food)
Similar to what I had previously guessed based on the text content.(P8, Food)
Because I am a design student, I can roughly imagine what kind of scene it is (P2 Fashion)
I can already guess what this picture looks like (P9 Art)

6.5.2.1.2.1 The effect of familiarity

When subjects are familiar with a topic, they can usually make rich associations and have clear questions. On the contrary, when the information is insufficient and the topic is unfamiliar, it will result in too much uncertainty and too much information to acquire and then subjects are unable to have imagination. For example, P1 is not familiar with the topic of politics, so even when the focus of the picture and activities were provided, she still felt that there was too much information to learn.

Quotes:
Pictures of salient objects or people are easier to understand completely and accurately (P2, Sports)
But specific information about masks is missing. The text content prompted me to think about information about the face masks. (P1, Fashion)
I will link the content of the article to the picture. The article mentioned ice cream so I would also imagine that there is ice cream in the picture. (P9, Food)
Similar to what I had previously guessed based on the text content.(P8, Food)
Because I am a design student, I can roughly imagine what kind of scene it is (P2 Fashion)
I can already guess what this picture looks like (P9 Art)

6.5.2.1.3 Go detail or go general

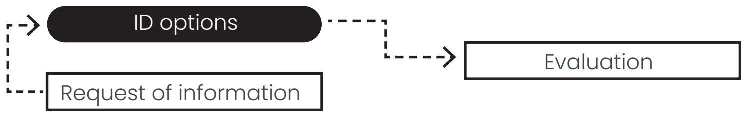


If a framework is established, users will explore the details of the image to get more information, thereby reducing uncertainty and supplementing details. The questions are always about the details a certain focus selected (e.g., the character's expression), but the focus is not necessarily the subject of the image.

When the subject is unable to build a framework or there is too much unknown information, the subject may give up the specific imagination of the visual content, and switch to more general content, especially the connection between the image and the content of the article.

Quote:
When not familiar with the topic, you would want to know the connection between the picture and the text more (P7, Politics)
I have a specific image in my mind. So, I also have some specific questions. (P1, Fashion)
I want to know too much, so I gave up on the details (P1, Politics)
I don’t think it’s possible for me to get more effective information, so it’s enough to learn some general information. (P2, Art)
Insufficient information so there is no imagination; insufficient imagination (P3, Politics)
There are many possibilities, I can't imagine what kind of picture this is (P5, economics)

6.5.2.1.4 Evaluation: Good > Perfect



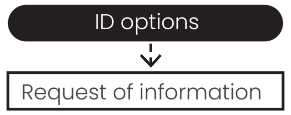
When the users have clear demands, they strongly hope that this information will be acquired and the availability of this information essentially determines the users’ final satisfaction. After getting the key information, there will be a high degree of satisfaction.

On the other hand, it is particularly difficult to ensure that users get all the information they want in only one round, because the description may also trigger additional questions for the user. Since the ID can never convey the effect of visual content 100%, when all the information is compatible with the surrounding text of the article and the core information has been conveyed, the remaining details will not affect the user’s evaluation of ID.

Quote:
When not familiar with the topic, you would want to know the connection between the picture and the text more (P7, Politics)
I have a specific image in my mind. So, I also have some specific questions. (P1, Fashion)

I want to know too much, so I gave up on the details (P1, Politics)
I don't think it's possible for me to get more effective information, so it's enough to learn some general information. (P2, Art)
Insufficient information so there is no imagination; insufficient imagination (P3, Politics)
There are many possibilities, I can't imagine what kind of picture this is (P5, economics)

6.5.2.2 The influence of the structured ID



The options provide users with a sense of control. They can actively choose the content that meets their demands instead of passively searching for the desired content in the process of receiving information.

In addition, it is reported that the **category options can help users think about what they want.** The data shows that the existing categories already have good coverage of users' required content.

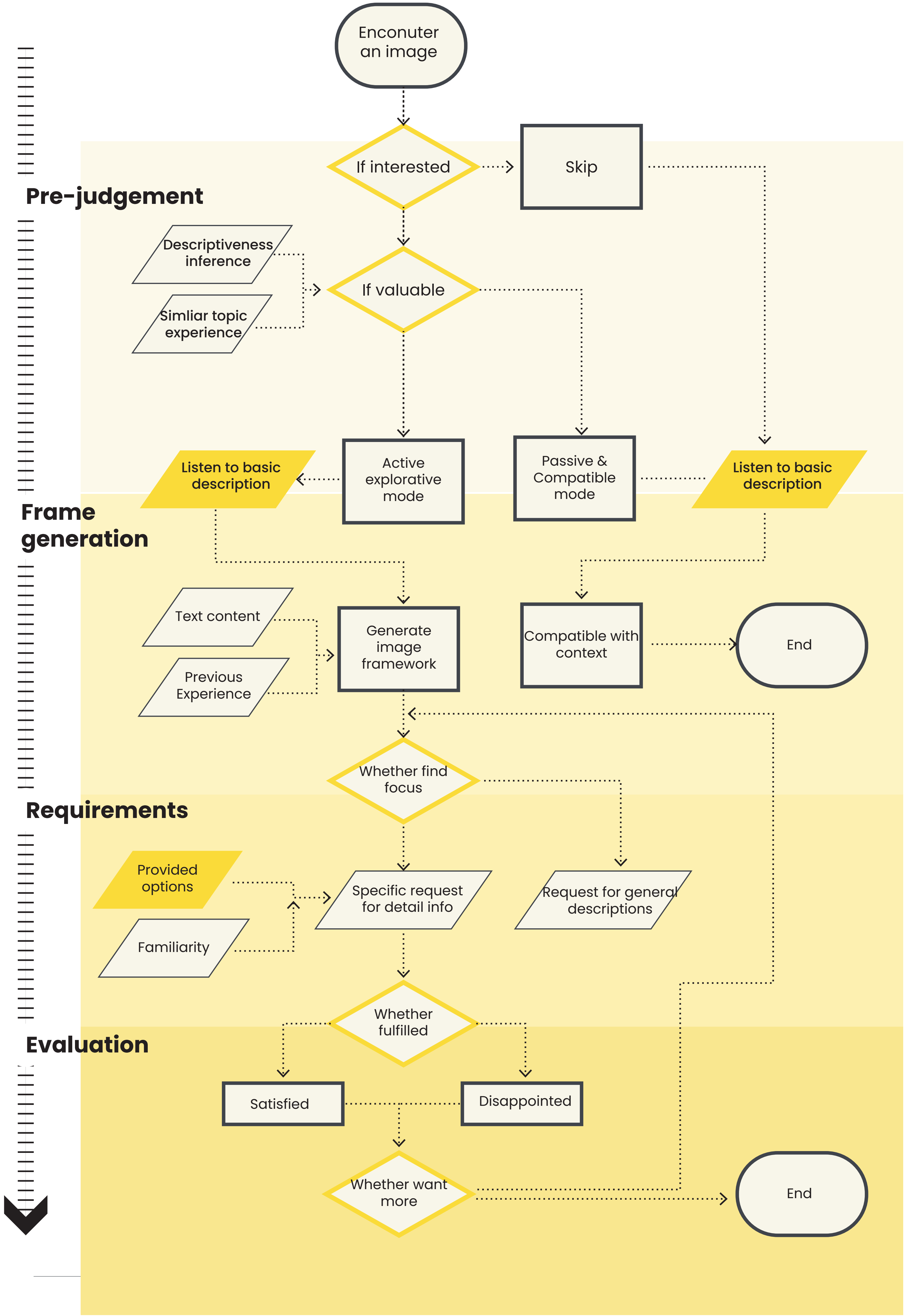
However, too many irrelevant options may increase the time for searching information and cause users to feel disappointed.

Quotes:
"Let me have a desire to learn this picture actively instead of passively listening to information (P1, Art)
It feels like ordering dishes through the menu (P3, Art)
"At first I didn't know how to ask, but these options really inspired me to sort out the structure of these pictures better" (P1, Art)
"Yes, and they stimulated my curiosity to ask more" (P9, economics)

6.5.2.3 Image perception model

In a conclusion, the qualitative analysis results are summarized as an image perception model as is shown in Figure 6.5.7.

Figure 6.5.7 Image perception Model



6.6 Transformation into design decisions

Through a lo-fi prototype, the research results in this chapter provide preliminary insights into how a progressive description system with structured works and its potential impact on the users. These insights can be transformed into design decisions to enrich the improve the original concepts:

For structured ID

The category options provided inspires and guides users to think about the content they want, which definitely gives them a sense of control and helps them establish the mental image while digesting the content of ID.

Meanwhile, due to the negative impact of too many irrelevant options, only relevant categories should be presented to the users.

In addition, according to the three tendencies for users to obtain additional descriptions (details of the subject, details of the others, and overall information), the category options can be organized better.

For progressive ID

The process by which users perceive pictures is dynamic. In the beginning, users need to build a framework to determine what additional content is needed. It helps clarifies how to divide the ID content into a brief and detailed parts.

How ID is prepared and evolved

Because the demand for a brief ID is consistent (6.5.2.1.1), the demand for details is relatively rare. Therefore, in order to achieve the best cost-efficiency, it is suggested that the brief description should be prepared in advance, while the detailed version is based on users' feedback. Considering the much longer time to generate detailed IDs, from this perspective, the distinction of brief and detailed part of interactive ID also helps describers save part of the cost of generating descriptions while maximizing the effect.

Based on this, instead of making all the content available in advance and ID evolving based on the user's behavioral data (Figure 6.6.1 proposal-1), or displaying empty options to let user actively choose what content they need (Figure 6.6.1 proposal-2), a hybrid strategy is taken: Users actively report which images need progressive description*. Then learn how to better describe the image in a specific situation based on the results of the crowd worker and the user's behavior data. Users' requests for certain information should still be allowed.

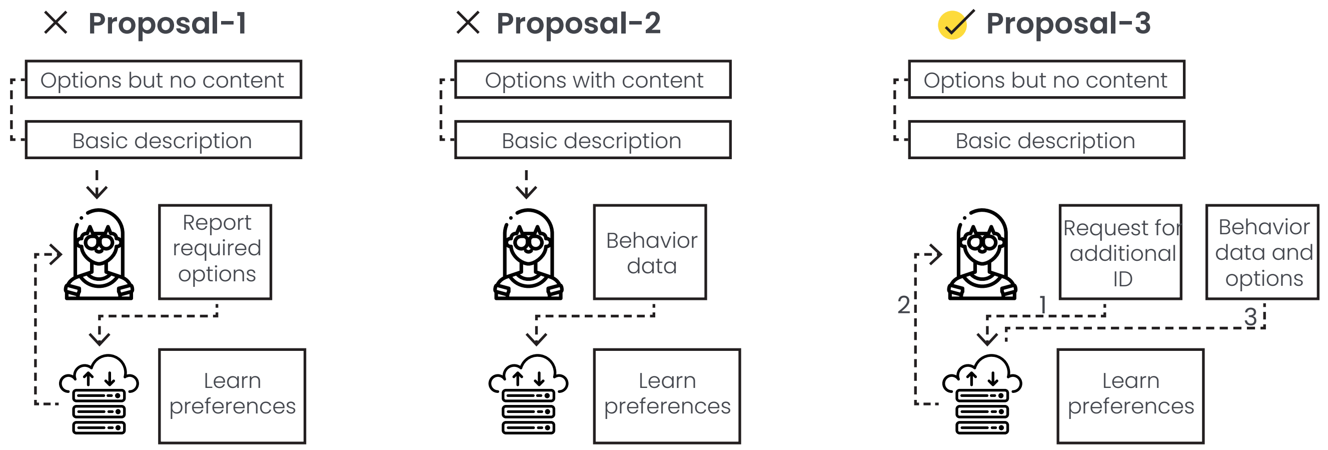


Figure 6.6.1 Three proposals and the final decision

*This procedure could be can be achieved by internal testing in a small range to mitigate the potential negative effects.

For crowd workers

The original prototype refers to the research results of Stangl et.al [2], which take image focus and source as the main reference to decide what categories of ID content should be described. Through the research activities of this phase, it is found that textual focus* and image type* should be taken into consideration. Therefore, these 2 factors should be integrated into the workflow of the crowd worker and data collection.

*see 6.5.1.1. When the textual focus is different from the visual focus of the image, users may want to know the details about contextual focus.

*When the image type is "drawing" or "artwork", users tend to pay attention to the overall feeling of the image, especially the perceptual information

MOVING -ON

Conclusions for this chapter:

1. The results of this research verify that user needs can be collected through the proposed system.
2. Progressive and structured ID can provide users with autonomy, making them feel "have control on the ID". The procedure of interacting with ID is also more in line with the dynamic process of users receiving information, generating needs, and seeking information (image perception model).
3. Textual focus and image type should also be considered to decide the ID content. In addition, for artworks a separate set of categories should be concluded.
4. Future research should also pay attention to whether the image needs a detailed description

Next chapters

1. Translate the current insights into a demonstrator of how this system will work

07

Demonstrator

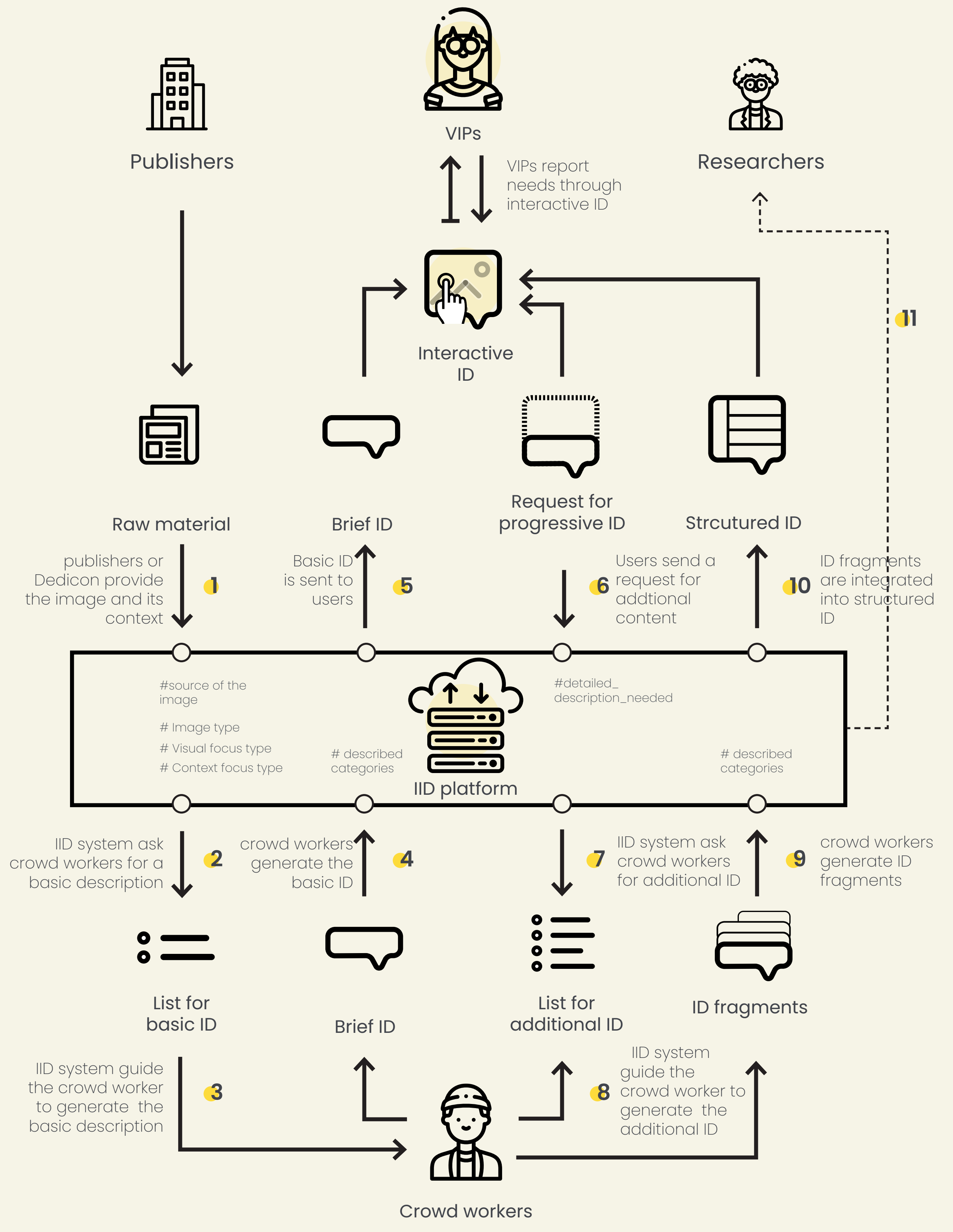

Publishers
VIPs
Researchers
VIPs report needs through interactive ID
Interactive ID
11
Raw material
Brief ID
Request for progressive ID
Strcutured ID
publishers or Dedicon provide the image and its context
1
Basic ID is sent to users
5
6
Users send a request for addtional content
10
ID fragments are integrated into structured ID
#source of the image
# Image type
# Visual focus type
# Context focus type
# described categories
#detailed_description_needed
# described categories
IID platform
IID system ask crowd workers for a basic description
2
crowd workers generate the basic ID
4
7
IID system ask crowd workers for additional ID
9
crowd workers generate ID fragments
List for basic ID
Brief ID
List for additional ID
ID fragments
IID system guide the crowd worker to generate the basic description
3
8
IID system guide the crowd worker to generate the additional ID
Crowd workers

Figure 7.1.1 Overview of the IID system

7.1 Overview of the Workflow

Figure7.1.1 shows how this system might work

First, publishers or Dedicon provide the image and its context (1). The interactive image description(IID) system asks crowd workers for a basic description (2). In this process, the system will record the #source of the image and the related topic #tags as metadata. The interfaces for crowd workers will guide them to report the #image_type, #visual_focus and #contextual_focus type, and then guide them to generate a basic description (3, 4).

Then, when the image of an article is described, it can be sent to a small group of users (5). During the reading process, these users report the images that need to be described in detail (6), and the system will thank the users after the feedback is completed. The system will mark this image as #detailed_description_needed.

When an image is confirmed to require a detailed description, it will be sent to the crowd worker again to generate a detailed description(7). The system will guide the crowd workers to describe in detail through three aspects: **1. details of the image focus, 2. details of sub-focuses (other salient objects/ persons) and 3. overall information** (8). The crowd worker responds to it and provide ID fragments (9), whcih will be integrated into structured ID.

The content is then sent to the user (10), and the user is allowed to choose the content they want through the structured ID. The system records the category that the user has selected and completely listened to. At the same time, users can feedback the specific information they need to the system through the interface.

Last but not least, according to the context factors, image factors and the ID selected by the user collected by the system. Researchers can study more systematically how to describe a picture in a specific situation, and summarize more precise guidelines to assist the work of crowd workers(11).

7.2 Interfaces for VIPs

7.2.1 Learn from Google Talkback

Since the design scope set mobile phones as the hardware to convey the image description, research on the Google talkback system is conducted to learn:

- How currently VIPs normally interact with mobile devices?
- How vibration and sound help VIPs operate on the mobile phone?

Google Talkback is an accessibility service for the Android operating system that helps blind and visually impaired users to interact with their devices. Sound, vibration and other audible feedback are used to allow the user to know the content and the activities on the screen. To be more specific, there are 3 most common gestures which allow VIPs to read information and switch among interfaces like normal people:

7.2.1.1 Swipe/ drag the finger for navigation

People with normal sights can directly find what they want on the screen and click to select. However, VIPs need to navigate through the screen to find the content they need.

The method designed for them is to drag the finger throughout the static screen. As long as the finger hovering over a component, the content/ metadata belonged to this item will be read through spoken words. When the moment the finger switching to another component, there will be vibration feedback. It is worth noting that finger's move, in this case, does not cause the page to scroll as usual.

In addition, the user can also select the previous and next space by swiping right or left.

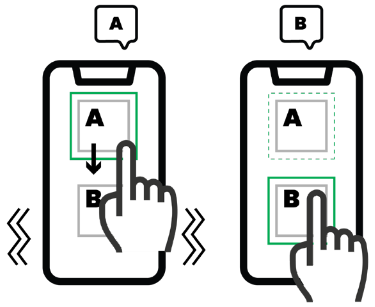


Figure 7.2.1 Illustration depicts how VIPs navigative through the screen

7.2.1.2 Two-finger swipe to scroll (scrolling sound)

So how do VIPs perform scroll operations to browse more content? The current solution is to use two fingers to swipe up and down. When the page scrolls, there is a scrolling sound effect to help users confirm their operation.

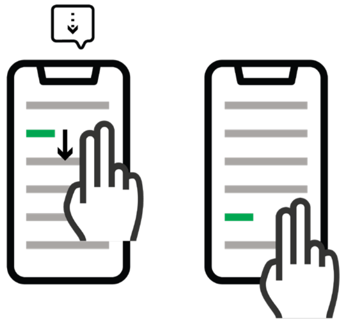


Figure 7.2.1 Illustration depicts how VIPs scroll the screen

7.2.1.3 Double-tap on everywhere of the screen to activate (vibration when activated)

People with normal sights can directly tap an item on the screen to activate it. For VIPs, it is difficult to tap a location accurately. Therefore, when Talkback users select the target item through navigation, they only need to double-click anywhere on the screen to activate the object. At this time, the phone will also confirm this operation by vibration

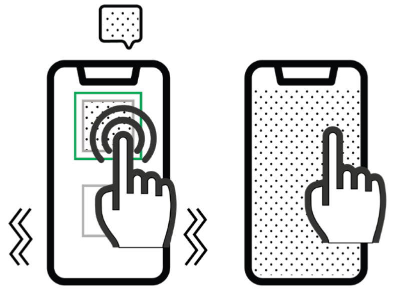


Figure 7.2.3 Illustration depicts how VIPs activate a component

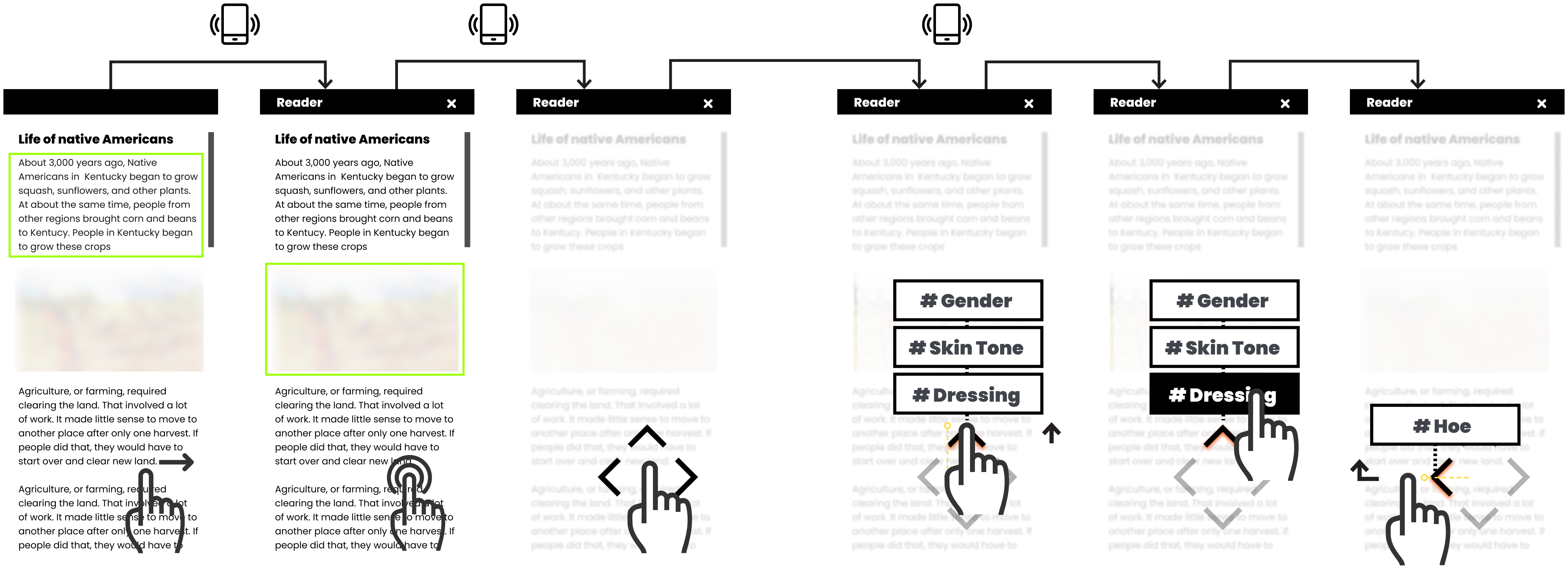
Summary

In summary, gestures, vibration, and sound play an important role in accessibility services. Different gestures and prompts correspond to different interaction logic, as is shown in Table 7.2.1

One finger swipe	Navigation
Two finger swipe	Scroll
Touch	Choose
Double tap	Activate
vibration	Switch
sound	Identify controls and reading contents

7.2.2 Design Interfaces for VIPs

Based on the learnings from Google Talkback, the design of the crowd workers interface also uses similar logic and operations



Life of native Americans

About 3,000 years ago, Native Americans in Kentucky began to grow squash, sunflowers, and other plants. At about the same time, people from other regions brought corn and beans to Kentucy. People in Kentucky began to grow these crops

Agriculture, or farming, required clearing the land. That involved a lot of work. It made little sense to move to another place after only one harvest. If people did that, they would have to start over and clear new land.

Agriculture, or farming, required clearing the land. That involved a lot of work. It made little sense to move to another place after only one harvest. If people did that, they would have to

Firstly, VIPs swipes through the screen to choose content

When an image is encountered, the system will prompt. At this time, the basic ID will be played.

Users are allowed to skip by swipe

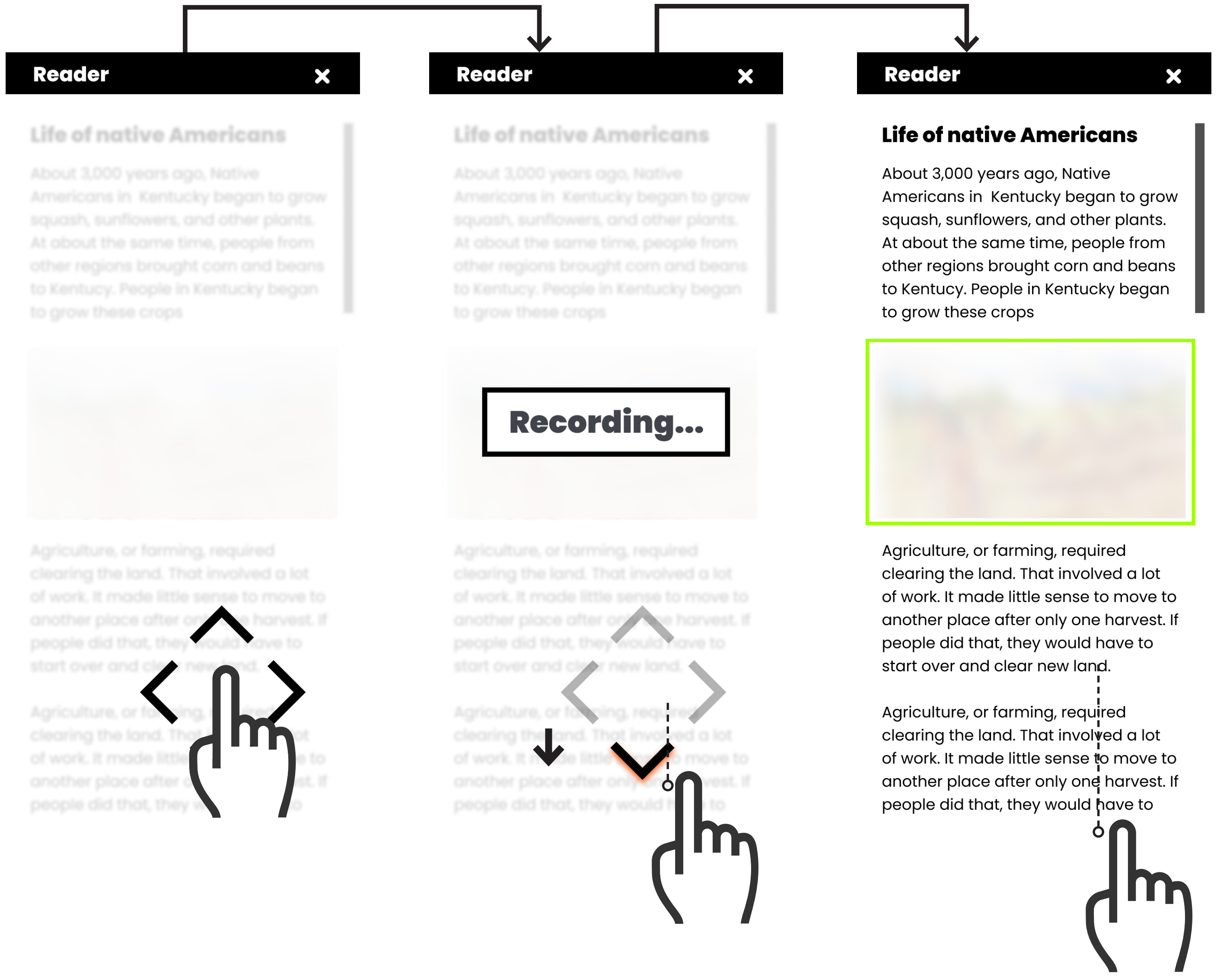
When the user is interested, he can **double-click to activate the additional structured ID** of the picture

At this time, a navigation dial will appear at the finger position. Users can get addtional ID through it

If the user moves fingers up at this time, accompanied by **a vibration**, the system will display the details for subjects options for additional ID

The user can continue to move the finger to obtain different content information

If the user swipes to the left at the beginning, then Details for sub-focuses will be triggered. Similarly, to the right is overall information



Life of native Americans

About 3,000 years ago, Native Americans in Kentucky began to grow squash, sunflowers, and other plants. At about the same time, people from other regions brought corn and beans to Kentucy. People in Kentucky began to grow these crops

Agriculture, or farming, required clearing the land. That involved a lot of work. It made little sense to move to another place after only one harvest. If people did that, they would have to start over and clear new land.

Agriculture, or farming, required clearing the land. That involved a lot of work. It made little sense to move to another place after only one harvest. If people did that, they would have to

Users can move fingers down to give feedback

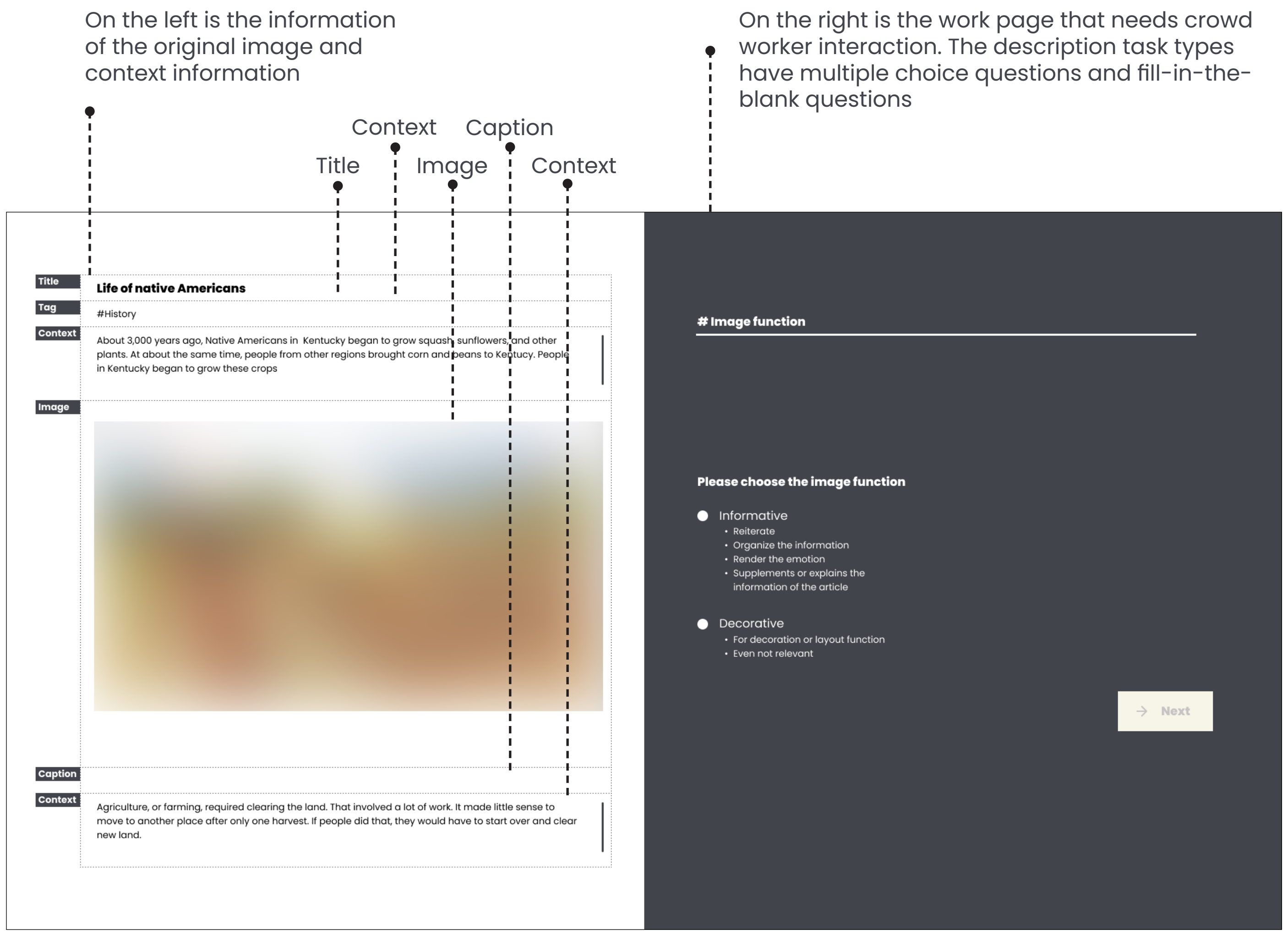
To return to the previous page, **the user only needs to release the finger, without any additional gestures or return button**

7.3 Interfaces for crowd workers

7.3.1 Basic Interface

This section introduces the basic composition of the interface for crowd workers. Basically, the left side of the interface displays information about the image, and the right side is the workspace for the crowd worker. The system will help the crowd worker determine what content needs to be described based on the attributes of the image.

See Appendix - 1 for what content(categories) the system will guide crowd workers to describe under various situations.


On the left is the information of the original image and context information

On the right is the work page that needs crowd worker interaction. The description task types have multiple choice questions and fill-in-the-blank questions

Life of native Americans

#History

About 3,000 years ago, Native Americans in Kentucky began to grow squash, sunflowers, and other plants. At about the same time, people from other regions brought corn and beans to Kentucy. People in Kentucky began to grow these crops

Agriculture, or farming, required clearing the land. That involved a lot of work. It made little sense to move to another place after only one harvest. If people did that, they would have to start over and clear new land.

# Image function

Please choose the image function

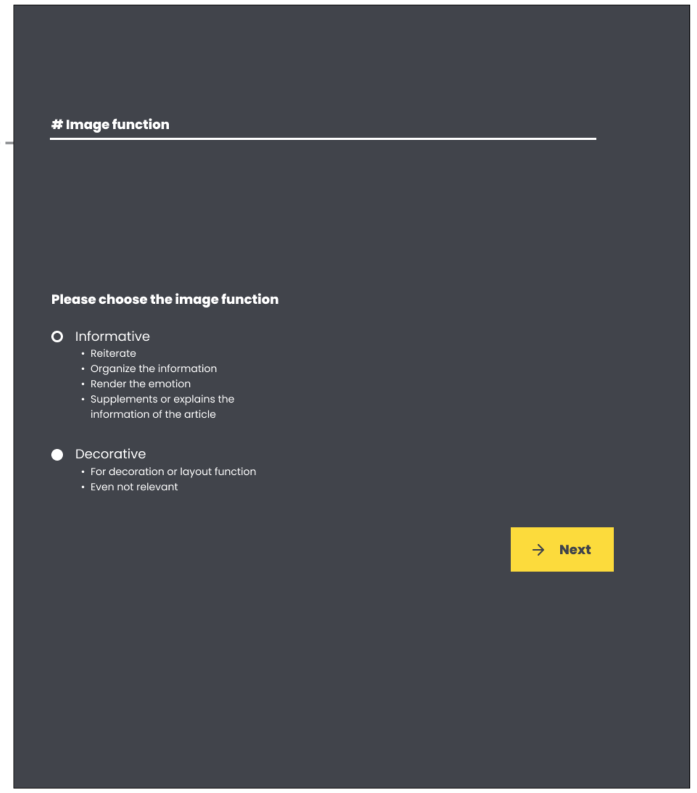
○ Informative
- Reiterate
- Organize the information
- Render the emotion
- Supplements or explains the information of the article

○ Decorative
- For decoration or layout function
- Even not relevant

7.3.2 Basic Descriptions

This section introcude the interfaces for basic description

The interface will first ask the crowd worker to determine some basic information about the image, including #image function, #image type, #visual focus and #context focus



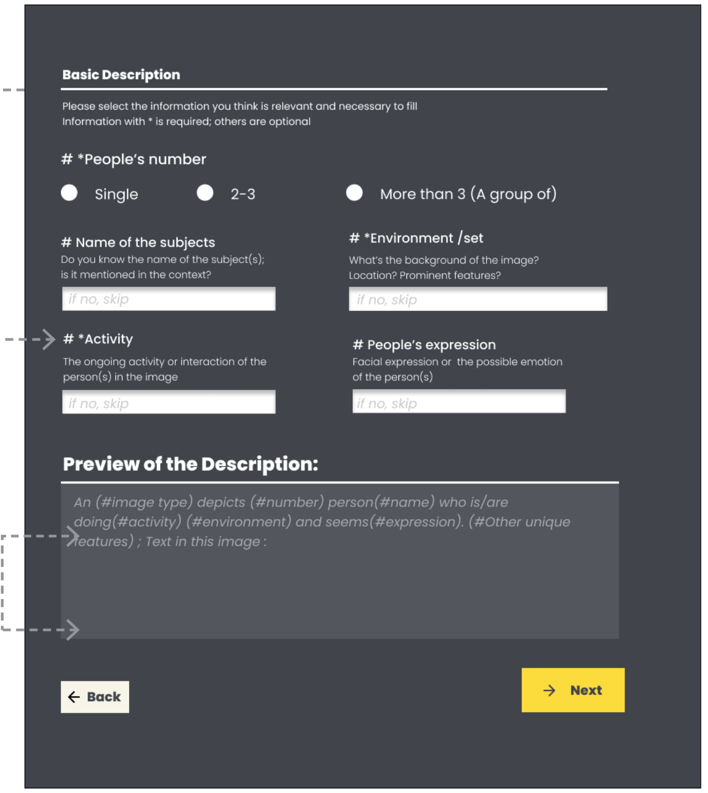
# Image function

Please choose the image function

○ Informative
- Reiterate
- Organize the information
- Render the emotion
- Supplements or explains the information of the article

● Decorative
- For decoration or layout function
- Even not relevant

Based on the user's input, the interface will generate the corresponding content that needs to be filled in by the user



Basic Description

Please select the information you think is relevant and necessary to fill
Information with * is required; others are optional

*People's number

● Single ● 2-3 ● More than 3 (A group of)

Name of the subjects
Do you know the name of the subject(s); is it mentioned in the context?
if no, skip

*Environment /set
What's the background of the image? Location? Prominent features?
if no, skip

*Activity
The ongoing activity or interaction of the person(s) in the image
if no, skip

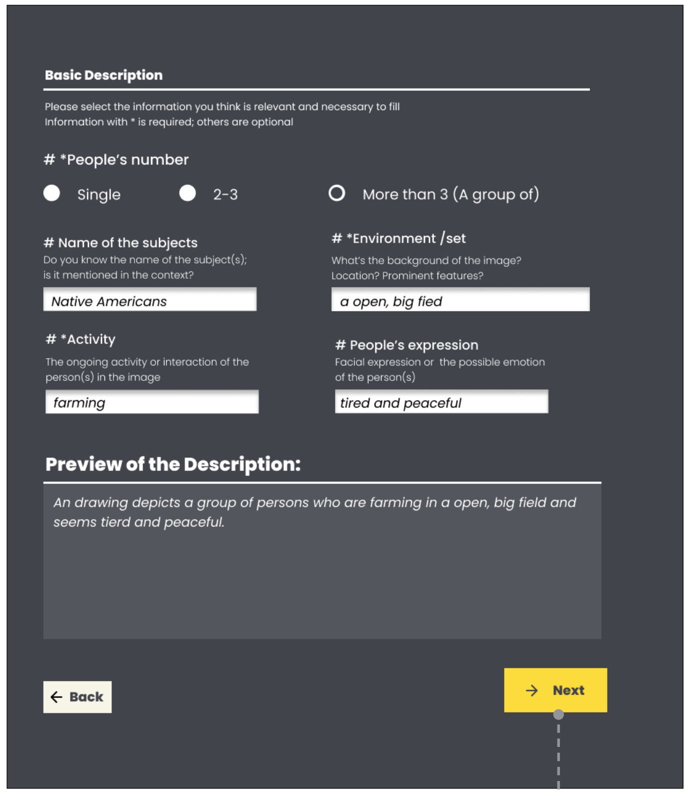
People's expression
Facial expression or the possible emotion of the person(s)
if no, skip

Preview of the Description:

An (#image type) depicts (#number) person(#name) who is/are doing(#activity) (#environment) and seems(#expression). (#Other unique features) ; Text in this image :

The system provides a range of description options, some are required, some need to be judged by crowd worker themselves

The user enters information through separate text boxes. For basic description, the system will integrate the content for the user and provide a preview



Basic Description

Please select the information you think is relevant and necessary to fill
Information with * is required; others are optional

*People's number

● Single ● 2-3 ○ More than 3 (A group of)

Name of the subjects
Do you know the name of the subject(s); is it mentioned in the context?
Native Americans

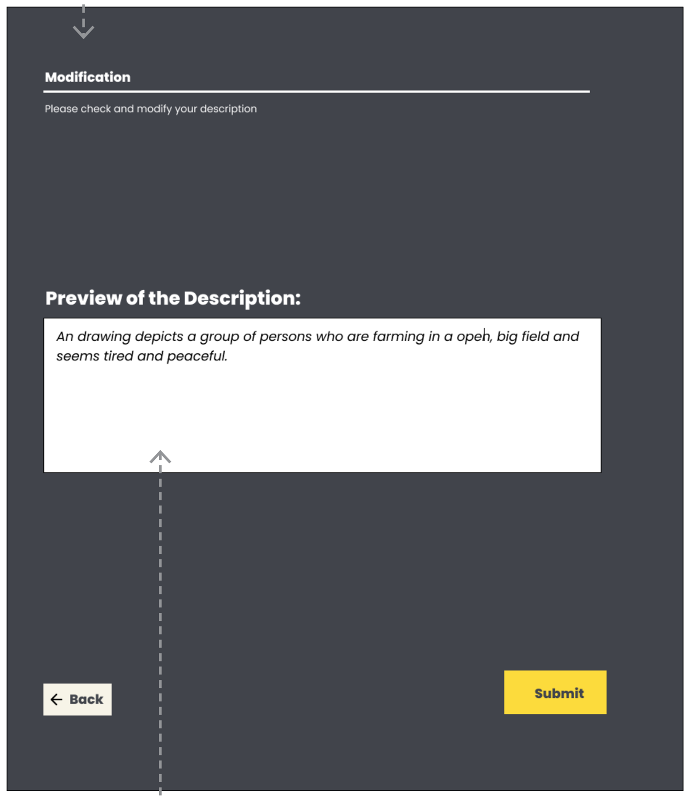
*Environment /set
What's the background of the image? Location? Prominent features?
a open, big fied

*Activity
The ongoing activity or interaction of the person(s) in the image
farming

People's expression
Facial expression or the possible emotion of the person(s)
tired and peaceful

Preview of the Description:

An drawing depicts a group of persons who are farming in a open, big field and seems tierd and peaceful.



Modification

Please check and modify your description

Preview of the Description:

An drawing depicts a group of persons who are farming in a open, big field and seems tired and peaceful.

After completing the input, the user can manually modify and then submit

7.3.3 Additional description

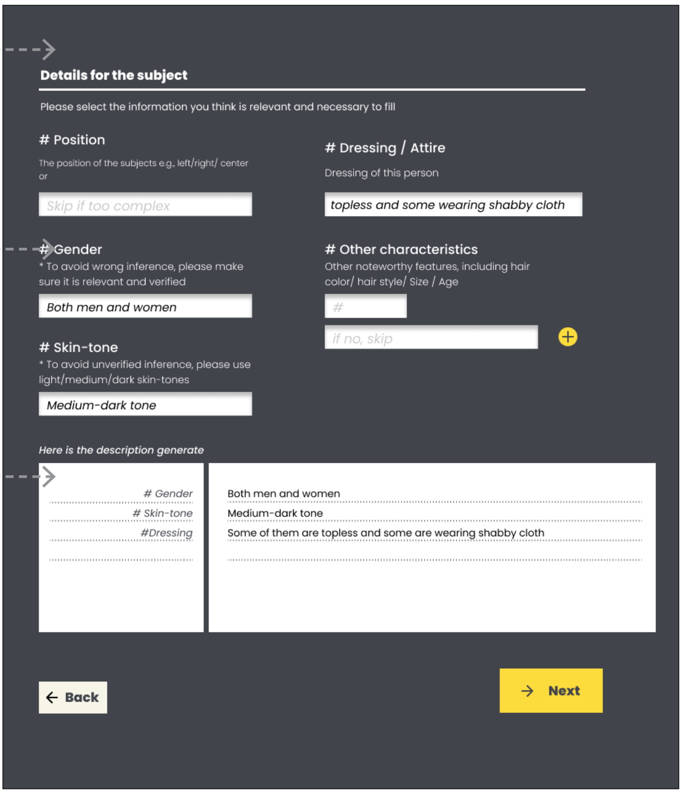
If an addtional ID is required, the system will send a new task to the crowd worker to generate the content. The generation of additional ID is divided into three parts: Details of the subjects (focus)/ Details of sub-focuses / Overall information

Details of the subject

The first part is to generate the content to describe the details of the subject

For contents needing attention in wording, the system provide extra instructions

For the content of the additional ID, they do not need to be integrated


Details for the subject

Please select the information you think is relevant and necessary to fill

Position
The position of the subjects e.g., left/right/ center or

Skip if too complex

Dressing / Attire
Dressing of this person

topless and some wearing shabby cloth

Gender
* To avoid wrong inference, please make sure it is relevant and verified

Both men and women

Other characteristics
Other noteworthy features, including hair color/ hair style/ Size / Age

#

if no, skip

Skin-tone
* To avoid unverified inference, please use light/medium/dark skin-tones

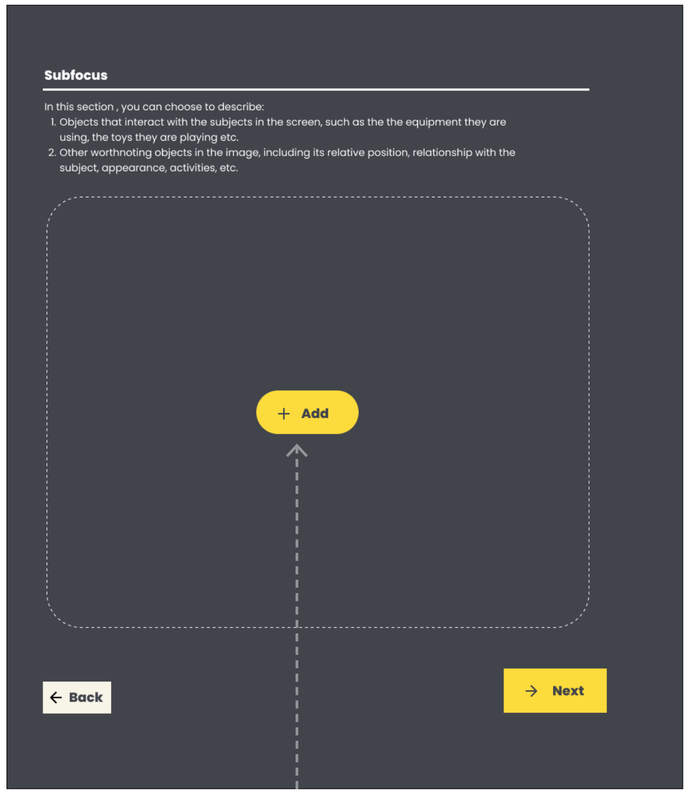
Medium-dark tone

Here is the description generate

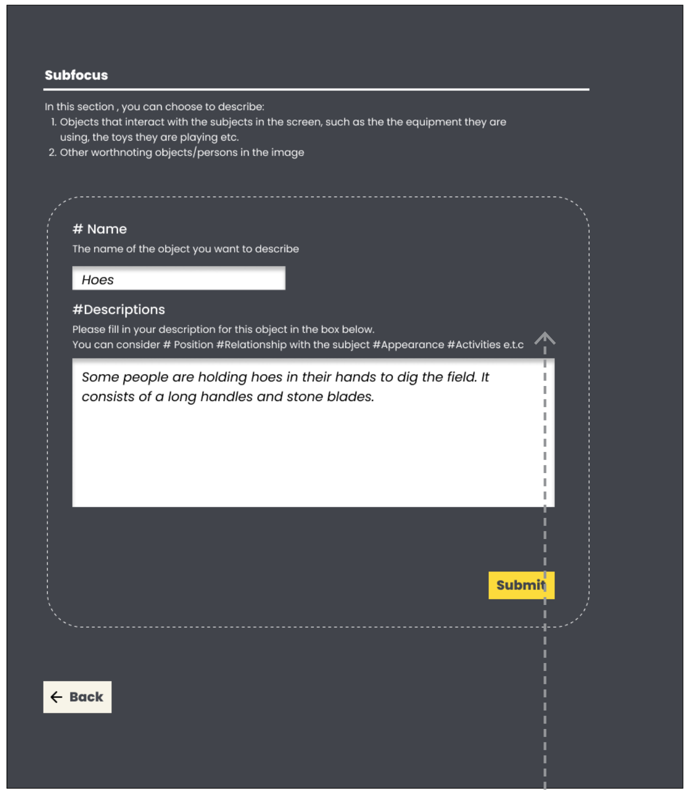
# Gender	Both men and women
# Skin-tone	Medium-dark tone
#Dressing	Some of them are topless and some are wearing shabby cloth

Details of the subfocuses

This part is a bit different from other parts. The crowd worker needs to add a sub-focus first and then describe it.


Subfocus

In this section , you can choose to describe:
1. Objects that interact with the subjects in the screen, such as the the equipment they are using, the toys they are playing etc.
2. Other worthnoting objects in the image, including its relative position, relationship with the subject, appearance, activities, etc.


Subfocus

In this section , you can choose to describe:
1. Objects that interact with the subjects in the screen, such as the the equipment they are using, the toys they are playing etc.
2. Other worthnoting objects/persons in the image

Name
The name of the object you want to describe

Hoes

#Descriptions
Please fill in your description for this object in the box below.
You can consider # Position #Relationship with the subject #Appearance #Activities e.t.c

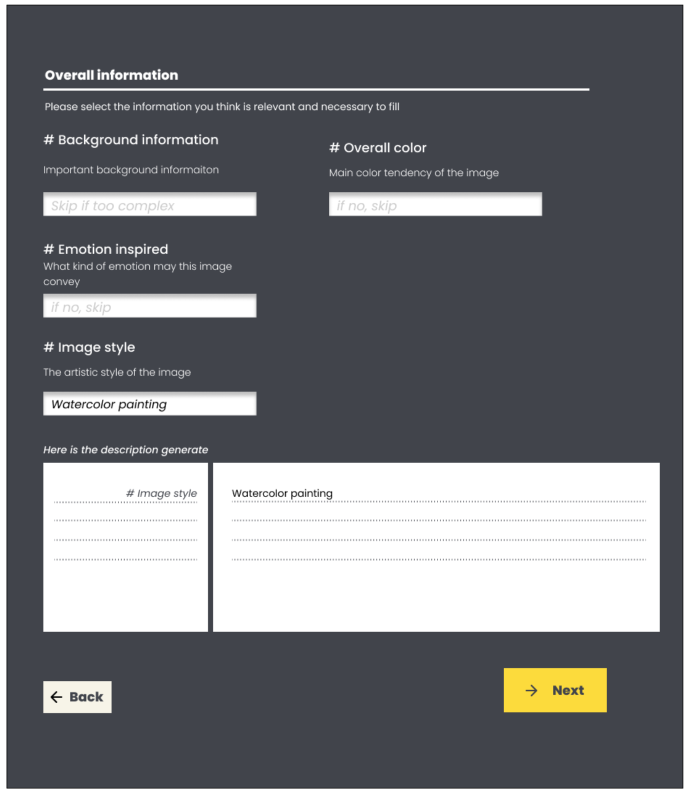
Some people are holding hoes in their hands to dig the field. It consists of a long handles and stone blades.

This part is a bit different from other parts. The crowd worker needs to add a sub-focus first and then describe it.

The system will prompt for information that may need to be described

Overall Information

The interface to fill in overall information is similar to that for details of the subject.


Overall information

Please select the information you think is relevant and necessary to fill

Background information
Important background informaiton

Skip if too complex

Overall color
Main color tendency of the image

if no, skip

Emotion inspired
What kind of emotion may this image convey

if no, skip

Image style
The artistic style of the image

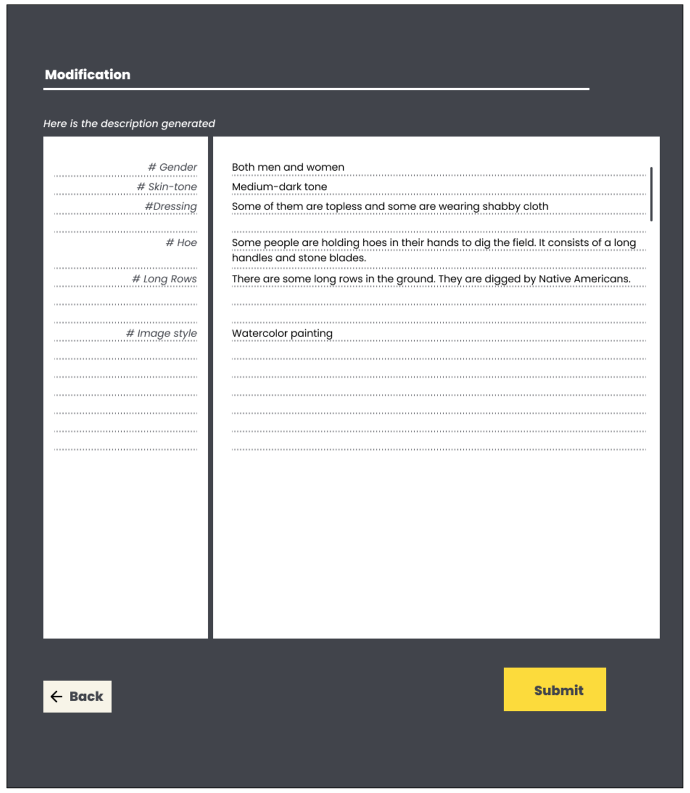
Watercolor painting

Here is the description generate

# Image style	Watercolor painting

Confirmation and thanks

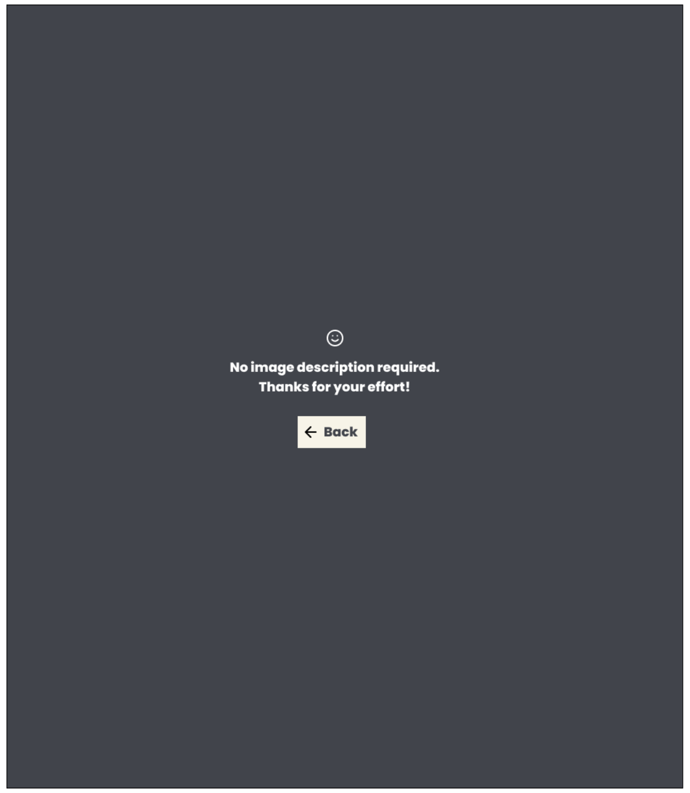
Finally, crowd workers also have the opportunity to confirm and manually modify the description.



Modification

Here is the description generated

# Gender	Both men and women
# Skin-tone	Medium-dark tone
#Dressing	Some of them are topless and some are wearing shabby cloth
# Hoe	Some people are holding hoes in their hands to dig the field. It consists of a long handles and stone blades.
# Long Rows	There are some long rows in the ground. They are digged by Native Americans.
# Image style	Watercolor painting



No image description required.
Thanks for your effort!

08

Evaluation

Due to the limitations of the COVID-19 epidemic, it's difficult to approach VIP users and test the interactions of the prototype offline. However, a pilot test with people with normal sights indicates that they are not familiar with the logic of accessibility services, so the feedback from simulated VIPs is hardly meaningful.

Limited by the remaining time of the project, it is decided to postpone the evaluation phase after the end of this graduation project. I will keep in touch with experts from Dedicon and KB to find an opportunity as soon as possible.

09

Conclusion

9.1 Review

The initial goal of this project was to research the requirements of the visually impaired people (VIPs) on image description (ID). Since the human in the loop approach is the most promising way to generate ID in the near future, research results on this topic can guide crowd workers' work and set the basis for a scalable way to generate ID, which is essential for VIPs' acquiring information in their daily life.

- With this goal, literature, desk, and field research are conducted to 3 questions regarding VIPs' requirements
- What images should be described?
- What content of an image should be described?
- How image should be presented?

The research results indicate it's difficult to find comprehensive and clear answers for these questions through normal approaches, considering the complexity of the image-text relationship (Question 1), the various factors influencing the preference (Question 2), and the enormous ways of organizing language (Question 3).

However, several opportunities are found as well. The first one is VIPs' **consistent needs of having control** over the presentation of ID (both presence and content). The second one is **the introduction of interactive ID**, which sets the conditions for VIPs' active expression of their needs. The third one is **the possibility of mapping variables and ID content** (e.g., to describe an image with people as a visual focus, you need to describe the activity, expression, etc.), which directly connects the needs that users actively report through the structured ID with the tasks that the crowd worker should perform.

Depending on these findings, it is proposed that **interactive ID can be leveraged as a novel approach to collect VIPs' needs and directly transform them into straightforward tasks for the crowd workers**. As a result, the design goal was defined as:

To develop a system that enables VIPs to have control on the ID content and is able to collect VIPs' requirements that can be transformed into straightforward description tasks for crowd workers

In the design phase, the prototype is developed to **verify the design goal and serve as a probe to learn users' dynamic process of perceiving the information of interactive ID**. Through a set of comparative experiments, the research results confirm the systems' function to collect user preferences and improve the content of image description accordingly. More importantly, the research uncovers how users utilize structured ID to acquire and digest information, which is a different procedure compared with that of current one-shot descriptions. It is also pointed out that structured ID can be organized in a way more line with people's mental model of accepting information (set the focus to develop the details/ acquire overall information).

As a result, the learning of this phase, combined with the knowledge about ID from the research phase is synthesized as a demonstrator of an IID (interactive image description) system.

Finally, let's review the original 3 questions:

What images should be described?

This research did not find a better answer to this question. But at least, now only a brief description is needed in advance.

What content of an image should be described?

The system can make a preliminary prediction on the required ID content through factors, and thus eliminate part of crowd workers' efforts to make the decision on these.

With users' feedback through interactive ID, we can continue to learn how to describe an image (regarding the length and content of ID) and gradually get a better answer.

How image should be presented?

Through interactive ID the requirements to organize ID content have been much lower. At the same time, interactive ID is a way that fits people's information mode well and provides users with autonomy.

9.2 Contribution

The research results of this project main contributes to the following two topics:

9.2.1 About how to generate ID

This thesis demonstrates the feasibility to collect user needs through interactive ID, and thus improves the current guideline for crowd workers. It can benefit the development of the human in-the-loop-approach in a long run.

This thesis verifies the feasibility of directly present separate ID content of different categories in the form of structured ID, which reduces crowd workers' efforts of making decisions and organizing language as well as improve end-users' experience. This finding is helpful to develop a new procedure to produce ID.

The research results also reveal the impact of contextual focus and image type on ID requirements, which can be a supplement to the existing research. The research results point out the deficiencies of the existing taxonomy that defines image focus and provide directions for improvement.

9.2.2 About the impact of interactive ID

Through a prototype that presents interactive ID together with context, this thesis reveals the influence of context on users' tendency to request information through structured ID. The perception model explains the user's mental activities when acquiring and evaluating additional descriptions, which is essential for the preparation and presentation of structured ID. These findings are helpful to understand how interactive ID may have an impact on the user experience, the requirements for description and the entire process of ID production.

In addition, through progressive ID, it is found that the demand for basic description is pervasive while the demand for detailed content is relatively rare, which provides a new opportunity to reduce workload.

9.2.3 Summary

In summary, the final outcome of this project can be regarded as a new approach to research VIPs' image description needs, and a new opportunity to improve the current way of generating ID.

For visually impaired people, the IID (interactive image description) system reveals a new opportunity to enhance their experience of reading image descriptions. More importantly, it allows them to enjoy the autonomy of controlling picture descriptions and gives them a way to actively express their needs.

For crowd workers, compared to the existing ways of providing them with instructions, templates or question lists, the results of this research propose a possibility to simplify their tasks and turn their work into an interactive and guided process.

For further researchers on VIPs' requirements on ID, the results of this project can be further developed as a new approach to conduct large-scale quantitative research on VIPs' requirements

For designers, the image perception model and the final design can serve as a starting point for future explorations of interactive images.

10

Reflection

10.1 Limitations

Participants

Due to the COVID-19 epidemic, I had no opportunity to contact Visually impaired people offline. Therefore, I did not have the opportunity to observe their reading activities and neither a chance to verify my prototype with them. Most feedbacks within these projects are from experts or simulated blind users. Obviously, their feedback cannot replace first-hand feelings from visually impaired people, and genuine details are missed. Therefore, further verification with visually impaired users is still required.

Prototype

The prototype of this prototype is not fully functional and it executes in the most ideal situation. Even though the images are presented with the context, the test setting is not able to simulate the state of concentration when reading long contents. Researchers' presence may also influence participants' will to interact with the system.

Long-term influence

Because all the tests are completed in a short time through online interviews. The freshness of the first contact with interactive ID may prompt users to be more willing to interact with it and have better patience. However, in the course of long-term use, the additional interaction threshold brought by interactive ID may be magnified. Further researches are still needed to learn its long-term usage rate and user experience.

Variety of test materials

Eight sets of images and their contexts are used in this thesis for investigation. Even though it already takes a long time to conduct the test for every participant because of the number of tested images, the variety of the test material is far from being able to cover all types of images and image-text relationships.

10.2 Recommendations

Current design verification

Interface for the crowd workers. Although the project proposed a preliminary idea for the crowd workers' interface, its impact on the crowd worker is not verified. Future work can elaborate on this to learn about possible problems that crowd workers may encounter in the production of structured ID. A comparison between the cost of generating ID through the IID system and normal approaches is also important.

Since there is no condition to run the IID system under a large-scale condition, what kind of data this system may generate and how the data could be used in that case is still unknown. Future researches can continue work on this to evaluate the IID system's value for requirements collection

Verification with visually impaired users. As is stated in the last section, due to the limitations of the COVID-19 epidemic, current research hasn't conducted tests with visually impaired people. It is of great value to approach them and collect their feedbacks.

Further explorations

Further exploration with structured ID. The current system's taxonomy for structured ID is completely based on the existing research from the picture description. Through experiments it is found that it is far from being able to handle all situations and includes all the possible image description content. It is suggested that experts from other fields should be invited to further development.

In addition, how to integrate AI captioning system to assist or replace part of crowd workers work is also an interesting direction. Since structured ID split the ID content, AI captioning is possible to deal with part of the description tasks. How to utilize the advantage of AI captioning and avoid the potential risks they may cause will be a valuable direction to explore.

10.3 Personal reflections

The process of exploring and completing this project is definitely challenging while valuable experience, especially under the COVID-19 situation. Although it is still a pity that I miss the opportunity to finalize this project with more comprehensive outputs in time, I do learn a lot from this unique design practice, which is completely different from the design projects I have done in the past in multiple aspects.

This project is my first design for visually impaired users, and actually the first time for the so-called disadvantaged group. Because of a lack of knowledge about this group and similar design experience, at the beginning of the project, I felt a bit uncertain and frustrated about how to approach the context of this project and conduct research activities. The COVID situation also added additional difficulties to this process. Fortunately, there are plentiful academic research results in this field. Experts from Dedicon and KB patiently brought me a lot of valuable insights as well. Although the process is completely different from what I expected at the beginning, in the end, I did accumulate a complete and in-depth knowledge about this field, which I am quite proud of. It is not only helpful for me to complete this project but also adaptable to related projects that I may encounter in the future. I even want to find a job in this field in the future or apply for a Ph.D. to study this topic in depth. This is a brand new field, and this project leads me to open the door to it.

Secondly, due to the COVID-19, in the research phase, I don't have much opportunity to approach my target user and start field research. For me, as a designer who highly relied on user feedback to conduct project analysis, synthesis, and even ideation, it is a huge challenge. It turns out that this challenge became an opportunity for me to learn how to study a topic systematically through literature review. Considering that this project is related to a lot of areas that I am not familiar with before (image description technology, visually impaired groups, crowdsourcing), I did spend enormous efforts to sort out a coherent framework for the literature I have read, which results in the literature review part of my report. (I would also like to thank Alessandro for his advice to me so that I can do this better) In this process, I learned how to explore the literature to understand the background and ask questions, and how to expand the reading materials of a certain topic for comparison, and how to summarize and reframe what I have learned and adapt them into the project context. This struggling procedure is like an unprecedented challenge to intelligence and energy for me, which is a feeling that the RAM of your brain has been overloaded. I am so thankful that I can go through these in the end.

Last but not least, in addition to the improvement in skills and knowledge, the most valuable learning from this project is the transformation of awareness. I originally regarded my target group as a "special group" in need of help, who may have different living habits or preferences from "normal" people. But this project taught me that they are absolutely normal people just like me, and this view applies to all disadvantaged groups. It should be believed that all groups enjoy similar fundamental needs. It is just that the norm of this society does not provide equal opportunities for some groups, and this is precisely what should be changed. This recognition also allows me to better accept myself. Coupled with the difficulties experienced in this project, as well as the staged failures and depressions, I have become more empathetic and grateful to people around me as well.

Finally, in this project, I also realized that I still have many shortcomings and there are various areas that I need to improve (such as time management, communication with multiple clients, and timely document and wrap-up for phased results). Knowing what to learn is as important as what has been learned. I believe that I am still in a process of constantly improving myself, and I will be able to do better in these aspects in the future.

Thanks for this project.

11

Acknowledgment

This work was carried out at TU Delft. I would like to express my gratitude and maybe also some apologies to all the people who supported and accompanied me.

Firstly of all, I want to say thanks to all of my supervisors: Prof.dr.ir. Alessandro Bozzon, Dr. Verma, H, Jeff Love, Dr. ir. Vermeeren, and Sepideh Mesbah. It's an extremely difficult challenge to proceed with this project under COVID-19. Your patience and suggestions are the most important factor for me to complete this project.

Thanks to experts from KB and Dedicon: Ted, Anne, Koen, and Anneke. Your input is critical in various stages of project completion.

Also, I want to express gratitude to everyone who has participated in my project. Special thanks to my friends: Bao Baihong, Li Junyao. Thank you for keeping communicating with me, helping me with valuable suggestions, and continuous encouragement.

Finally, I want to say my thanks to my parents Zhou Qin, Chu Qingsong. Without you, I could not complete this project. Your company and support are my last motivation to overcome the despair and advance this project in the most difficult moments.

Reference

1.Miltenburg, E. Van (no date) Pragmatic factors in [automatic] image description.

2.Stangl, A., Morris, M. R. and Gurari, D. (2020) ‘"Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions’, pp. 1–13. doi: 10.1145/3313831.3376404.

3.Petrie, H. et al. (1999) ‘Describing images on the Web : a survey of current practice and prospects for the future Centre for Human Computer Interaction Design City University London Northampton Square 2 The importance of describing images on the Web’.

4.Morash, V. S. et al. (2015) ‘Describe STEM with Template --- Guiding novice web workers in making image descriptions using templates’, ACM Transactions on Accessible Computing, 7(4). doi: 10.1145/2764916.

5.Morris, M. R. et al. (2016) ‘"With most of it being pictures now, I rarely use it": Understanding Twitter’s evolving accessibility to blind users’, Conference on Human Factors in Computing Systems - Proceedings, pp. 5506–5516. doi: 10.1145/2858036.2858116.

6. MacLeod, H. et al. (2017) ‘Understanding blind people’s experiences with computer-generated captions of social media images’, Conference on Human Factors in Computing Systems - Proceedings, 2017-May, pp. 5988–5999. doi: 10.1145/3025453.3025814.

7.Guinness, D., Cutrell, E. and Morris, M. R. (2018) ‘Caption Crawler: Enabling reusable alternative text descriptions using reverse image search’, Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3174092.

8.Nengroo, A. S. and Kuppusamy, K. S. (2018) ‘Accessible images (AIMS): a model to build self-describing images for assisting screen reader users’, Universal Access in the Information Society. Springer Berlin Heidelberg, 17(3), pp. 607–619. Doi: 10.1007/s10209-017-0607-z.

9.Salisbury, E., Kamar, E. and Morris, M. R. (2017) ‘Conversational Crowdsourcing as a Tool’, Aaai Hcomp 17, (Hcomp), pp. 147–156.

10.Von Ahn, L. et al. (2006) ‘Improving accessibility of the Web with a computer game’, Conference on Human Factors in Computing Systems - Proceedings, 1, pp. 79–82. doi: 10.1145/1124772.1124785.

11.Brady, E., Morris, M. R. and Bigham, J. P. (2015) ‘Gauging receptiveness to social microvolunteering’, Conference on Human Factors in Computing Systems - Proceedings, 2015-April, pp. 1055–1064. doi: 10.1145/2702123.2702329.

12.Saleous, H. et al. (2016) ‘Read2Me: A cloud-based reading aid for the visually impaired’, 2016 International Conference on Industrial Informatics and Computer Systems, CIICS 2016. IEEE. doi: 10.1109/ICCSII.2016.7462446.

13.Bigham, J. P. et al. (2006) ‘WebInSight: Making web images accessible’, Eighth International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2006, 2006, pp. 181–188. doi: 10.1145/1168987.1169018.

14. Salisbury, E., Kamar, E. and Morris, M. R. (2018) ‘Evaluating and

complementing vision-to-language technology for people who are blind with conversational crowdsourcing’, IJCAI International Joint Conference on Artificial Intelligence, 2018-July, pp. 5349–5353. doi: 10.24963/ijcai.2018/751.

15. Improving automatic image description in EPUB using Artificial Intelligence, Gregorio Pellegrino, https://www.youtube.com/watch?v=XZpgGNoBQoo&ab_channel=EDRLab

16. Bhowmick, A. and Hazarika, S. M. (2017) ‘An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends’, Journal on Multimodal User Interfaces. Springer International Publishing, 11(2), pp. 149–172. doi: 10.1007/s12193-016-0235-6.

17. Morris, M. R. et al. (2018) ‘Rich representations of visual content for Screen reader users’, Conference on Human Factors in Computing Systems - Proceedings, 2018-April. doi: 10.1145/3173574.3173633.

18. Morris, M. R. (2020) ‘AI and accessibility Discussion of ethical concern’, Communications of the ACM, 63(6), pp. 35–37. doi: 10.1145/3356727.

19. Power, C. et al. (2012) ‘Guidelines are Only Half of the Story : Accessibility Problems Encountered by Blind Users on the Web’, pp. 433–442.

20. Hodosh, M., Young, P. and Hockenmaier, J. (2015) ‘Framing image description as a ranking task: Data, models and evaluation metrics’, IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua, pp. 4188–4192.

21. Yao, T. et al. (2017) ‘Boosting Image Captioning with Attributes’, Proceedings of the IEEE International Conference on Computer Vision, 2017-Octob, pp. 4904–4912. doi: 10.1109/ICCV.2017.524.

22. Fang, H. et al. (2015) ‘From captions to visual concepts and back’, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June, pp. 1473–1482. doi: 10.1109/CVPR.2015.7298754.

23. Wu, S. et al. (2017) ‘Automatic alt-text: Computer-generated image descriptions for blind users on a social network service’, Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, pp. 1180–1192. doi: 10.1145/2998181.2998364.

24. Tariq, A. and Foroosh, H. (2017) ‘A Context-Driven Extractive Framework for Generating Realistic Image Descriptions’. IEEE, 26(2), pp. 619–632.

25. Act, E. A. (2019) ‘European Directive to Improve the Accessibility of Mainstream Ebooks’, pp. 1–3.

26. Kasdorf, B. B., The, C. and Union, E. (2019) ‘Make E-books Accessible Now’, pp. 1–4.

27. Virgili, G. et al. (2018) ‘Reading aids for adults with low vision (Review)’. doi: 10.1002/14651858.CD003303.pub4.www.cochranelibrary.com.

28. Fisher, D. (no date) ‘Barriers faced by blind and partially sighted people - RNIB Strategic prioritisation research Authors’.

29. Moraes, M. and Bruno, G. (no date) ‘Accessibility Policy : what people with visual impairment say’, pp. 1–15.

30. Limburg, H. and Keunen, J. E. E. (2020) ‘Blindness and low vision in The Netherlands from 2000 to 2020 — modeling as a tool for focused intervention Blindness and low vision in The Netherlands from 2000 to 2020 — modeling as a tool for focused intervention’, 6586. doi: 10.3109/09286580903312251.

31. WebAIM (2012) ‘Screen Reader User Survey #4 Results’, Screen Reader User Survey #8 Results, (September), pp. 1–30. Available at: http://webaim.org/projects/screenreadersurvey4/.

32. Vision, A. and Health, E. (2018) ‘A review of visual impairment’, pp. 1–4.

33. Zhong, Y., Matsubara, M. and Morishima, A. (2018) ‘Identification of Important Images for Understanding Web Pages’. IEEE, pp. 3568–3574.

34. Bigham, J. P. (2007) ‘Increasing Web Accessibility by Automatically Judging Alternative Text Quality’, pp. 349–352.

35. Huang, T. H. K. et al. (2017) ‘Is there anything else i can help you with?&#x00022;: Challenges in deploying an on-demand crowd-powered conversational agent’, arXiv.

36. Olson, J. S. and Editors, W. A. K. (no date) Ways of Knowing in HCI.

37. Grier, D. A. (2013). When computers were human. Princeton University Press.

38. Ibáñez, L. D., Reeves, N., & Simperl, E. (2020). Crowdsourcing and Human‐in‐the‐Loop for IoT. The Internet of Things: From Data to Insight, 91-105.

39. COOPER HEWITT GUIDELINES FOR IMAGE DESCRIPTION, Cooper Hewitt Guidelines for Image Description | Cooper Hewitt, Smithsonian Design Museum

40. Stappers, P. J., & Giaccardi, E. (2017). The Encyclopedia of Human-Computer Interaction.

41. Dunne, A., & Raby, F. (2013). Speculative everything: design, fiction, and social dreaming. MIT press.

42. World report on vision (WHO), 2019

43. SSMR. (2009) ‘Understanding the Needs of Blind and Partially Sighted People : their experiences , perspectives , and expectations’.

44. West, S. K. et al. (2002) ‘How Does Visual Impairment Affect Performance on Tasks of Everyday Life?’, 120(June).

45. Langelaan, M. et al. (2009) ‘Impact of Visual Impairment on Quality of Life : A Comparison With Quality of Life in the General Population and With Other Chronic Conditions Impact of Visual Impairment on Quality of Life : A Comparison With Quality of Life in the General’, 6586. doi: 10.1080/09286580601139212.

46. Binns, A. M. et al. (2012) ‘How Effective is Low Vision Service Provision ? A Systematic Review’, Survey of Ophthalmology. Elsevier Inc, 57(1), pp. 34–65. doi: 10.1016/j.survophthal.2011.06.006.

47.[Oogvereniging, 2018]Oogvereniging (2018). Slechtziendheid - Oogvereniging. [online] Oogvereniging. Available at: https://www.

oogvereniging.nl/oogaandoeningen/oogaandoeningen-overzicht/blind-doofblind-of-slechtziend/slechtziendheid/

48.Optelec International. (2018). Eye conditions. [online] Available at: https:// in.optelec.com/eyeconditions [Accessed 15 Feb. 2018].People: their experiences, perspectives, and expectations. University of Surrey RNIB.

49. Hypertext Markup Language - 2.0, https://tools.ietf.org/html/rfc1866]

50. HTML longdesc attribute https://www.w3resource.com/html/attributes/html-longdesc-attribute.php]

51. EPUB 3 http://diagramcenter.org/59-image-guidelines-for-epub-3.html]

52. DAISY https://daisy.org/activities/standards/daisy/]

53. Beginner's Guide to Image SEO – Optimize Images for Search Engineshttps://www.wpbeginner.com/beginners-guide/image-seo-optimize-images-for-search-engines/#:~:text=What%20is%20the%20Difference%20Between%20Alt%20Text%20vs%20Caption,are%20visible%20below%20your%20images.]

54. Zhao, Y. et al. (2017) '[05-16]The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments', Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), pp. 1–22. doi: 10.1145/3134756.

55. Bigham, J. P. et al. (no date) '[01-4]WebinSitu : A Comparative Analysis of Blind and Sighted Browsing Behavior', pp. 51–58

56. KB, no date, Introduction of KB https://www.kb.nl/organisatie

57. CaptionBot – For pictures worth the thousand words, 2017. https://www.captionbot.ai.

58.Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

59. H., S. Kawanaka, M. Kobayashi, T. Itoh, and C. Asakawa. Social accessibility: achieving accessibility through collaborative metadata authoring. ASSETS 2008, 193–200, 2008.

60. Bigham, J. P. et al. (2010) 'VizWiz : Nearly Real-time Answers to Visual Questions'.

61. Avila, M. et al. (2016) 'Remote assistance for blind users in daily life: A survey about be my eyes', ACM International Conference Proceeding Series, 29-June-20. doi: 10.1145/2910674.2935839.

62. Brady, E. et al. (2013) 'Investigating the Appropriateness of Social Network Question Asking as a Resource for Blind Users'.

63. Bridgwater, A. (2017) 'Definition – What is Human In The Loop ?', pp. 1–4.

64. Marsh, E. E. and White, M. D. (2003) 'A taxonomy of relationships between images and text', 59(6), pp. 647–672. doi: 10.1108/00220410310506303.

65. POET Image Description Guidelines, no date, http://diagramcenter.org/table-of-contents-2.html

66. COOPER HEWITT GUIDELINES FOR IMAGE DESCRIPTION, no date, https://www.cooperhewitt.org/cooper-hewitt-guidelines-for-image-description

67. Slatin, J., & Rush, S. (2002). Maximum accessibility: Making the web more usable for everyone.

68. Hudson, R (2003). Text Alternatives for Images. Retrieved 27 February 2005, from http://www.usability.com.au/resources/image-text.cfm

69. Bartolome, J. I. et al. (2019) 'Exploring aRt with a voice controlled multimodal guide for blind people', TEI 2019 - Proceedings of the 13th International Conference on Tangible, Embedded, and Embodied Interaction, pp. 383–390. doi: 10.1145/3294109.3300994.

70. Zhong, Y. et al. (2015) 'Regionspeak: Quick comprehensive spatial descriptionsof complex images for blind users', Conference on Human Factors in Computing Systems - Proceedings, 2015-April, pp. 2353–2362. doi: 10.1145/2702123.2702437

71. How blind people interact with visual content on social networking services', Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 27, pp. 1584–1595. doi: 10.1145/2818048.2820013

72. Gurari, D. et al. (2018) 'VizWiz Grand Challenge: Answering Visual Questions from Blind People', Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3608–3617. doi: 10.1109/CVPR.2018.00380.

73. Gregorio Pellegrino,2019, DPUB SUMMIT 2019 - 6 - Improving automatic image description in EPUB using Artificial Intelligence, , https://www.youtube.com/watch?v=XZpgGNoBQoo&t=672s

74. Zhong, Y. et al. (2015) 'Regionspeak: Quick comprehensive spatial descriptionsof complex images for blind users', Conference on Human Factors in Computing Systems - Proceedings, 2015-April, pp. 2353–2362. doi: 10.1145/2702123.2702437