Enhancing the Human Experience of Learning with Technology:
New challenges for research into digital, open, distance & networked education
European Distance and E-Learning Network (EDEN) Proceedings
2020 Research Workshop | Lisbon, 21-23 October, 2020
ISSN 2707-2819
doi: 10.38069/edenconf-2020-rw0040

REFLECTION ON HUMAN AND AUTOMATIC IMAGE DESCRIPTION IN ELEARNING CONTEXT

Manuela Francisco, Distance Education and eLearning Lab, Universidade Aberta, Portugal

Abstract

Images have a strong presence in educational contexts, particularly in online environments. When images add vital information for the understanding of a given subject, this information must be perceived by all students, including those who have a visual impairment. Although most web tools and platforms have a field for the description or alternative text, most images do not contain this information. Since 2016, some platforms and web services have been providing features, based on Artificial Intelligence, which present a brief description of the images to screen readers. Are these descriptions effective for people with visual impairment, in an eLearning context? Are they enough for a blind person to create a mental image? How do they differ from the description performed by humans? To answer these questions, it is necessary to have an idea of how we perceive images and how they make sense in our brain, according to our values and culture. Thus, in this work we present a reflection related to these questions, using examples of descriptions obtained by the google chrome feature "Get image description" and descriptions made by people in the various editions of the MOOC "Image description in web context".

Keywords: Inclusive eLearning, image description, automatic image description generation, web accessibility, visual impairment.

Introduction

The text/image binomial has been widely investigated in areas such as communication, cognitive psychology, and education. Mayer and Anderson (1992) conducted a study with animation and narrative, having concluded that text and image are more effective when they occur simultaneously in space and time, just as words are more effective when spoken than when written. Robinson and Nagar (2010), Spindler, Klaus, and Weber, (2010), Jonassen, Carr, and Yueh (1998), Jonassen (1996), and Newby et al. (1996) consider that the integration of different elements, such as audio, text, and image, into educational resources makes learning more effective and responds more effectively to different student

profiles. Aware that the diversity of formats and channels increases the students' motivation and the learning effectiveness, it is important to bear in mind the different functions that images can assume in this context. Bodmer (1992) states that images in didactic resources or used in class as a pedagogical strategy (complementary to written text or oral explanation) serve to enlarge, explain, interpret or decorate a written text, performing specific functions different from other contexts, such as in a painting or photography exhibition. Rodriguez Dieguez (1977) also mentions that one of the main tasks of didactics is the selection of communication codes, explaining that most materials use verbal codes. According to the author, different codes must be combined in pedagogy, namely the iconic code properly combined with the verbal code. Carney and Levin (2002), consider that images assume 5 functions when they are used simultaneously with the text:

- decorative function, when only decorates the text, that is, assumes a role of embellishment, may or may not be related to the text;
- representative function, when representing a part or all of the textual content;
- organizing function, when presented in a charismatic way an idea or a structure, a route;
- interpretive function, when it helps to clarify more complex text;
- transformative function, when reinforcing the memorization of textual information.

These authors also refer to a study conducted by Levin et al. 1987, where the benefits of the different image functions (in textual content) were analysed. They concluded that images with decorative function did not bring any benefit, unlike the images with the other 4 functions, namely the transformative function that presented greater benefits in learning.

Considering that blind people and some people with low vision do not have visual access to images, a textual equivalent should be given, particularly in eLearning context. This text must provide information about the elements contained in the image. If this textual equivalent does not exist or is poorly elaborated, the person who does not have visual access to the image may be at a disadvantage compared to those who have visual access to the whole content. Although accessibility guidelines for web content (WCAG) explain how to fill image/figure HTML attributes, the way it is described and what is written always depends on the visual literacy and interpretation of the *descriptor* (who describes it). This means that the same image can have numerous descriptions and may give too much information or not give enough and/or useful information. According to some studies (Francisco, 2008; 2015; Petrie, Harrison, & Dev, 2005) and blind people statements regarding their preferences (Kleege, 2008; Edison, 2012; NVDA discussion group and WAI discussion group), the image description must comply with certain parameters and be as objective as possible. Francisco (2015) conducted a study involving blind/low vision and

sighted people to verify the efficacy of the parameterized description. This study validated a matrix of parameters that helps the "descriptor" to create long descriptions and should be applied only to images considered vital in an educational resource. At this point, we may ask whether the automatic description generation is appropriate and whether it can replace the description performed by humans. When we refer to automatic description generation, we are considering the process referred by Brownlee (2019):

> *"Automatic image captioning is the task where, given a photograph, the system must generate a caption that describes the contents of the image".*

When we refer to the manual or human description we are considering the text that people write in HTML attributes whenever they insert an image into a digital content or webpage). Most of the time, these texts are only available to screen readers.

Image perception

According to Berger (1972), the act of seeing involves not only the optical function but also a set of information that allows us to identify what is seen. DeWitt (2013) explains that the constitution of the eye allows to perform the optical and perceptual functions, that is, the optical function captures the images focused by the retina and the perceptual function processes the luminous stimuli, transmitting them to the brain in a coded way. Optic nerve fibres have the function of transporting the information perceived by the retina to the brain. This means that to recreate a perceptive image a complex combination between the eyes and the brain is necessary. In psychological terms, this process is known as visual perception, however, and according to Collignon et al. (2011), Kupers et al. (2010; 2011), Bedny et al. (2009; 2011) and Burton, Diamond, and McDermott (2003), the absence of this process (e.g. blindness) doesn´t mean that some visual functions, such as the pupillary reflex or the visual cortex activation, are not used. The authors also report that in the absence of a sense or function, namely vision, there is a relocation due to the plasticity of the brain. Sacks (2010) adds that this plasticity can provide a blind person with a "hyperacuity" that will be far beyond the visual capacity of a person with vision, being possible to create mental images or visual representations through other sensory stimuli. However, individuals may not be "aware" of brain-activated functions, as reported by Damásio (2000; 2010).

Based on the assumption that the brain creates mental representations from the various stimulus, we can question whether the text can trigger the process of creating mental images. Humphreys et al. (2013) consider that language is often used to describe real-world situations, as such, words and phrases evoke mental representations of objects and experiences perceived by the senses. However, the authors report that this perspective is not consensual among researchers, considering that there are 3 different lines of thought

regarding semantic representations and descriptions of actions related to movement: (a) there is an organization around linguistics and the representation of action concepts (movement) and that is distributed along the sensory cortexes, but this doesn´t mean that linguistic representations are identical to the perceptions; (b) there is a neuronal connection between words and experiences related to these words, that is, words related to actions are learned in the context of the execution or observation of these actions; (c) representations are independent of the perception.

Despite the different perspectives, it seems evident that there is a strong relationship between the word and the imagery of the real world, whether it is perceived by vision or another sense, even if this relationship is not consciously presented to individuals.

Empirical study: Automatic description versus Human description

Since 2010 we have been conducting studies on image description, in workshops and MOOC. Participants are asked to select an image from the web and describe it. When no indications are provided about how and what to describe an image, most people create a brief description. The description is usually presented in a single sentence, consisting of 4 to 10 words.

In 2016, Facebook announced the incorporation of artificial intelligence that allowed to describe images to blind users. (Wu, Wieland, Farivar, Omid, & Schiller, 2017). On the Facebook website, they explain how does it work:

> *"Automatic alternative (alt) text uses object recognition technology to create a description of a photo for the blind and vision-loss community."*

This alternative text was formed by isolated words, as pointed out by Mazzoni (2019), Brownlee (2019), Karpathy (2016):

> *"The descriptions generated by artificial intelligence are based on labels that fall on the objects with greater prominence in the images."*

Since then, we have been testing, with screen reader users, different types of photos shared on this social network. In 2019 Google announced the new accessibility feature to describe images, and they explain on their support website how does it works:

> *"When you use a screen reader in Chrome, you can get descriptions of unlabelled images, for example, images that don't have alt text. Images are sent to Google to create the descriptions. If Google cannot describe an image, the screen reader will say 'No description available'."*

Although we are still testing this feature, we consider it appropriate to make a comparative analysis between the type of information presented in the descriptions, we already analysed:

- 100 descriptions made by people in the activity proposed in the open online courses "Describing images in a digital context" (2018/2019) and "Web accessibility: where to start" (2019/20120) (https://www.nau.edu.pt/cursos/);
- 30 descriptions obtained by Google Chrome feature "Get Image Descriptions", about images available in different Open Educational Resources (https://www.casadasciencias.org), using NVDA screen reader.

Table 1: Type of information presented in human descriptions and descriptions

Elements	Human description	Google automatic description
Sentence Beginnings	Photography of … Image of … In this image we see …	Appears to be …
Use of Adjectives	Yes	No
Colours	Yes	No
Environment	Sometimes	Sometimes
Main Objects/figures	Yes	Yes
Hidden information	Yes (e.g. Cities, dates, names)	No
Spatial references	No	No

In addition to these elements, it was found that some descriptions made by people (without previous indications of how to describe) repeat the title of the news or page where the image is inserted. The automatically generated descriptions refer to the objects highlighted in the image, only if there is a high contrast between the objects and the background.

Final considerations

Virtual environments can contain decorative or motivational images for those who can see, also educational content can use reinforcement images to a written subject. In these cases, it will make no sense to use long and complex descriptions. Only concise and succinct texts should be used to allow people to create a general idea of what is represented in the image. From the brief analysis carried out in our study, we can consider that the automatic description obtained with Google Chrome feature may have advantages over descriptions made by people who are not aware of how to write an image description. Artificial intelligence only interprets and associates the contrasts of spots, while people, as mentioned in the Image perception section, interpret what they see according to their experience and values.

However, we've found out that Google's feature doesn't display descriptions of images available on LMS platforms. Many web platforms, in the most current versions, force the user to write a description whenever uploading an image or to alternatively mark it as

decorative. In the latter case, screen readers do not detect the image, so they do not receive the description from google. We also found that artificial intelligence currently used does not interpret images based on schemes, drawings, complex or detailed photographs. In images with diagrams that display text, only the text is read, so, the context is lost. We consider that this study should be deepened since automatic descriptions are increasingly well structured and can be a good tool for creating alternative text in the eLearning context.

References

Berger, J. (1972). *Ways of Seeing.* London: British Broadcasting Corporation.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. In M. Merzenich (Ed.), *Proceedings of the National Academy of Sciences, 108*(11), 4429-34. doi:10.1073/pnas.1014818108

Bedny, M., Pascual-Leone, A., & Saxe, R. (2009). Growing up blind does not change the neural bases. In M. Merzenich (Ed.), *Proceedings of the National Academy of Sciences, 106*(27), 11312–11317. doi:10.1073/pnas.0900010106

Bodmer, G. R. (1992). Approaching the illustrated text. In G. E. Sadler (Ed.), *Teaching children's literature: Issues, pedagogy, resources* (pp. 72-79). New York: The Modern Language Association of America.

Brownlee, J. (2019, August 7) How to Automatically Generate Textual Descriptions for Photographs with Deep Learning. [Blog post] Retrieved October 3, 2020, from https://machinelearningmastery.com/how-to-captionphotos-with-deep-learning/

Burton H., Diamond J. B., & McDermott, K. B. (2003) Dissociating cortical regions activated by semantic and phonological tasks: A fMRI study in blind and sighted people. *Journal of Neurophysiology, 90*, 1965–1082. doi:10.1152/jn.00279.2003

Carney, R., & Levin, J. R. (2002) Pictorial illustrations still improve students learning from text. *Educational Psychology Review, 14*(1), 5-26. doi:10.1023/A:1013176309260

Collignon, O., Vandewallec, G., Vossa, P., Albouyc, G., Charbonneaua, G., Lassondea, M,. & Leporea, F. (2011). Functional specialization for auditory–spatial processing in the occipital cortex of congenitally blind humans. *Proceedings of the National Academy of Sciences (PNAS), 108*(11), 4435-40. doi:10.1073/pnas.1013928108

Damásio, A. (2000). *O mistério da consciência: Do corpo e das emoções ao conhecimento de si.* São Paulo: Companhia das Letras.

Damásio, A. (2010). *O livro da consciência: A Construção do Cérebro Consciente.* Lisboa: Círculo de Leitores.

DeWitt, D. (2013, July 1). Visual Perception: More than Meets the Eye. [Website] Retrieved October 3, 2020, from https://answersingenesis.org/human-body/eyes/visual-perception-more-than-meets-the-eye/

Edison, T. (2012, December 4). Describing Colors to Blind People. [YouTube]. Retrieved from https://www.youtube.com/watch?v=59YN8_lg6-U

Francisco, M. (2008). *Contributos para uma Educação Online Inclusiva: Estudo aplicado a casos de Cegueira e Baixa Visão* (Master dissertation). Universidade Aberta, Lisboa. Retrieved from http://hdl.handle.net/10400.2/1273

Francisco, M. (2015). *A descrição parametrizada da imagem para um eLearning acessível e inclusivo* (Doctoral dissertation). Universidade Aberta, Lisboa. Retrieved from http://hdl.handle.net/10400.2/4957

Humphreys, G. F., Newling, K., Jennings, C., & Gennari, S. P. (2013). Motion and actions in language: Semantic representations in occipito-temporal cortex. *Brain & Language, 125*(1), 94-105. doi:10.1016/j.bandl.2013.01.008

Jonassen, D. (1996). The Challenges of Teaching with Mindtools: Supporting Mindfulness and Self-Regulation. In D. Jonassen (Ed.), *Computers in the Classroom: Mind tools for critical thinking* (pp. 257-268). Columbus: Merrill/ Prentice Hall.

Jonassen, D., Carr, C., & Yueh, H.-P. (1998). Computers as Mindtools for Engaging Learners in Critical Thinking. *TechTrends, 43*(2), 24-32. doi:10.1007/BF02818172

Karpathy, A. (2016). NeuralTalk2. [GitHub] Retrieved October 3, 2020, from https://github.com/karpathy/neuraltalk2

Kleege, G. (2008). Blind Imagination: Pictures into Words. *Southwest Review, 93*(2), 227-239. Retrieved October 3, 2020, from http://www.jstor.org/stable/43473516

Kupers, R., Chebat, D. R, Madsen, K. H., Paulson, O. B., & Ptito, M. (2010). Neural correlates of virtual route recognition in congenital blindness. *Proceedings of the National Academy of the Sciences of the United States of America, 107*(28), 12716-12721. doi: 10.1073/pnas.1006199107

Kupers, R., Pietrini, P., Ricciardi, E., & Ptito, M. (2011). The nature of consciousness in the visually deprived brain. *Front Psychology, 2*, 19. doi:10.3389/fpsyg.2011.00019

Mayer, R. E., & Anderson, R. B. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology, 84*(4), 444-452. doi:10.1037/00220663.84.4.444

Mazzoni, D. (2019, October 11). Using AI to give people who are blind the “full picture”. [Blog post]. Retrieved October 3, 2020, from https://blog.google/outreach-initiatives/accessibility/get-image-descriptions/

Newby, J., Stepich, A., Lehman, D., & Russell, D. (1996). *Instructional technology for teaching and learning: Designing instruction, integrating computers, and using media.* Englewood Cliffs, New Jersey: Prentice Hall.

Petrie, H., Harrison, C., & Dev, S. (2005). Describing images on the Web: a survey of current practice and prospects for the future. In C. Stephanidis (Ed.), *Proceedings of 3rd International Conference on Universal Access in Human Computer Interaction, part of HCI International.* [CD-ROM]. Mahwah, NJ: Lawrence Erlbaum. Retrieved from http://www-users.cs.york.ac.uk/~petrie/HCII05 _alt_text_Paper.pdf

Robinson, T., & Nagar, A. K. (2010). Tactile Graphic Tool for Portable Digital Pad. In K. Miesenberger, J. Klaus, W. Zagler & A. Karshmer (Eds.), *Computers Helping People with Special Needs* (pp. 403 406). doi:10.1007/978-3-642-14100-3_60

Rodriguez-Dieguez, J. L. (1977). *La función de la imagen en la enseñanza. Coleção Comunicación visual.* Barcelona: Editorial Gustavo Gili.

Sacks, O. (2010). *The Mind's Eye.* New York: Knopf Publishers.

Spindler, M., Kraus, M., & Weber, G. (2010). A Graphical Tactile Screen-Explorer. In K. Miesenberger, J. Klaus, W. Zagler, & A. Karshmer (Eds.), *Computers Helping People with Special Needs* (pp. 474-481). doi:10.1007/978-3-642-14100-3_71

Wu, S., Wieland, J., Farivar, O., & Schiller, J. (2017). Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1180-1192.doi:10.1145/2998181.2998364