Rodrigues, J., Lopes, C.T. (2024). Automatic Description of Research Images: Utopia or Reality?. In: Garoufallou, E., Sartori, F. (eds) Metadata and Semantic Research. MTSR 2023. Communications in Computer and Information Science, vol 2048. Springer, Cham. https://doi.org/10.1007/978-3-031-65990-4_18

Automatic Description of Research Images: Utopia or Reality?

Joana Rodrigues[1,2](✉) and Carla Teixeira Lopes[1,2]

[1] Faculty of Engineering of the University of Porto,
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
joanasousarodrigues.14@gmail.com, ctl@fe.up.pt
[2] INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

Abstract. Data description is a fundamental step in Research Data Management (RDM). When it comes to images, the challenge is increased, as they have characteristics that differentiate them from other typologies. We conducted a study in which we obtained a set of 27 images described according to their content, by researchers of the projects where they are inserted. After obtaining the ground-truth that would support the analysis, we proceeded to two more stages of description, one through an automatic processing tool (Vision AI) and the other through researchers with no knowledge of the images. We concluded that the human description is more elucidative of the images' content, namely at a semantic level. In turn, the automatic tools enhance a more literal description. This study allowed us to reflect on the description of images in a research context and to discuss the potential of formal analysis and analysis of the semantic expression of images.

Keywords: Research Data Management · Images · Image description

1 Introduction

In Research Data Management (RDM), data description is one of the core tasks to ensure that all data are properly interpreted and understood and that they can be cited and reused. However, in many cases, the process involved in this task is time-consuming and can sometimes be discarded by researchers [2].

Scientific images constitute a great challenge, as there are still no sufficiently enlightening guidelines for orienting the process of describing images, taking into account their content and the sense and meaning by which they were produced.

Science benefits from images since the possibility of accurate registration enhances scientific research. If, in its beginnings, the image records were related to tribal peoples or new species discovered, with the passage of time it allowed to evolve in studies in which successive photographs were shot (one of the most famous studies in this regard is from 1881, by researcher Eadweard Muybridge, in which he took photographs in a sequence of running horses, in order to study the gallop), even others in which the technology associated with this documentary

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
E. Garoufallou and F. Sartori (Eds.): MTSR 2023, CCIS 2048, pp. 205–212, 2024.
https://doi.org/10.1007/978-3-031-65990-4_18

typology has boosted innovative studies in the scope of brain mapping and in the human body assessment, for example [6].

The production of images is, more than ever, in a growing phase. Nowadays, through various technological devices, such as mobile phones, tablets, and drones, it is possible to capture and broadcast images with great ease.

Vision is known as one of the most refined human senses and, for this reason, images are considered essential in human perception. Unlike human vision, which is limited to the visible band of the electromagnetic spectrum, computers, and imaging devices cover almost all of the electromagnetic spectrum [1]. Thus, imaging devices can process images generated from ultrasound, electron microscopy, and computer-generated images. This factor motivated the use of digital image processing in a wide field of applications [3].

The automatic description of images is a significant advance for the visualization and interpretation of images. It allows us to go beyond the human vision which, although very accurate, is fallible. Image processing tools enable us to observe in greater detail and can also be decisive in diagnoses, assessments, and decision-making [4].

But can image description depend solely on automatic processing tools? We believe that we are closer to obtaining the expected results when it comes to analyzing the visible content of an image, but when it comes to analyzing the semantics behind that image, namely its messages, and meanings, the path to follow seems longer. With this work, we intend to put this question in perspective, comparing the labels assigned to a set of images by the researchers that produce the images, researchers not associated with the context in which images were produced, and an automatic image processing tool (Vision AI). With these three perspectives, we intend to study two essential phenomena in the process of image description: formal analysis and semantic expression analysis.

2 Description of Images

The description of images implies a set of tasks aimed at obtaining results that allow a faithful representation of the images. Therefore, image description must involve the reliable creation of access points to information and the use of controlled languages, based on authority control. The use of controlled languages is useful because of a very justifiable benefit since an information system that does not have control of access points does not offer the user the guarantee that he retrieves all the information that actually exists and is relevant to him, even though, at first glance, the search seems to be effective. When we talk about controlled languages, we are referring to the sensible use of vocabulary expressions that properly represent the content without creating dual interpretations. It is important to note that there is no specific vocabulary for the representation of concepts and, although it must be effectively controlled, it cannot deviate from the original content of the document under analysis [5,12,13].

The representation of concepts can be performed through indexing terms, access points, descriptors, metadata, or classifications, for example. However, regardless of their typology, these vocabulary expressions must be able to represent concepts, the selected terms must not correspond to more than one concept and their presentation can be made as a simple term or a compound term (it is formed by two parts: nucleus and modifier) [9].

2.1 Formal Analysis

Formal image analysis is, above all, founded on two major pillars that allow it to be carried out properly. The first is the identification of the elements that are general properties of the image, namely the environment and the composition of the image, the spaces and scenarios (whether natural, urban, or peri-urban), the human relations in that space, the static aspects (such as gesture, body expressions, and clothing), and the objects and living beings. The second is about choosing the concepts that best represent the content of the image. These selected labels or expressions should be unambiguous and well-known terms. This selection is challenging because it is necessary to guarantee the interpretation of the image (do not select few concepts) and that the user does not have an overload of information regarding the image that may be redundant, duplicated, or not necessary (do not select too many concepts) [11].

2.2 Semantic Expression Analysis

The reading of images has a polysemic character, that is, the multiplicity of meanings existing in this source of information can be varied. As in the written document, the images require analysis, establishing comparisons, similarities, coincidences, and repetitions [14]. In the analysis of images' semantic expression, it is necessary to determine the aspects that are not visible in a literal and raw reading of the image, such as ideological messages, advertising messages, satire, and others [8].

A large portion of society manages to make only a superficial reading of an image, understanding its objective and superficial characteristics, leaving aside what is expressed in a semantic way [14]. For this reason, a greater understanding of the image and its importance is necessary [8].

The idea that non-specialists have the right to read images like someone who reads text is increasingly present. Manguel [7] even says that images are like stories waiting for a narrator. And the spectator must discover the explicit or implicit stories. However, these images are not always easy to read, so much so that, in order to try to read them, it is necessary to know how to see and interpret them. Therefore, one cannot work with the image as if it were totally transparent, but understand it as a language, produced within a socio-historical and cultural context.

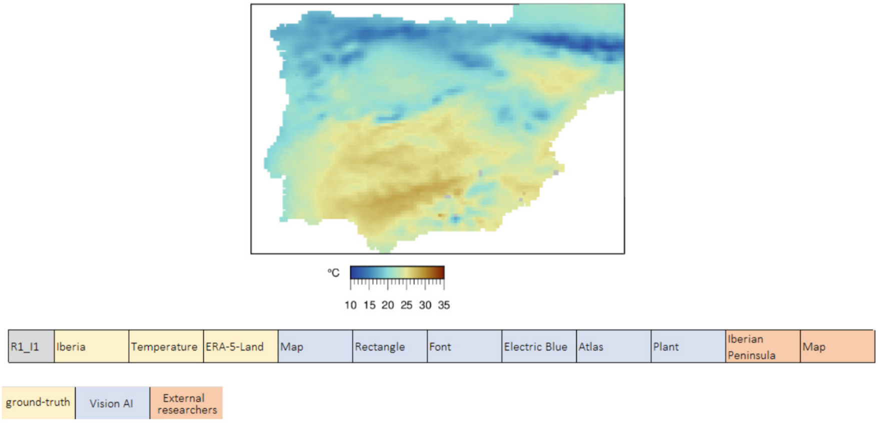
3 Methodology

The sample of images used for the study consists of 27 images, from five researchers from different projects and research domains. Of the total, 22 are in .png format and 5 in .jpg format.

The interaction with the researchers took place through an online form (Google Forms), which included the images that the researchers previously sent and which they should write down according to the content of each one.

A dataset was created that compiled all the images and their descriptions. We understand that the analysis of the results would work if we compared the first descriptions (which we classify as the ground-truth of the study) with descriptions made by automatic processing tools and by researchers who were not familiar with the images. For the first one, we use Vision AI[1], a tool that integrates computer vision models and Machine Learning in order to automatically obtain image labels. For the second, we selected researchers who had no previous knowledge about the images, nor who belonged to the same research domain. In the construction of the dataset, we normalized all labels (capitalization of the first letter and reduction of all labels to the singular) so that the analysis was more thorough.

The dataset is available at INESC TEC's data repository [10]. The dataset referring to the comparative analysis of the ground-truth with Vison AI and with external researchers will also be deposited in the same data repository. While the article is under revision, the dataset can be viewed here[2]. In the Fig. 1 we can see an example of an image and its respective descriptions.



R1_I1	Iberia	Temperature	ERA-5-Land	Map	Rectangle	Font	Electric Blue	Atlas	Plant	Iberian Peninsula	Map

ground-truth	Vision AI	External researchers

Fig. 1. Example description of an image

4 Results

Based on ground-truth, it was realized that the description obtained through Vison AI mostly did not correspond to what had been previously identified. As

[1] https://cloud.google.com/vision?hl.

[2] https://shorturl.at/aimv6.

we can identify in Fig. 2, only image R5_I4, obtained a totally coincident term between the researcher and the tool. This term was “map”. All the rest did not converge. Although most of the time, Vison AI collected more labels compared to the original description, this did not guarantee greater proximity of description.

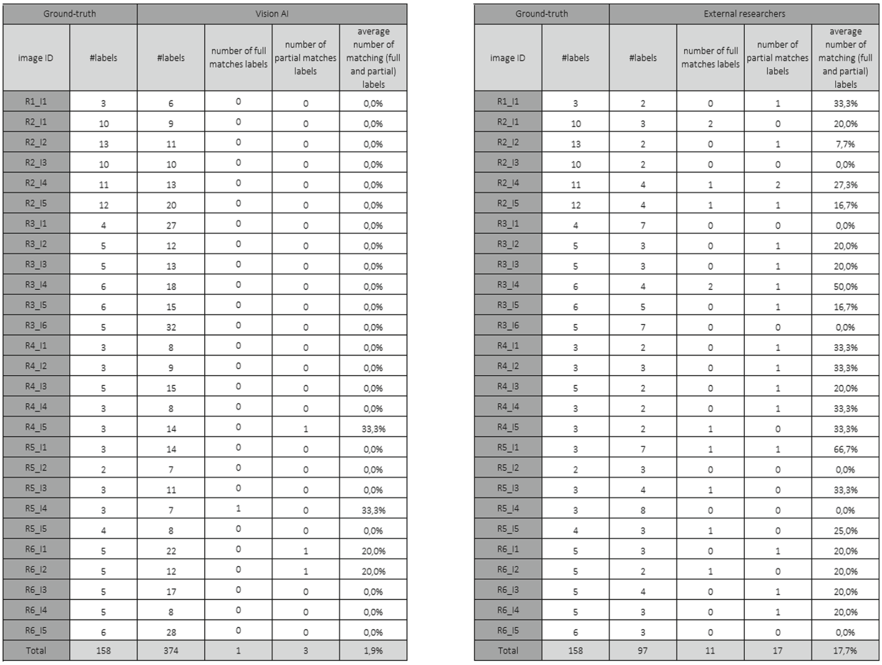
Given the results, we decided to find out if there were labels that, although not totally the same, had some similarities, named by partial matches. We checked this in three images, namely R4_I5, R6_I1, and R6_I2. In the first case, the researcher annotated the “line chart” and the Vison AI “line”. In the second case, the researcher annotated the “mouse cage” and the Vison AI “cage”. In the third case, the researcher annotated “animal welfare” and the Vison AI “animal”.

Descriptions by researchers from outside the projects where the images were produced present different results. Regarding the labels that are fully compatible with the original ones (named by full matches), eleven labels were found to be equivalent, and these labels pertained to nine images (R2_I1, R2_I4, R2_I5, R3_I4, R4_I5, R4_I6, R5_I3, R5_I5 and R6_I2). For R2_I1, both researchers identified “ArchoOnto” and “CIDOC-CRM”. For R2_I4 and R2_I5 “Linked Data”. For R3_I4 “comment” and “anotation”. For R4_I5 “Publications per year”. For R5_I1 and R5_I5“digital twin”. Para R5_I3 “blockchain”. Para R6_I2 “hen”.

As for the partial matches between the ground-truth and the external researchers, 16 of the 27 images under analysis have a coinciding label. This means that in more than half of the images under analysis (59.26%), the external researchers were able to find a label that matches the original description. Among the many examples are “B2Note tool use” and “B2Note”; “Metadata set” and “Metadata”; “DMP creation” and “DMP”, “class” and “class diagram”; “map” and “european map”, “Iberian” and “Iberian Peninsula”, among others that can be seen in the published dataset.

It was also found that in images R1_I1 and R2_I3, there was, in each of them, a coincident label when comparing the descriptions of the Vison AI and the external researchers. In the first case, the term used was “map” and in the second “diagram” (Fig. 2).

As for the average number of labels, with the exception of Researcher 2 (R2) who has an average of 11 labels per image, the others are similar, having an average of 4 labels per image. As for Vison AI, it was the one that obtained higher numbers, with an average of 14 labels per image. As for external researchers, they show very similar numbers to ground-truth, with an average of 4 labels per image. In the last column of the tables in the Fig. 2 (average number of matching (full and partial) labels), we can see that Vison AI presents averages with little similarity, with only 4 images having matching labels (1.9%). The external researchers show more similarity of labels, and only 6 images do not have any matching of labels (17.7%).



Ground-truth		Vision AI			
image ID	#labels	#labels	number of full matches labels	number of partial matches labels	average number of matching (full and partial) labels
R1_I1	3	6	0	0	0,0%
R2_I1	10	9	0	0	0,0%
R2_I2	13	11	0	0	0,0%
R2_I3	10	10	0	0	0,0%
R2_I4	11	13	0	0	0,0%
R2_I5	12	20	0	0	0,0%
R3_I1	4	27	0	0	0,0%
R3_I2	5	12	0	0	0,0%
R3_I3	5	13	0	0	0,0%
R3_I4	6	18	0	0	0,0%
R3_I5	6	15	0	0	0,0%
R3_I6	5	32	0	0	0,0%
R4_I1	3	8	0	0	0,0%
R4_I2	3	9	0	0	0,0%
R4_I3	5	15	0	0	0,0%
R4_I4	3	8	0	0	0,0%
R4_I5	3	14	0	1	33,3%
R5_I1	3	14	0	0	0,0%
R5_I2	2	7	0	0	0,0%
R5_I3	3	11	0	0	0,0%
R5_I4	3	7	1	0	33,3%
R5_I5	4	8	0	0	0,0%
R6_I1	5	22	0	1	20,0%
R6_I2	5	12	0	1	20,0%
R6_I3	5	17	0	0	0,0%
R6_I4	5	8	0	0	0,0%
R6_I5	6	28	0	0	0,0%
Total	158	374	1	3	1,9%

Ground-truth		External researchers			
image ID	#labels	#labels	number of full matches labels	number of partial matches labels	average number of matching (full and partial) labels
R1_I1	3	2	0	1	33,3%
R2_I1	10	3	2	0	20,0%
R2_I2	13	2	0	1	7,7%
R2_I3	10	2	0	0	0,0%
R2_I4	11	4	1	2	27,3%
R2_I5	12	4	1	1	16,7%
R3_I1	4	7	0	0	0,0%
R3_I2	5	3	0	1	20,0%
R3_I3	5	3	0	1	20,0%
R3_I4	6	4	2	1	50,0%
R3_I5	6	5	0	1	16,7%
R3_I6	5	7	0	0	0,0%
R4_I1	3	2	0	1	33,3%
R4_I2	3	3	0	1	33,3%
R4_I3	5	2	0	1	20,0%
R4_I4	3	2	0	1	33,3%
R4_I5	3	2	1	0	33,3%
R5_I1	3	7	1	1	66,7%
R5_I2	2	3	0	0	0,0%
R5_I3	3	4	1	0	33,3%
R5_I4	3	8	0	0	0,0%
R5_I5	4	3	1	0	25,0%
R6_I1	5	3	0	1	20,0%
R6_I2	5	2	1	0	20,0%
R6_I3	5	4	0	1	20,0%
R6_I4	5	3	0	1	20,0%
R6_I5	6	3	0	0	0,0%
Total	158	97	11	17	17,7%

Fig. 2. Correspondence of labels between the three types of description

5 Discussion and Conclusion

Although images have an undeniable capacity for representation, having access to the information they make available is no guarantee of full knowledge of what they translate. People are familiar with operations related to written text, but there are still many things to be studied when reading images, be it a painting, a film, a photograph, a graphic, or another. It is necessary to understand that the image is no longer just art and has become information and knowledge.

With technological advances, namely, the countless possibilities provided by Artificial Intelligence, much progress has been made in the sense of collecting and representing the content of images in an automatic way. It is undeniable the important role that these tools have in the most diverse areas, however, it is necessary to analyze all the challenges and limitations they include.

With this work, we realize that automatic image processing tools obtain better results when used on images such as photographs because they more easily collect the features present in the image. However, when we talk about graphs or diagrams the scenario is different. Besides the text inserted in the image, other labels are less consistent. We found that this analysis is done in a very literal way, getting labels like "rectangle" or "number" when it was a graph in a rectangular position, with associated numbers, leaving the theme and the variables it had to be described.

It was also noticeable that the formal analysis is much more present in the automatic processing tools, instead of the semantic expression analysis. This only confirms that these technologies are not yet fully available to delve into the semantic and meaning layer inherent in images. We have an example when one of the images was intended to reinforce the idea of "animal welfare", however, this was not possible to identify with a formal analysis of the image and therefore was not a concept identified by Vision AI.

Is necessary to raise awareness of the fact that human inclusion in the description process is important, as it allows validation of the results, guaranteeing their reliability. Automatic processing tools can be great helpers for researchers looking for help in the first layer of description, even more so when the volume of images is large, however, these should always be included in the workflow, as their knowledge of all the details inherent to the images is fundamental to a good description.

Acknowledgements. Joana Rodrigues is supported by research grant from FCT - Fundação para a Ciência e Tecnologia: PD/BD/150288/2019.

References

1. Acharya, T., Ray, A.: Image Formation and Representation, pp. 17–36 (2005). https://doi.org/10.1002/0471745790.ch2
2. Amorim, R.C.E.A.: A comparison of research data management platforms: architecture, flexible metadata and interoperability. Univers. Access Inf. Soci. **16**(4), 851–862 (2017)
3. Faria, D.: Análise e Processamento de Imagem. Faculdade de Engenharia da Universidade do Porto (2010)
4. Gonzalez, R., Woods, R.: Digital Image Fundamentals. 2 edn. (2000)
5. Gorman, M.: Authority control in the context of bibliographic control in the electronic environment. Cataloging Classif. Q. **38**, 11–22 (2004). https://doi.org/10.1300/J104v38n03_03
6. Huds, D.: Our Pastimes (2019). https://ourpastimes.com/the-impact-of-photography-on-society-12377030.html
7. Manguel, A.: Lendo imagens: uma história de amor e ódio. Companhia das Letras (2001)
8. Meiyu, L., Junping, D., Yingmin, J., Zengqi, S.: Image semantic description and automatic semantic annotation. In: ICCAS 2010, pp. 1192–1195 (2010). https://doi.org/10.1109/ICCAS.2010.5669742
9. Mendes, M.T.P., Simões, M.d.G.: Indexação por assuntos: princípios gerais e normas. Gabinete de estudos a&b (2002)
10. Oliveira, P., Rodrigues, J., Lopes, C.T.: Images annotated according to their content: a study on the description of data in image format in multiple domains [dataset]. INESC TEC research data repository (2021). https://doi.org/10.25747/szch-ve91
11. Pinto Molina, María, J.G.M.F., del Carmen Agustín, M.L.: Indización y Resumen de Documentos Digitales y Multimedia. 2 edn. (2002)
12. Ribeiro, F.: Indexação e Controlo de Autoridade em Arquivos. Câmara Municipal do Porto (1996), https://hdl.handle.net/10216/10721

13. Rodrigues, J.: Tendências atuais e perspetivas futuras em organização do conhecimento. Atas do III Congresso ISKO Espanha-Portugal and XIII Congresso ISKO Espanha (2017)
14. Stumpf, K.: Ensino da leitura de imagem: expressão semântica e conteúdo sintático. Revista Científica Multidisciplinar Núcleo do Conhecimento **7** (2020)