Identifying Metadata Elements in Photographic Repositories by Semantic Segmentation of Images with Deep Learning

Sinaí López-Castillo
Center of Innovation in Competitive Technologies (CIATEC, A.C.)
Isnardo Reducindo
Faculty of Information Science, Autonomous University of San Luis Potosí
Francisco Benita
Singapore University of Technology and Design

Introduction

Over time, humans have expressed information on different media in order to leave evidence of his activities, interests and way of thinking. The ways in which humans express the information have been evolving and adapting according to the different needs they have, among them we found walls, clay tablets, papyrus, parchment, paper, etc. However, one of the most interesting media of the nineteenth century is photography, because it allowed to leave evidence of an event with the exact context in which it happened. Various areas of science put their attention on photography as it gave the possibility of providing realities embodied in images due its probative character. Before the World War I, the value that users gave to photography was mainly artistic, but after the war the large number of news and people interested in visualizing in detail what was happening photography began to obtain a documentary character and becomes an indispensable communication form. Therefore, the information containing in the photographs attained historical and evidential value. At this point, the photography began to reach many parts of the world, revolutionizing the way people communicate, evidencing and informing others about a variety of events. Archival

documents and photographic archives started to spread worldwide. Moreover, the generalized access to the photographic cameras in the last decades increased the volume of photographs that existed in different private and public organizations and other parties that used this support as archival document related to its infrastructure, monuments, cultural objects, and so forth. In recent years, the access to these repositories became more complex due to the vast amount of photography stored by the institutions. Therefore, different standards and regulations guidelines started to emerge with the aim to organize and standardize the photography archival.

On the other hand, in the wake of the rapid expansion of Information Systems, mobile and web technologies, photography collections were evolving with the constant change of technologies. Digital photographic repertoires and digital photographic archives of public organizations have grown rapidly as a response to such innovative new technologies. To date, digital repositories are the common framework for users to consult photography archival. However, in order to preserve and have a well organization of the photographic archival trough digital repositories, the use of established standard metadata records for image description is required (Crawford 1984). The metadata fields provide different access points such that users can retrieve the photographs by applying different filtering criteria.

In this work, we propose an automated image description method for the photographic archival. In order to obtain a semantic description of the photographs, we provide a normalized access points through metadata elements using a deep learning algorithm for image segmentation. In the next sections, we give more details about appropriate image description, theoretical terms and technologies employed in this work, and then we explain our proposed framework and show the obtained results to validate our approach.

Background

The digital repertoires facilitate documents organization and its access, but it is necessary to provide them with an adequate description so that image retrieval is simple and fast. The description, in

both physical and digital formats, is a complicated process that must be carried out with caution. However, the large documentary volumes together with the little time that organizations have for organizing their repositories usually leads to outputs that complicate the retrieval process and make the photography description inefficient. For example, photography description is often performed by individuals that are not familiar with the description and standardization of terms through Knowledge Organization Structures (KOS) such as undergraduate interns and other non-information professionals. In addition, when the image description task is carried out by multiple non-professional individuals within a short period of time, it is likely that the quality of the recorded information will be inadequate

Despite working in a digital environment much of work today related to the archival (image) description is manual. On top of this, the old practice (Floyd and Oram 1992) of using undergraduate student employees to perform the routinized task of images description is still common nowadays. Hence, one of the challenges faced by public organizations such as government agencies or universities is the issue relating to the standardization of image description.

In the context of Information Systems, ontologies, thesaurus, vocabularies, terminologies and other types of KOS are widely used to facilitate and standardize the archival image description. In particular, ontologies have been viewed as a new type of vocabulary that ensure semantic interoperability with other vocabularies (e.g., the capacity of two or more systems to exchange and use the information that has been exchanged) through its ISO standard (ISO 25964-2). Semantic heterogeneity causes serious problems as it might happen that the same image might have different description depending on the person who is executing the task. It also might happen that the same semantic term is used to describe completely different properties. For example, one can think of "plant" which can refers either to factory or herbal. Semantic interoperability is an ontology-based approach that deals with this problem (Doerr 2003).

The KOS is a generic term used for referring to a wide range of items that models mutual relationships between elements of the repository (e.g. subject headings, thesauri, classification schemes and

ontologies). They are characterized by different specific structures and functions, and they are used in a plurality of contexts by diverse communities. Although different in nature, what they all have in common is that they are designed to support the organization of knowledge and information in order to make their management and retrieval of information easier (Mazzocchi 2018). Despite the tools provided by regulations, standards, KOS and technologies, the description of photographic collections and repositories has not evolved, that is, it still requires the support of many collaborators for its description and the result remains the same.

Another of the advantages that technology evolution allows is the image labelling using tools that allow us to know the most relevant characteristics of a photograph (image) by the scene segmentation. Image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels). Then, after the implementation of computer vision tools the image is characterized by means of a conceptual representation using labelling (Peres et al. 2010). In recent years several powerful tools performing image segmentation (*SEGNET* by (Badrinarayanan, Kendall, and Cipolla 2017) or *DeepLab* by (Chen et al. 2017)) and object detection (Mask R-CNN by (He et al. 2017)) have been developed. Image segmentation has achieved very good results in the field of medicine, urbanism, construction and robotics. Image segmentation is a well-studied computer vision task which has been significantly improved with the boom of convolutional neural networks (Krizhevsky, Sutskever, and Hinton 2012). This technique has been used to look at the problem of automatically describing archival images, at least in its conceptual-representational mode (Frigui and Caudill 2006).

With the analysis of the needs in the photographic collections and the advantages provided by the image labelling, the opportunity offered by using the latter tool is identified in order to start automating the description process. This approach not only helps to reduce the time needed to describe and image but also improves the normalization description and leads to a better retrieval performance. In this work, we chose the Dublin Core standard (Weibel 1997). which is a simple and effective 15-element set for describing images. The 15

core elements are: title, description, subject (keyword), type, relations, creator, contributor, publisher, rights, date, format, identifier and language.

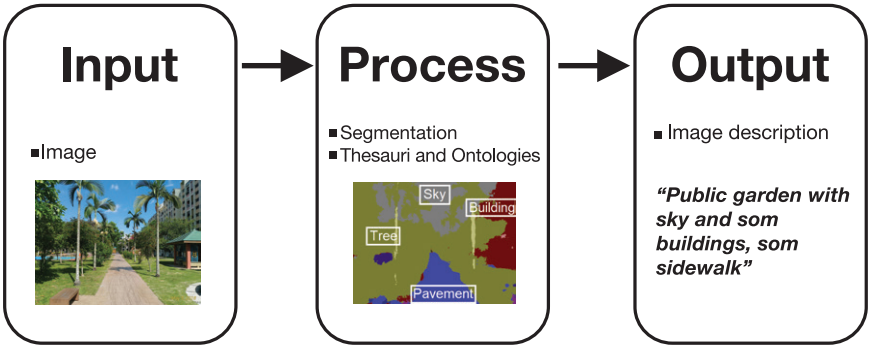
Nevertheless, the image segmentation does not provide the necessary elements to define all the metadata that make up the standard. Conversely, the tool can support the definition of the most time-consuming elements such as title and description. That is, from the elements that make up the image, image segmentation may allow the creation of access points so that users can identify the images with respect to their content.

Proposed framework

To test our approach, we used 24 photographs of a residential area in Singapore. The "description" element of the Dublin Core standard was generated in three different ways. First, a final-year undergraduate student from library management background, i.e., "Person", described each image according to her own criteria. Note that this process is commonly adopted by public institutions as it is inexpensive and fast, but it leads to semantic heterogeneity issues. Second, a graduate student of archival management, i.e, "Archivist", described the images using a thesaurus. The Archivist description can be considered as a standardized description that generates normalized access points. This is also a common practice within institutions that aim improving the quality of their digital repositories. Third, an expert in library management and archival management systems, i.e., "Target", described the images following the standardized thesaurus and ontologies (UNESCO, n.d.). This Target constitutes the gold standard of the normalized descriptions. Fourth, captions were generated after implementing a Deep Learning tool of image segmentation. In this work open source SegNet algorithm was employed (Badrinarayanan, Kendall, and Cipolla 2017). SegNet allows the identification of 12 different classes related to outdoor environments, namely, sky, building, pole, road, road marking, pavement, tree, sign symbol, fence, vehicle, pedestrian and bike. But the convolutional neural network can be trained to identify new segments such as person, different ani-

mals, everyday items, food or indoor furniture (chairs, lamps, tables, floor, wall, etc.). *Figure 1* illustrates the description approach based on image segmentation and Deep Learning techniques.

Figure 1. Process description.



Next, the semantic similarity among all three descriptions versus the Target description (expert normalized description) is performed adapting the semantic interoperability metrics detailed (Yahia, Aubry, and Panetto 2012). Given a set A composed by n semantic concepts ci (i = 1,..,n), we can define the lexical AL and nonlexical ANL subsets as follows:

$$A_L = \{c_i \,|c \text{ is a lexical concept } \forall c_i \in A\}$$

$$A_{NL} = \{c_i \,|c \text{ is a nonlexical concept } \forall c_i \in A\}$$

such that

$$A_L \cup A_{NL} = A \qquad \text{and} \qquad A_L \cap A_{NL} = \emptyset$$

We can define a lexical concept as a term that can be written down, consist of letters, numbers, among other characters. A non-lexical concept is a concept that cannot be written down and is named by lexical concepts (Lezoche, Aubry, and Panetto 2012). Then, given two semantic sets A and B, we obtain the semantic relations set R as fallows

$$R = \{\langle c_i^b, (c_1^a, ..., c_j^a)\rangle \mid sem(c_i^b) \subseteq sem(c_1^a, ..., c_j^a)\ \forall c_j^a \in A, \forall c_i^b \in B\}$$

where sem(c) represents the semantic interpretation of the concept c. Now, removing the non-lexical subset of A we can define the lexical relations set RL as fallows

$$R_L = \{\langle c_i^b, (c_1^a, ..., c_j^a)\rangle \mid sem(c_i^b) \subseteq sem(c_1^a, ..., c_j^a)\ \forall c_j^a \in A, \forall c_i^b \in B_L\}$$

Finally, we defined the semantic similarity potential V between two semantic sets A and B as

$$V_{A \to B} = \frac{|R_L|}{|R|}$$

where $|\cdot|$ represents the set cardinality. It is important to note that the semantic similarity potential is not bijective, for this reason we compute both sides potential in all cases.

RESULTS

In order to appreciate the differences between the tree description cases (under graduated student (Person), graduated student (Archivist) and automated (SegNEt)) versus the target (expert normalized description) Figure 2 displays the boxplot of the semantic similarity potential between all six types of descriptors (both sides of potential).

By visual inspection it is not clear if mean values are different among types of descriptors, hence Table 1 summarizes the p-values from the Kruskal-Wallis test. This is a non-parametric test than help us to decide whether types of descriptors came from the same distribution. The table compares all three combinations of descriptions and, with the exception of pairs (Target-Person, Person-Target), (Target-Person, Automated-Target), (Target-Archivist, Person-Target) and (Target-Automated, Archivist-Target), all other pair of combination between descriptors came from the different distributions. In other words, we found evidence that the description generated by the different individuals and automated tools is semantically different.

Figure 2. Boxplot of semantic similarity potential.

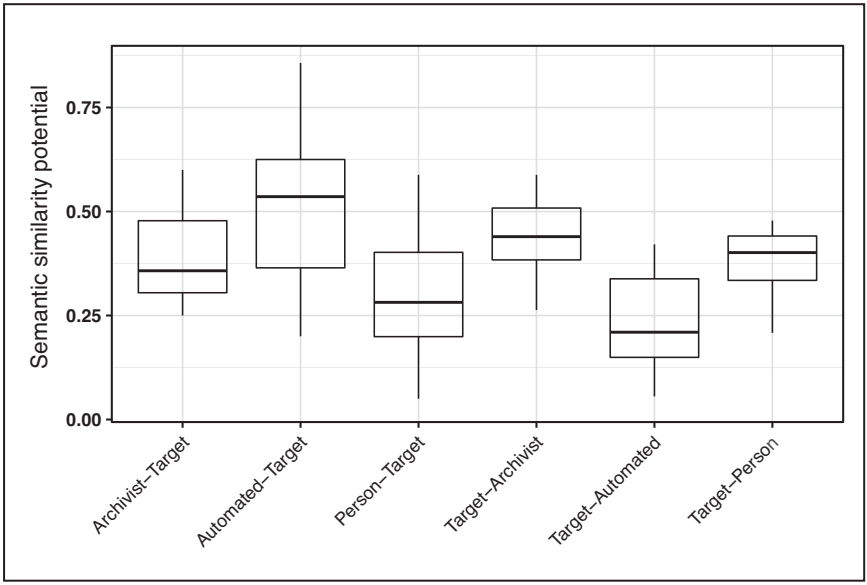


Table 1. P-values of the Kruskal-Wallis test of semantic similarity potentials
Note: Estimates in bold are statistically significant (p-value < 0.05).

	Person-Target	Archivist-Target	Automated-Target
Person-Target	0.0664	**0.0005**	0.0928
Archivist-Target	0.9589	**0.0487**	**0.0004**
Automated-Target	**0.0108**	0.2307	**0.0004**

In addition to the presented semantic similarity potential results, it is worth noticing that the automated photographs description took a computation time of 2.79 seconds for all 24 images. Alternatively, the description generated by the by under graduated student, e.g., "Person", required 4 hours (one day of work), the graduated normalized student description, i.e., "Archivist", required 16 hours (four days of work), and the expert description, (i.e., "Target"), took 8 hours (two days of work).

Concluding remarks

In this work, we propose the use of image segmentation combined with Deep Learning techniques as a feasible framework to derive automated metadata descriptors for photographs. The experimental results reveal that our automated descriptors have high semantic similarity compared with expert normalized description which evidence the effectivity of tour approach. Furthermore, the automated description is not only considerably faster (couple of seconds vs couple of hours) but also could help improving knowledge organization of the digital photographs repositories as it provides well normalized access points. Our approach also removes the ambiguities in the descriptions that can arise when the task is performed individuals that are non-experts in the field of photograph documentation such as undergraduate students.

Finally, a natural extension of this work is the improvement of the semantic concepts in the automated description by adding other characteristics of the objects located in the images such as colours or position in the scene. In doing so, the convolutional neural networks need to be trained to detect other type of objects. The implementation of natural language processing techniques is also required in order to derive more elaborated image descriptors.

Bibliography

Badrinarayanan, V., A. Kendall, and R. Cipolla. (2017). "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–95.

Chen, L. C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2017). "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4): 834–48.

Crawford, W. (1984). MARC *for Library Use: Understanding the USMARC Formats.* USA: Knowledge Industry Publications.

Doerr, M. (2003). "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine* 24 (3): 75.

Floyd, B., and R. Oram. (1992). "Learning by Doing: Undergraduates as Employees in Archives." *The American Archivist* 55 (3): 440–52.

Frigui, H., and J. Caudill. (2006). "Unsupervised Image Segmentation and Annotation for Content-Based Image Retrieval." *Proceedings of the IEEE International Conference on Fuzzy Systems* (ICFS 2006), 72–77.

He, K., G. Gkioxari, P. Dollár, and R. Girshick. (2017). "Mask R-CNN." *Proceedings of the IEEE International Conference on Computer Vision* (ICCV 2017), 2961–69.

Krizhevsky, A., I. Sutskever, and G.E. Hinton. (2012). "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems (NIPS 2012)*, 1097–1105.

Lezoche, Mario, Alexis Aubry, and Hervé Panetto. (2012). "Formal Fact-Oriented Model Transformations for Cooperative Information Systems Semantic Conceptualisation." *Lecture Notes in Business Information Processing,* 117–31. https://doi.org/10.1007/978-3-642-29958-2_8.

Mazzocchi, Fulvio. (2018). "Knowledge Organization System (KOS)." *ISKO Encyclopedia of Knowledge Organization.* Birger Hjørland and Claudio Gnol.

Peres, F A, F R Oliveira, L A Neves, and M F Godoy. (2010). "Automatic Segmentation of Digital Images Applied in Cardiac Medical Images." *Pan American Health Care Exchanges,* 38–42. https://doi.org/10.1109/PAHCE.2010.5474606.

UNESCO. n.d. "UNESCO Thesaurus." http://vocabularies.unesco.org/browser/thesaurus/es/?clang=en.

Weibel, S. (1997). "The Dublin Core: A Simple Content Description Model for Electronic Resources." *Bulletin of the American Society for Information Science and Technology* 24 (1): 9–11.

Yahia, Esma, Alexis Aubry, and Hervé Panetto. (2012). "Formal Measures for Semantic Interoperability Assessment in Cooperative Enterprise Information Systems." *Computers in Industry* 63 (5): 443–57. https://doi.org/10.1016/j.compind.2012.01.010.