Coded Visions: Addressing Cultural Bias in Image Annotation Systems with the Descriptions and Situations Ontology Design Pattern

Delfina Sol Martinez Pandiani[1] and Valentina Presutti[2]

[1]Department of Computer Science and Engineering (DISI), University of Bologna
[2]Department of Modern Languages, Literatures, and Culture (LILEC), University of Bologna

1 Introduction

Rapidly expanding image collections, including visual catalogues from the Cultural Heritage field, have attracted significant research efforts for the automatic understanding, retrieval, and management of images (Llamas et al., 2016; Saleh, 2018; Wang et al., 2021). Graph-based resources are used to organize and assign meaning to the raw content of images, as well as to interconnect such meanings to other images and datasets. Scene-graphs, ontologies and other graph-based linguistic resources, such as WordNet (Miller, 1995), are considered the best models for semantically annotating complex image scenes, to support tasks such as their automatic understanding and interpretation. These models enable the extraction of fine-grained information from image collections, which is crucial for obtaining relevant semantics as well as relational information between recognized objects (Sager et al., 2021). Some image labelling and retrieval systems rely on large, strongly connected Knowledge Graphs, such as ConceptNet (Liu and Singh, 2004) and Framester (Gangemi et al., 2016), to assign and/or refine the meaning of raw features of images (Martinez Pandiani and Presutti, 2021; Samih et al., 2021; Tariq and Foroosh, 2017). In turn, these data are used to capture and organize the content of images and as ground truths for training computer vision systems.

However, meanings assigned to images are not culturally agnostic. Rather, visual forms illustrate, and circulate, concepts both by providing links to depicted objects through raw features–such as lines, color, shape, and size–as well as through what Barthes called an image's 'connotation': a second layer of meaning made from culturally coded elements (Barthes, 1981). These culturally coded elements guide humans in making decisions of where to identify objects, how to label them, how to attribute features to them, and how to establish relationships between them.

In fact, classification systems which provide ground-truth meanings to widely used repositories of images, such as ImageNet (Deng et al., 2009), may stabilize contested political categories in ways that are difficult to see. For example, an image of an

indigenous person wearing a traditional garment may be classified as "half naked", or an image of a person sleeping on a bench as "homeless". Because classification systems decide which features make a difference, if the culturally coded aspects of meaning assignment and organization are not incorporated into the representation of the knowledge itself, there is a significant pitfall of implicit bias and unawareness of how image classification systems may be echoing, amplifying, or restructuring certain interpretations, or visions, of the world. As such, the additional layer of culturally-coded interpretation demanded by visual material ought to be made explicit, instead of directly algorithmically coded.

This paper proposes an ontology-based module that specifically aims to represent the culturally-bound processes of annotating images, as a promising tool for more faithfully representing the multi-layered process of visual understanding. We propose a specialisation of the Description and Situations ontology design pattern (Gangemi and Mika, 2003), which models non-physical objects whose intended meanings results from statements, to account for 1) the explicit tracking of how meaning(s) associated to images in the annotation processes come from culturally coded annotation situations and 2) how these assignments of meaning can be compared.

2 Motivation

Increasingly, graph-based models are used to organize meaning of the raw content of both photographic and art images, for example in the form of scene graphs (e.g., Visual Genome (Krishna et al., 2017)) or taxonomies (e.g., the Tate collection[1]). Many of these structures are used to provide labels for objects detected by an annotator–be it artificial, human, or a combination of both. These assigned labels are then integrated into graph-based structures connecting images at face value, i.e., they are treated as facts, rather than as beliefs or assertions bounded to a specific annotation context.

However, the extraction and representation of semantic elements from visual material constitutes the construction of a code system that can–and should–be made explicit. This is because visuality, different from the purely biological process of vision, is flexible and encompasses "the way that we encounter, look at, and interpret images based on the social, cultural, technological, and economic conditions of their viewing" (Giotta, 2020, 32). That is, visuality is a cultural practice with a history marked by different habits or ways of seeing, as well as different types of spectators (Foster, 1988).

2.1 Cultural Biases in Computer Vision

Importantly, the cultural-boundedness of visuality remains when images are passed through computer vision pipelines, including through the seemingly straightforward process of object detection. As argued by Arnold and Tilton (2019) in their distant viewing methodological and theoretical framework for large-scale study of visual materials, there is a code system required, which is culturally and socially constructed. As such, labelling and classification systems can act as forms of centralized power reflecting a specific group's or culture's values.

Nevertheless, humans tend to show high confidence in the integrity and objectiveness of benchmark datasets' image labels, including a naive faith both in photographic truth and in the unretouched quality of images discovered online (Giotta, 2020). This faith in

[1] https://github.com/tategallery/collection/issues/27

the cultural agnosticism of images' semantic labeling is intricately related to the topic of machine-learning bias, currently leading to worldwide discussions on the importance of representation in artificial intelligence systems. Practically and generally speaking, models will always incorporate some form of bias, since it is unrealistic to build a complete representation of the real world. However, images and the field of computer vision is especially vulnerable, as bias can occur in a variety of stages, from human reporting and selection bias to algorithmic and interpretation bias (Laranjeira et al., 2021).

The idea that automated systems are not inherently neutral and instead reflect the priorities, preferences, and prejudices of those who have the power to mold artificial intelligence is an increasingly public topic of discussion. For example, recent works have demonstrated that the geographical sampling of Flickr images as well as the use of English as the primary language for dataset construction and taxonomy definition result in inherent cultural bias within the datasets (de Vries et al., 2019), with work being done to design new annotation procedures that enable fairness analysis (Schumann et al., 2021). Another example is Gender Shades[2] (Buolamwini and Gebru, 2018), an investigation from the Massachussets Institute of Technology on the false assumption of machine neutrality, and the coded gaze–this algorithmic 'way of seeing' which classifies content through researcher- and machine-labeled categories–which "reflects both our aspirations and our limitations" according to its coiner (Buolamwini, 2017, 44).

2.2 Identifying and Representing Cultural Bias

The bias that is the focus of this paper is not to be confused with the bias term in machine learning models or prediction bias, which mathematically speaking is an intercept or offset from an origin. Instead, we intend cultural bias in the sense presented by The American Psychological Association (2015):

> the tendency to interpret and judge phenomena in terms of the distinctive values, beliefs, and other characteristics of the society or community to which one belongs. This sometimes leads people to form opinions and make decisions about others in advance of any actual experience with them (prejudice).

This article focuses specifically on how to technically account for the way that cultural bias permeates the moment of assigning semantic labels to pixel areas of images during the creation of computer vision datasets. This specific moment of human-led or human-evaluated annotation is critical, as these labels become part of input data of widely used models across many different fields, and in that way the "data itself" can host a lot of human biases, such as stereotyping, prejudice or racism. In this sense, this paper is concerned with the intersection of cultural and measurement bias. Measurement bias refers to faulty, low quality or unreliable measures when collecting data, such as inconsistent or unreliable labeling of samples, which can have many causes such as insufficient label options (e.g. binary gender (Scheuerman et al., 2019)) or from subjective views from labelers.

This paper acknowledges the gap of technical frameworks to concurrently model how meaning is culturally-bound and assigned, via beliefs, to raw visual content depending on the annotation situation within image annotation pipelines. It hypothesises that network-based structures can provide technical frameworks in order to mitigate harm

[2] http://gendershades.org/, https://www.media.mit.edu/projects/gender-shades/overview/

from implicit cultural bias.

2.3 Ontologies for Digital Hermeneutics

Specifically, it sees the use of ontology-based representations for image label data as promising, since associating a meaning with complex scenes may require an explicit and symbolic representation of the domain of knowledge, and the ability to reason over it. As such, ontologies provide a powerful framework to address this issue for image analysis and interpretation (Bannour and Hudelot, 2011); one of their biggest benefits is that the semantics is made explicit and allows queries to be formulated in terms of concepts and their relationships.

Ontologies for modeling interpretation related to cultural artifacts, including visual works of art, are more widespread. The IECA ontology[3]provides a conceptual and ontological model to represent interpretative encounters between human observers and cultural artifacts, and to investigate the content, context, and relationships between alternative interpretations of cultural objects. A recent work by Baroncini et al. (2021) about modelling interpretation and meaning for art pieces, presents a data model for describing iconology and iconography. Previously, Daquino and Tomasi (2015) proposed the Historical Context Ontology (HiCO), an ontology aiming to outline relevant issues related to the workflow for stating, and formalizing, authoritative assertions about context information of cultural heritage artifacts. Also recently Carboni and de Luca (2019) have presented the VIR (Visual Representation) ontology, constructed as an extension of CIDOC-CRM (CIDOC Conceptual Reference Model), which sustains the recording of statements about the different structural units and relationships of a visual representation, differentiating between object and interpretative act.

3 The IAS Ontology Module

This paper discusses the potential of using an ontology-based technique for the explicit tracking of how meaning(s) associated to images come from culturally coded annotation situations, and of how these assignments of meaning can be compared. Our approach is based on the idea that an image's semantic labels depend upon the specific annotation situation under which it is interpreted. Specifically, the idea is that an `Image` can be involved in multiple annotation situations, which may differ on account of one or many aspects: when, where, and why they were performed, by whom, whether and how the annotator was remunerated, etc. In order to represent this model formally, we designed the Image Annotation Situations (IAS) ontology module, based on the Descriptions and Situations (DnS) ontology pattern (Gangemi and Mika, 2003), which supports a first-order manipulation of theories and models. DnS was chosen as a core design pattern because it allows for the modeling of non-physical objects, such as an image annotation system or situation, whose intended meaning results from statements, i.e. they arise in combination with other entities.

Influenced by the work in Vacura et al. (2008), in the IAS module we consider that the image annotation process is a situation (i.e. a context reified in the class `ImageAnnotationSituation`) that needs to be described via an `ImageAnnotationDescription`, and that represents the state of affairs of all entities related to that state of affairs: the actual `Image` to be annotated, the `Annotator` and information related to them, the time and place of the annotation situation, whether

[3] https://ieca-ontology.github.io/

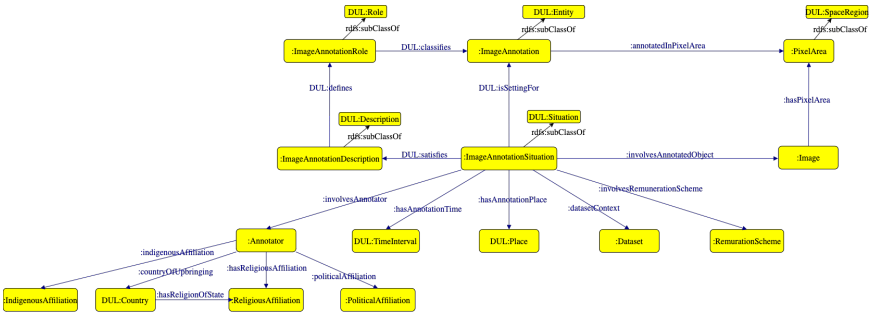

Figure 1: The IAS ontology module is aligned to and reuses patterns from DOLCE+DnS Ultralite (DUL) foundational ontology (Gangemi et al., 2002) in order to represent and give meaning to the culturally-bound context in which annotation data was created during an image annotation process. All un-prefixed classes belong to the namespace of the IAS ontology.

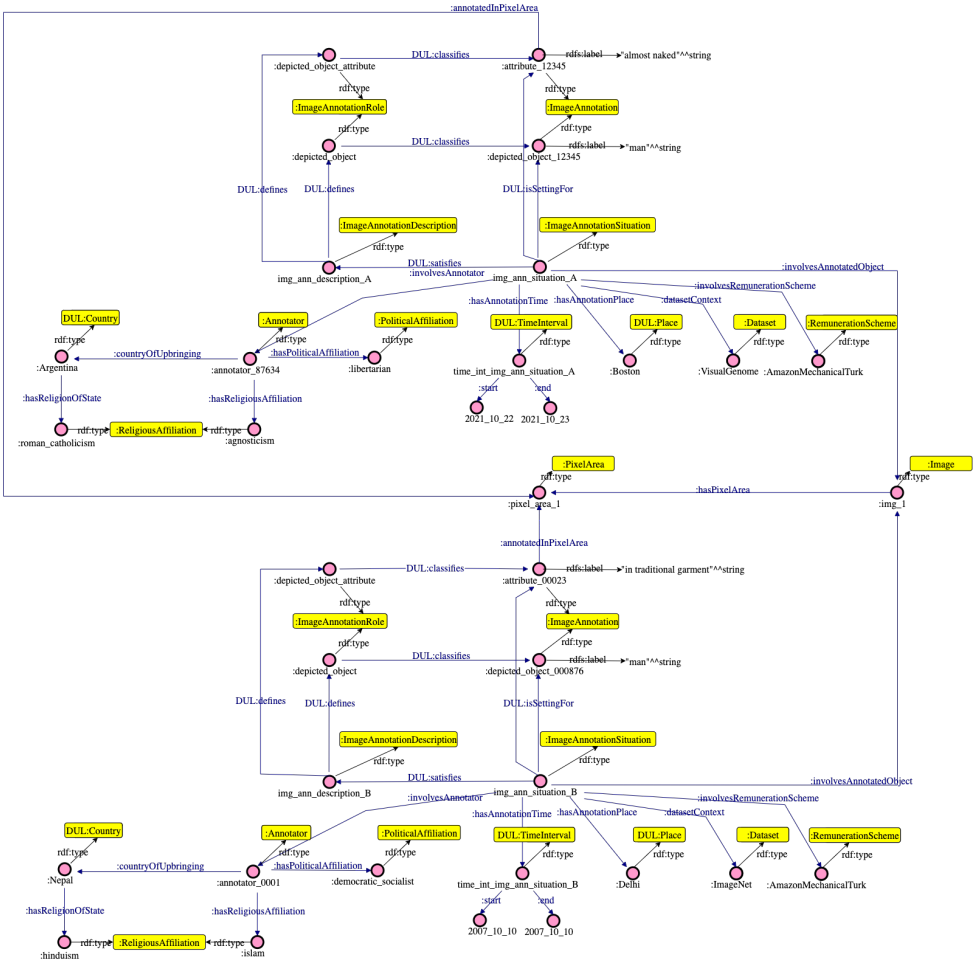

Figure 2: Example use of the IAS module to formally represent a single image being involved in two different `ImageAnnotationSituation`s, which lead to different attributes ("almost naked" and "in traditional garment") to the same area of pixels.All un-prefixed classes belong to the namespace of the IAS ontology.

and how the annotator was remunerated, whether the annotation was done for the creation of a specific dataset, etc. The `ImageAnnotationSituation` class applies the Situation pattern to allow the contextualization of things that have something in common, or are associated: a same place, time, view, causal link, systemic dependence, etc. In the case of IAS, an `ImageAnnotationSituation` provides a context and is the setting for a variety of things that share a same informational space. In Figure 1, IAS and its representation of culturally-coded annotation contexts are shown.

The IAS module not only allows the explicit integration of cultural contextual information regarding image annotation situations, but also provides the infrastructure to compare different image annotation situations connected to the same image object. In other words, an IAS-based knowledge graph (see Figure 2) can be SPARQL queried in order to retrieve all the `ImageAnnotationSituation`s that an image may be involved in, and more specific queries can be performed to compare and better understand the contexts in which (potentially contradictory) interpretations of the same image, or parts of it, were produced.

4 Conclusion

The Image Annotation Situation (IAS) ontology module addresses cultural bias in image annotations systems by providing a graph-based technical representation of the culturally-bound process of annotating images. In doing so, it lays the ground for a more faithful representation of the multilayered process of visual understanding, and of the interpretative situations underlying widely spread computer vision pipelines. It is crucial to continue research on how image classification systems may be echoing, amplifying, or restructuring certain visions of the world, a task requiring an ongoing partnership between humanists and computational scholars.

References

Arnold, T. and L. Tilton (2019). Distant viewing: analyzing large visual corpora. Digital Scholarship in the Humanities 34(Supplement_1), i3–i16.

Bannour, H. and C. Hudelot (2011). Towards ontologies for image interpretation and annotation. In 2011 9th International Workshop on content-based multimedia indexing (CBMI), pp. 211–216. IEEE.

Baroncini, S., M. Daquino, and F. Tomasi (2021). Modelling art interpretation and meaning. a data model for describing iconology and iconography. arXiv preprint arXiv:2106.12967.

Barthes, R. (1981). Camera lucida: Reflections on photography. Macmillan.

Buolamwini, J. and T. Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pp. 77–91. PMLR.

Buolamwini, J. A. (2017). Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Ph. D. thesis, Massachusetts Institute of Technology.

Carboni, N. and L. de Luca (2019). An ontological approach to the description of visual and iconographical representations. Heritage 2(2), 1191–1210.

Daquino, M. and F. Tomasi (2015). Historical context ontology (hico): a conceptual model for describing context information of cultural heritage objects. In Research Conference on Metadata and Semantics Research, pp. 424–436. Springer.

de Vries, T., I. Misra, C. Wang, and L. van der Maaten (2019). Does object recognition work for everyone? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 52–59.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee.

Foster, H. (1988). Preface. vision and visuality. ed. hal foster.

Gangemi, A., M. Alam, L. Asprino, V. Presutti, and D. R. Recupero (2016). Framester: A wide coverage linguistic linked data hub. In European Knowledge Acquisition Workshop, pp. 239–254. Springer.

Gangemi, A., N. Guarino, C. Masolo, A. Oltramari, and L. Schneider (2002). Sweetening ontologies with dolce. In International Conference on Knowledge Engineering and Knowledge Management, pp. 166–181. Springer.

Gangemi, A. and P. Mika (2003). Understanding the semantic web through descriptions and situations. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 689–706. Springer.

Giotta, G. (2020). Ways of seeing… what you want: flexible visuality and image politics in the post-truth era. Fake News: Understanding Media and Misinformation in the Digital Age, 29.

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123(1), 32–73.

Laranjeira, C., V. Fernandes Mota, and J. A. dos Santos (2021). Machine learning bias in computer vision: Why do i have to care? 34th SIBGRAPI conference on graphics, patterns and images.

Liu, H. and P. Singh (2004). Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal 22(4), 211–226.

Llamas, J., P. M. Lerones, E. Zalama, and J. Gómez-García-Bermejo (2016). Applying deep learning techniques to cultural heritage images within the inception project. In Euro-Mediterranean Conference, pp. 25–32. Springer.

Martinez Pandiani, D. S. and V. Presutti (2021). Automatic modeling of social concepts evoked by art images as multimodal frames. arXiv e-prints, arXiv–2110.

Miller, G. A. (1995). Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41.

Sager, C., C. Janiesch, and P. Zschech (2021). A survey of image labelling for computer vision applications. Journal of Business Analytics, 1–20.

Saleh, E. I. (2018). Image embedded metadata in cultural heritage digital collections on the web: An analytical study. Library Hi Tech.

Samih, H., S. Rady, M. A. Ismail, and T. F. Gharib (2021). Semantic graph representation and evaluation for generated image annotations. In International Conference on Advanced Machine Learning Technologies and Applications, pp. 369–384. Springer.

Scheuerman, M. K., J. M. Paul, and J. R. Brubaker (2019). How computers see gender: An evaluation of gender classification in commercial facial analysis services. Proceedings of the ACM on Human-Computer Interaction 3(CSCW), 1–33.

Schumann, C., S. Ricco, U. Prabhu, V. Ferrari, and C. Pantofaru (2021). A step toward more inclusive people annotations for fairness. arXiv preprint arXiv:2105.02317.

Tariq, A. and H. Foroosh (2017). Learning semantics for image annotation. arXiv preprint arXiv:1705.05102.

The American Psychological Association (2015). Cultural bias.

Vacura, M., V. Svátek, C. Saathoff, T. Franz, and R. Troncy (2008). Describing low-level image features using the comm ontology. In 2008 15th IEEE International Conference on Image Processing, pp. 49–52. IEEE.

Wang, X., N. Song, X. Liu, and L. Xu (2021). Data modeling and evaluation of deep semantic annotation for cultural heritage images. Journal of Documentation.