Exploring Naming Inventories for Architectural Elements for Use in Multi-modal Machine Learning Applications

R. Utescher[1,2], A. Patee[3], F. Maiwald[1], J. Bruschke[4], S. Hoppe[3], S. Münster[1], F. Niebling[4], and S. Zarrieß[2]

[1]FSU Jena
[2]Bielefeld University
[3]LMU Munich
[4]JMU Würzburg

1 Abstract

The study of architectural art history has greatly benefited from innovative, computer-aided approaches in recent years. From high resolution two-dimensional (2D) photos of building edifices, to three-dimensional (3D) models of entire structures, these emerging techniques are laying the foundation for new methodologies in researching architecture (Lutteroth and Hoppe, 2018; Sapirstein, 2016). Provided the three-dimensional nature of buildings, research projects have appropriately focused upon techniques that produce digital replicas of their forms, such as Structure-from-Motion (SfM) Photogrammetry and Terrestrial Laser Scanning (TLS) (Lercari, 2016). However, one critical aspect in the study of architecture has largely been dormant since the emergence of these technologies, namely, the computer-aided description of architectural elements. 2D images and 3D models were the logical first steps in devising new methodologies for identifying architectural elements, providing precise calculations of their dimensions and structures, buttressed by the unique capability to virtually study a building. What normally follows is a traditional text description of the discoveries achieved by the employment of these techniques, relegating these digital applications as mere means to a more enlightened end. The lamentable result is a digital purgatory of 3D models awaiting their fate in repositories or online databases.

This paper presents one aspect of a larger project seeking to utilize 2D images and 3D models as essential components of a search engine for architectural elements. These digital objects can serve as reference points for future research in architectural art history and archaeology. What is required, is a systematic identification of the elements themselves using text descriptors, in order that the digital representations of the elements can be efficiently explored. In many ways, this avenue of research is an evolution of Passonneau et al. (2008), which had expert and non-expert annotators chose phrases from longer textual descriptions with which to index paintings in a collection. For this purpose, we implement the methodology of recent research (Wu et al., 2021) as the foundation of the link between text and the digital representation. The work by Wu et al. (2021) represents an important step in integrating 3D models, collections of photographs and written descriptions of historic building using state-of-the-art machine learning methods. Given a multi-modal collection of cathedrals, the model learns to detect and classify 11 kinds of architectural objects, for example *portals* and *columns*. These classes of objects are however limited by which terms terms are frequently associated with images in the original source.

In this paper, we take a closer look at the first two steps in their workflow; (1) curating images and their descriptions from large, open-source collections and (2) selecting the vocabulary of architectural elements for the machine-learning model to classify. We argue that these are important design decisions that have ramifications for the output of the model down the line. Both Wu et al. (2021) and the authors of this paper source their images from Wikimedia Commons. While Commons is a free and abundant source of images, the indexing and naming it provides for individual images is comparatively limited and unsystematic.

Our case study, like Wu et al. (2021), is abstract as it is neither limited to a specific building, nor a design implemented by a specific architect. Rather, it is a collection of baroque monumental buildings largely built between the late 17th and mid-18th centuries, ranging geographically from Portugal to Russia. In effect, we construct a new collection which parallels the one introduced in Wu et al. (2021). The majority of the collection consists of 2D images, though this could be supplemented in future research by existing high-resolution TLS models of historic buildings in the German city of Dresden, such as the iconic Zwinger.

Acknowledgements

This work was supported by a grant from the Federal Ministry of Education and Research (BMBF, grant No. 01UG2120A).

References

Nicola Lercari. 2016. Terrestrial laser scanning in the age of sensing. *Digital methods and remote sensing in archaeology*, pages 3–33.

Jan-Eric Lutteroth and Stephan Hoppe. 2018. Schloss friedrichstein 2.0-von digitalen 3d-modellen und dem spinnen eines semantischen graphen. In *Computing art reader : Einführung in die digitale Kunstgeschichte*, pages 184–198.

Rebecca J Passonneau, Tom Lippincott, Tae Yano, and Judith L Klavans. 2008. Relation between agreement measures on human labeling and machine learning performance: results from an art history domain. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Philip Sapirstein. 2016. Accurate measurement with photogrammetry at large sites. *Journal of Archaeological Science*, 66:137–145.

Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. 2021. Towers of babel: Combining images, language, and 3d geometry for learning multimodal vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 428–437.