The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-2-2023
29th CIPA Symposium "Documenting, Understanding, Preserving Cultural Heritage:
Humanities and Digital Technologies for Shaping the Future", 25–30 June 2023, Florence, Italy

INTRODUCING A MULTIMODAL DATASET FOR THE RESEARCH OF ARCHITECTURAL ELEMENTS

J. Bruschke[a], C. Kröber[b], F. Maiwald[b], R. Utescher[b], A. Pattee[c]

[a]Human-Computer Interaction, Univerität Würzburg, Germany – jonas.bruschke@uni-wuerzburg.de
[b]Digital Humanities, Friedrich-Schiller-Universität Jena, Germany – (cindy.kroeber, ferdinand.maiwald, ronja.utescher)@uni-jena.de
[c]Art History, Ludwig-Maximilians-Universität München, Germany – aaron.pattee@lmu.de

KEY WORDS: Multimodal data, artificial intelligence, computer vision, art history, annotations.

ABSTRACT:

This article looks at approaches, software solutions, standards, workflows, and quality criteria to create a multimodal dataset including images, textual information, and 3D models for a small urban area. The goal is to improve art historical research on architectural elements relying on the three data entities. A specific dataset with manually created annotations is introduced and made available to the public. The paper provides an overview of the available data and detailed information on the preparation of the different types of data as well as the process of connecting everything through annotations. It mentions the relevance and creation of a controlled vocabulary. Furthermore, point cloud processing as well as neural network approaches are discussed which may replace manual labeling. Another focus is the analysis of linguistic similarities to identify whether annotations are actually connected and therefore relevant. Additionally, research scenarios will highlight the relevance of the approach for art history and the contributions, which come from computer linguistics and computer science.

1. INTRODUCTION

Photographs and other images as well as textual information and documents serve as important source materials and provide a foundation for many subject-orientated and theory-based investigations within historical studies, e.g., architectural studies, art history, and cultural studies. Among other scenarios, the sources may be used for (digital) reconstructions or to investigate different buildings, their construction history, and the impact they had on a city. A lot of sources are needed to get a thorough understanding of a building. Searching for images and texts but also the contextualization and the evaluation of them can prove challenging (Beaudoin and Brady, 2011). On the one hand, there is a vast amount of data available online and, on the other hand, filtering is usually not satisfying or avoided altogether. Additionally, 3D data is increasingly used while researching sites with an archaeological and art-historical background (Münster, 2016).

The project HistKI is working on a solution that connects different data in a research platform and, therefore, supports the benefit of multimodal data for investigation (Münster et al., 2022). Usually, different data types are researched and analyzed separately, and, in a next step, the researcher will connect everything and use the entirety of the gathered knowledge for context and discovery. Additionally, scholars will elaborate on source criticism in order to evaluate and verify data and findings. New technologies like artificial intelligence (AI) can be of help when dealing with the variety of data.

1.1 The Research Platform 4D Browser

A prototypical web application called 4D Browser[1] currently presents historical images of the city of Dresden in a virtual 3D city model (cf. Fig. 1). A timeline introduces the fourth dimension (4D) and provides information on the development of the city by filtering photos and 3D models according to any selected point in time. The features support image searching and filtering, data analysis, data visualization based on acquisition habits, and contextualization via linking photos and their spatio-temporal location within the model (Dewitz et al., 2019). Textual information will be available soon.

Figure 1. Interface of the 4D Browser showing spatio-temporal photographs in a 3D city model.

The different data entities text, image and 3D model will be linked in order to support investigation and connections. When researching architectural elements of a building, the user is supposed to query and find data entities that refer to the same or similar elements. Therefore, annotating specific architectural elements within all three types of data entities (text, 2D images, and 3D models) is necessary to build a query system which draws implicit links between the entities based upon the annotations.

Additional tools and functionalities on top of the annotations aim to support further research. For implementation and testing purposes, we introduce a multimodal dataset that contains texts, images, and 3D models that has been annotated for architectural elements of a small urban area. Requirements for data preparation will be considered as well as the workflow of integration

[1] https://4dbrowser.urbanhistory4d.org/

into the platform and the possibilities of novel functionalities for investigation.

1.2 Significance of the Zwinger for Art History Research

For the purpose of the project and the sake of efficiency, we selected a monumental palatial building from which we already own a detailed 3D model. The site is the Zwinger and its associated buildings located near the royal residence in Dresden, Germany. The site was commissioned by King August II the Strong of Saxony and Poland during the Baroque period during the last decades of the 17th century and the first decades of the 18th century. It is a peculiar design featuring four buildings, or Pavilions, bound together in the shape of a large rectangle with a main axis orientated from Northwest to Southeast. However, the most iconic element is undeniably the monumental gate along the southwestern side featuring a large crown atop a crown-shaped dome, aptly named the Kronentor (Fig. 2).

Figure 2. Historical photograph of the Dresden Zwinger, Germany taken before 1900. Photographer unknown. © Deutsche Fotothek

All of the buildings are highly decorated with ornate pilasters, columns, and statues. The area within these pavilions and their associated connections is uncovered, as it previously served as a parade ground for state events such as royal weddings or military demonstrations. The name of the site is also significant as a Zwinger once referenced an outer wall of a castle, though in this case, the outer wall of the residence had been completely repurposed to accommodate a large palatial complex in the relatively small environment to the southwest of the royal palace. The Zwinger is bounded along its western side by water that used to serve as the moat of the outer wall, though the area had been turned into a large garden during the remodeling (Fig. 3).

Besides it visual appeal, the Zwinger is also one of the most iconic Baroque buildings in the world, featuring a complex history of destruction and reconstruction. There is also a large collection of (historical) photographs from the Saxon State and University Library Dresden (SLUB) available with some images already geo-referenced within the 3D city model.

2. PREPARATION OF THE DATASET

2.1 Using Controlled Vocabularies

The desire to actively interconnect the data in order to support new insights and discover certain phenomena drove this project. Architectural elements of the same type are encoded differently in the three types of data entities, i.e., as words in texts, as group of pixels in images, and as group of vertices and faces in 3D

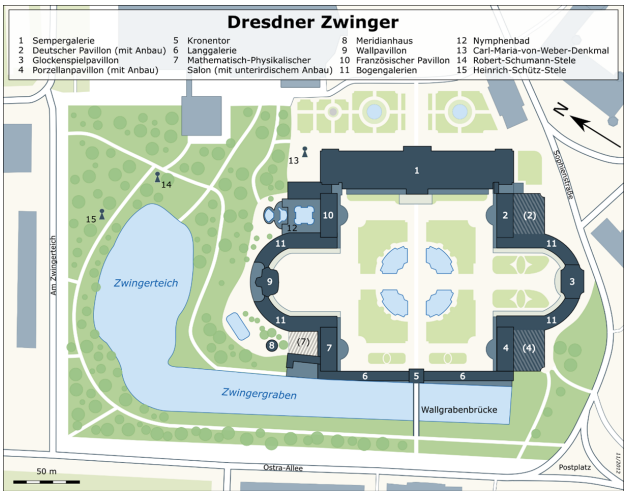

Figure 3. Map of the Zwinger in Dresden with labels. Source: `de.m.wikipedia.org/wiki/Datei:Karte-Zwinger-Dresden.png`

models. To this end, we incorporate controlled vocabularies that define language-agnostic identifiers for specific elements and are used to classify the annotations of the data entities.

In the domain of digital humanities, the Getty Art & Architecture Thesaurus (AAT) is a vocabulary that contains an extensive, hierarchical set of classes of architectural elements (Baca and Gill, 2015). Other Getty vocabularies are the Union List of Artist Names (ULAN), the Cultural Objects Names Authority (CONA), and the Thesaurus of Geographic Names (TGN). However, most of their classes might only be encoded in text and are, hence, only of little interest when researching architectural elements. Another huge knowledge base is Wikidata that defines a large variety of entities and classes including art and architectural elements (Vrandečić and Krötzsch, 2014). Beyond the hierarchical structure similar to the AAT, semantic relationships between the entities are also stored forming an ontology. This involves links to other knowledge bases including the Getty vocabularies. Wikidata is also increasingly used in the cultural heritage domain (Schmidt et al., 2022).

The number of classes of architectural elements in both the AAT and Wikidata is very large. However, only a subset is primarily used in texts about an architectural object. To narrow down the set to what is actually used, the first step is to identify those architectural terms by analyzing texts describing the Zwinger and related architecture (Utescher et al., 2022). This requires foreknowledge of art-historical terminology and experience pertaining to construction history and research (Baugeschichte and Bauforschung). Additionally, it requires a series of software to first make text readable and then to analyze the text in order to gather relevant terminologies together with their respective normative data. The AAT’s and Wikidata’s entity ids were tested to see how well they align with the textual documents of the project. Some relevant terms are currently missing on either one or both lists and need to be added. However, the majority and most relevant terms are included.

The analysis of the textual documents yielded a condensed list of vocabulary entries with, if possible, both AAT and Wikidata identifiers that is used for annotating our data. It comprises about 400 entries, from which about 140 are architectural elements and are of higher relevance.

2.2 Textual Resources

With respect to its application in the 4D, annotated text data is one of the areas that is currently under development. The methodology for annotating the text included in this dataset is geared towards being simple but robust.

Data Selection For quantitative analysis, we collected 321 Wikipedia articles. We use all Wikipedia articles which are linked on the German-language Zwinger, Dresden article. These texts cover parts of the Zwinger complex, important personalities connected to the Zwinger as well as articles about the Baroque or specific architectural elements. In addition, we include two scientific publications about the Zwinger in order to test our methodology on longer, more complex texts.

Annotation Protocol In the textual annotations, we catalog mentions of all terms in our vocabulary as well as terms that have a high semantic similarity. This semantic similarity measure is computed at word-level using German or English fasttext embeddings (Bojanowski et al., 2016). It is possible for a token to have more than one label. For each term-annotation, we provide a semantic similarity score $s \in [0, 1]$. We suggest a division of these annotations into three groups: (1) identical matches, (2) very similar terms, and (3) loosely related terms.

Identical matches ($s = 1$) are instances of a term from the list, including capitalization and morphology such as plurals. We consider terms with a semantic similarity > 0.8 to be near-identical in meaning, differentiating them from the identical and loose matches. Terms with $0.8 > s > 0.6$ also are included in the annotations. These loose matches are semantically similar to the core term, which we include in order to flesh out exploration of the data. Although the vocabulary is limited to nouns and proper nouns, the word search searches all parts of speech. For example, a search for the term “Gebäude” (building) will yield “Haus” (house), but also “abgerissen” (torn down) and “erbaut” (built).

Besides their bibliographical information, the text annotations contain exact matches of the terms in our term list as well as annotations of semantically similar terms. We include a plain text file, an annotation JSON and a BibTex file for each text in the dataset. The annotation file contains a list of all annotations for each text file. The BibTex file contains metadata about the text itself including title, year of publication, author(s), but also the associated object/building (e.g. Nymphenbad).

Text Annotation Overview For the 107 terms in the vocabulary and 321 Wikipedia articles, we were able to assign 89k labels. The Wikipedia articles form the larger and more homogeneous part of our textual resources. Furthermore, both the German and English fasttext embeddings were trained on Wikipedia as a whole. Out of the 89k, 7434 instances are exact matches, 2286 are close matches and 79.3k are loose matches. This semantic similarity annotation leads to a larger vocabulary of 3973 terms (including the original 107).

Note that close matches do not denote that the original term and the related term are synonymous. For example, a search for “Zierbrunnen” yields the close matches “Marktbrunnen”, “Prachtbrunnen”, “Rathausbrunnen” and “Springbrunnen”. Incidentally, its loose matches contain various other kinds of fountain as well as terms such as water spout.

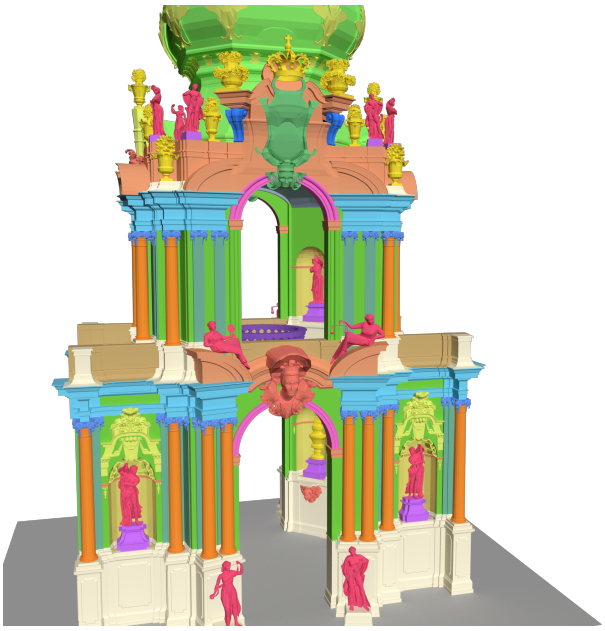

Figure 4. Annotated 3D model of the Kronentor of the Zwinger.

2.3 The 3D Model

3D models of buildings can be created in various ways: Existing buildings can be digitalized by terrestrial laser-scanning (TLS) or Structure-from-Motion (SfM). In contrast, lost or never-built (but also existing) architecture can be reconstructed from sources, e.g., plans, blueprints, and images. This reconstruction process involves an interpretation and thorough understanding of the source material, which may lead to new knowledge about the object that would otherwise have remained hidden (Echtenacher, 2011). A highly detailed 3D model of the Zwinger is available that is the outcome of a reconstruction project with the focus on its building history.

Considering ways to annotate the 3D model, there are approaches available which are able to label the architectural elements in a 3D point cloud (Morbidoni et al., 2020, Croce et al., 2021). In theory, mesh-based 3D models, like the available Zwinger model, can be sampled into point clouds in order to perform this labeling process. However, these approaches feature only a limited number of labels, i.e., architectural elements, and the training dataset, if publicly available, focuses on different architectural style. Whereas in baroque architecture as used in this example, other architectural elements are more predominant.

In this regard, the high-poly 3D model of the Zwinger was manually labeled (Fig. 4) in a 3D modeling environment and added to our dataset in order to start implementing and testing interlinking annotations. In this model, most of the architectural elements are separate objects that were labeled with respective AAT or Wikidata ids following the pattern `<englishName>_wd:<wikidataId>_aat:<aatId>`, which allows parsing the identifiers. Some objects are grouped to form a more general entity resulting in a hierarchy of elements. In the dataset, the 3D model is provided as FBX and XML-based Collada file.

With regard to the 4D Browser, the 3D model is too extensive that it can be displayed performantly in the 3D web interface.

Instead, a low-poly dummy object is used for display and navigation while the high-poly model remains in the backend for computation.

2.4 Images of the Zwinger

As the Zwinger in Dresden was already present at the beginning of the era of photography around 1850, a lot of historical but also contemporary images exist in various archives. The undertaken research focuses on ca. 10,000 historical images provided by the Saxon State and University Library Dresden (SLUB) and approximately 6,000 contemporary images from Wikimedia Commons. If these images have been spatially aligned to a segmented and annotated 3D model, these annotations could be automatically transferred from 3D model to image and vice versa by projection (Manuel et al., 2013). 3D models and images could complement each other (cf. Fig. 5).

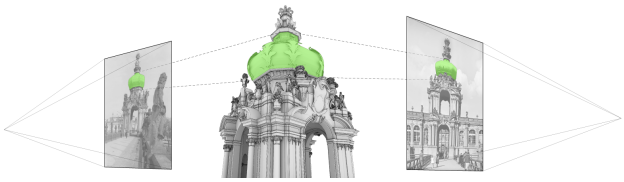

Figure 5. Projection of annotations between 3D model and spatially oriented images.

A fraction of the historical images is already oriented using the workflow described in (Maiwald, 2022) and (Maiwald et al., 2023). That means that the camera orientation parameters are estimated, and the location of the original photograph is determined in a global scale. As this is not an easy task for historical photographs with large geometric and radiometric differences, the pipeline uses a combination of Vanishing Point Detection for estimation of the principal distances (Maiwald and Maas, 2021), SuperPoint (DeTone et al., 2018) for feature detection, and SuperGlue for feature matching (Sarlin et al., 2020). Additionally for historical reconstructions, it is recommended to refine all derived keypoints, camera poses, and 3D points using Pixel-Perfect Structure-from-Motion (Lindenberger et al., 2021). The resulting model of approximately 200 historical photographs is depicted in Fig. 6.

For images that cannot be oriented, three alternative ways of segmentation of building parts are proposed. The first approach

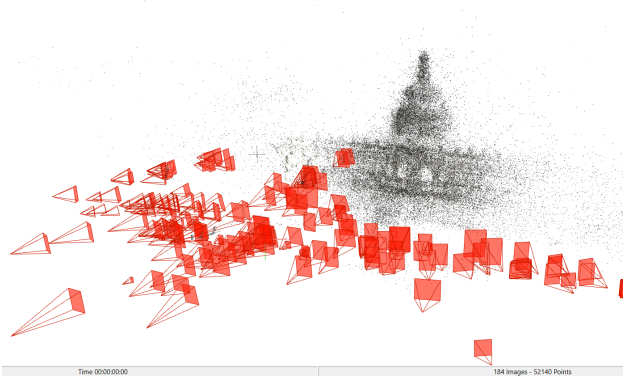

Figure 6. Structure-from-Motion reconstruction of the Kronentor of the Dresden Zwinger including the camera orientation of 184 historical images and a sparse point cloud.

uses images showing only specific building parts as query images. A convolutional neural network (CNN) based on VGG16 (Simonyan and Zisserman, 2015) searches in all provided images for these building parts and a frame is drawn around the image part that supposedly shows the respective building part of the query. For an effective search, the image is divided in smaller sub-images (Razavian et al., 2016). However, while the idea of re-using this neural network is convenient, the final bounding boxes are not congruent with the perfectly segmented building part and also sometimes the results are erroneous (Fig. 7).

Figure 7. Left side of the arrow shows the query image depicting the top of a tower. Right of the arrow are the best results for image parts found by the neural network in the archives. The right image shows a bounding box where mainly the sky is detected.

In order to achieve pixel-wise semantic segmentation, the second approach uses a residual neural network (He et al., 2016). Therefore, a ResNet50 is implemented that has been trained on contemporary images of facades using the well segmented eTRIMS Image Database with 8 classes (Korč and Förstner, 2009). As the dataset consists only of 60 images, the absolute number is enhanced using radiometric and geometric methods of augmentation. This is especially required for transformation of the contemporary data to historical facade images. As a result for most images, the sky is segmented quit well. However, pixel-perfect segmentation of windows, towers, statues, or other elements could also not be achieved using this method (Fig. 8).

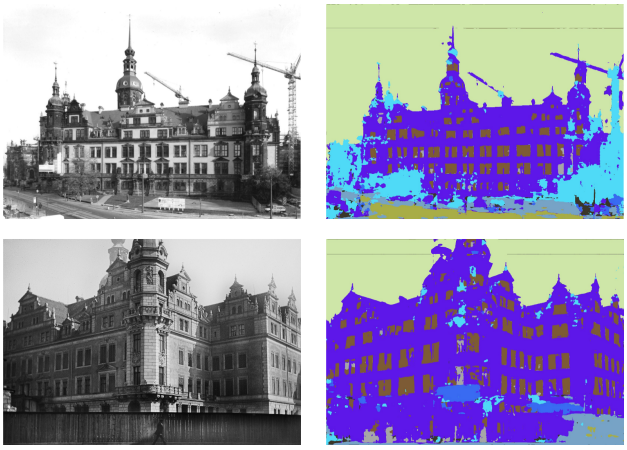

Figure 8. Two examples for the segmentation of historical facade images using a ResNet50.

The third approach is at the moment at a conceptual stage and uses the recently released Segment Anything (SAM) by Meta (Kirillov et al., 2023). For comparison, a segmentation without further modifications has been performed on the same images (Fig. 9). The advantage is that SAM is not initially restricted by classes and every object is treated individually. This allows a direct visual control of the results. As already seen in the ResNet50 approach, not every window is detected by SAM.

The results of the three different strategies presented show that

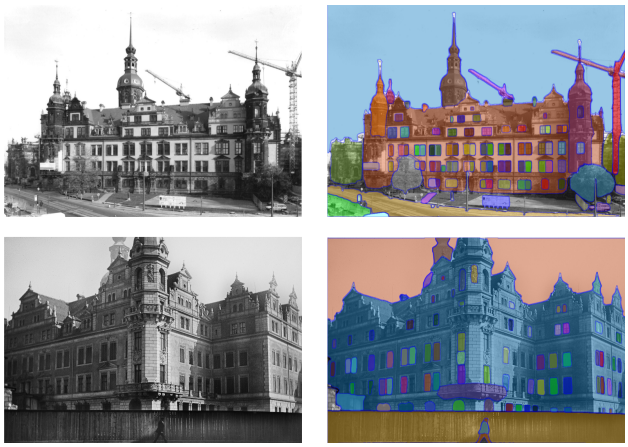

Figure 9. Two examples for the segmentation of historical facade images using Meta’s Segment Anything (SAM).

there is still a need for improvement on the segmentation of historical terrestrial images. The biggest challenge is that there does not exist a training dataset that contains the very domain specific labels of relevant architectural elements, but is however needed to meet the research scenarios from art and architectural history. For this purpose and following the reasons as for the 3D model, a set of images has been manually segmented and labeled using Label Studio[2], a multi-type data labeling and annotation tool with standardized output format (Fig. 10). Each annotation follows the same naming convention as for the 3D models incorporating both AAT and Wikidata ids. In the dataset, the images are provided together with a JSON file following the common COCO dataset segmentation format (Lin et al., 2015).

Figure 10. Manually annotated image of the Kronentor.

3. USE OF THE DATASET

The dataset is primarily used for implementing and testing tools to support art historical research. Currently, the database of the 4D Browser stores texts, 2D images, and 3D models as single instances with their metadata and, if available, their spatial information. Some relations are also stored, e.g., if it’s known that a building, resp. 3D model, is depicted in an image. Introducing annotations allows a more thorough description of a data entry leading, however, to a more complex database model. Hence, when adding annotations to an entry, information about the location and content of the segment as well as links to the identifiers of the vocabularies need to be stored too.

The main objective is to provide features that enable users to query similar data entities starting with an item of interest or connected annotations. There are several scenarios in art historical research when using interlinked data would be beneficial, e.g., when investigating architectural elements like a statue in a certain urban setting. In this scenario the focus and interest are on how the statue was created, what is depicted and how the piece of art was perceived or affected the surroundings. The 3D city model can be a starting point to locate the desired statue. The annotations provide the necessary connection to other data entities like images that give more information on the appearance. Even though it is already possible to use deep learning approaches to identify characters (Madhu et al., 2019), for scholars it is crucial to study textual information in order to verify assumptions about who is depicted and why. The relevance based on similarity matching offers a chance to keep search results low but still discover all relevant data. A different approach is to start with written accounts and track any mentions of relevant terms or location and though that discover connected images and even other textual resources. However started, a search for an architectural element will lead to insightful accounts on architects, clients, styles, and city history. In order to find those relevant data, an additional contextual analysis of each annotation is required that is eventually used as a base for similarity matching of the annotations and thus the data entries. Workflows on how to integrate this feature on a technical, but also user-interactive scale still need to be assessed. For example, users need to be made aware of which annotations are generated automatically in a way that is accessible without in-depth knowledge of the computational models. The text annotations require a nuanced presentation to users of the 4D Browser. Close and identical matches follow the general intuition of a regular search. The more loosely related terms, while presumably less transparent to users, generate many potential links in the dataset for users to explore.

4. CONCLUSION

During the research connected to multimodal data and their access and use for art historical research, it became evident that there have not been efforts so far to interconnect the three entities of images, texts, and 3D models to research architecture. For testing, evaluating, and verification purposes, it became obvious that a well-matched dataset is essential. The labeling of several images as well as the 3D model while ensuring to adhere to the controlled vocabulary and not miss relevant entries was a tedious undertaking. Since it was done by several different editors, quality control was crucial. The progress in AI and deep learning will hopefully lead to possibilities to automatically identify architectural objects or elements in images to the necessary extent and with the necessary reliability – ideally in the foreseeable future. Having spatially aligned images and the related 3D models will hopefully allow for an automatic segmentation of 3D model reducing human interventions to a minimum. Although the dataset is used as a test set for striving forward the development of a specific research platform, it may be of interest to researcher in other domains.

[2] https://labelstud.io/

At the time of publication, the current dataset contains only a 3D model and images of the Kronentor of the Zwinger, Dresden. We plan to extend the dataset by 3D models and images of other building parts of the Zwinger as well as other buildings. The annotated textural resources, which are currently mainly Wikipedia articles, will be extended by art historical papers. The dataset is available at `https://github.com/tudipffmgt/ArchiLabel`.

ACKNOWLEDGMENTS

The research upon which this paper is based is part of the research project "HistKI" which has received funding from the German Federal Ministry of Education and Research (BMBF), grant identifier 01UG2120.

REFERENCES

Baca, M., Gill, M., 2015. Encoding multilingual knowledge systems in the digital age: the Getty vocabularies. R. P. Smiraglia (ed.), *Proceedings from North American Symposium on Knowledge Organization*, 5, 41–63.

Beaudoin, J. E., Brady, J. E., 2011. Finding Visual Information: A Study of Image Resources Used by Archaeologists, Architects, Art Historians, and Artists. *Art Documentation: Journal of the Art Libraries Society of North America*, 30(2), 24–36. doi.org/10.1086/adx.30.2.41244062.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Croce, V., Caroti, G., De Luca, L., Jacquot, K., Piemonte, A., Véron, P., 2021. From the Semantic Point Cloud to Heritage-Building Information Modeling: A Semiautomatic Approach Exploiting Machine Learning. *Remote Sensing*, 13(3), 461. doi.org/10.3390/rs13030461.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-supervised interest point detection and description. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 337–33712.

Dewitz, L., Kröber, C., Messemer, H., Maiwald, F., Münster, S., Bruschke, J., Niebling, F., 2019. Historical Photos and Visualizations: Potential for Research. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W15, 405–412. doi.org/10.5194/isprs-archives-XLII-2-W15-405-2019.

Echtenacher, G., 2011. Wissenschaftliche erkenntnisse durch manuelles konstruieren von 3d-modellen. K. Heine, K. Rheidt, F. Henze, A. Riedel (eds), *Von Handaufmaß bis High Tech III*, Verlag Philipp von Zabern, Darmstadt/Mainz, 49–57.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. doi.org/10.48550/arXiv.2304.02643.

Korč, F., Förstner, W., 2009. eTRIMS image database for interpreting images of man-made scenes. Report TR-IGG-P-2009-01, Dept. of Photogrammetry, University of Bonn.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., Dollár, P., 2015. Microsoft COCO: Common objects in context. doi.org/10.48550/arXiv.1405.0312.

Lindenberger, P., Sarlin, P.-E., Larsson, V., Pollefeys, M., 2021. Pixel-perfect structure-from-motion with featuremetric refinement. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 5987–5997.

Madhu, P., Kosti, R., Mührenberg, L., Bell, P., Maier, A., Christlein, V., 2019. Recognizing characters in art history using deep learning. SUMAC '19, Association for Computing Machinery, New York, NY, USA, 15–22.

Maiwald, F., 2022. A window to the past through modern urban environments – developing a photogrammetric workflow for the orientation parameter estimation of historical images. PhD Thesis, TU Dresden. https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-810852.

Maiwald, F., Bruschke, J., Schneider, D., Wacker, M., Niebling, F., 2023. Giving Historical Photographs a New Perspective: Introducing Camera Orientation Parameters as New Metadata in a Large-Scale 4D Application. *Remote Sensing*, 15(7), 1879. doi.org/10.3390/rs15071879.

Maiwald, F., Maas, H.-G., 2021. An automatic workflow for orientation of historical images with large radiometric and geometric differences. *The Photogrammetric Record*, 36(174), 77–103. doi.org/10.1111/phor.12363.

Manuel, A., Gattet, E., De Luca, L., Véron, P., 2013. An approach for precise 2D/3D semantic annotation of spacially-oriented images for in-situ visualization applications. *Digtal Heritage International Congress*.

Morbidoni, C., Pierdicca, R., Paolanti, M., Quattrini, R., Mammoli, R., 2020. Learning from Synthetic Point Cloud Data for Historical Buildings Semantic Segmentation. *J. Comput. Cult. Herit.*, 13(4), Article 34. doi.org/10.1145/3409262.

Münster, S., 2016. *Interdisziplinäre Kooperation bei der Erstellung geschichtswissenschaftlicher 3D-Modelle*. Springer VS, Wiesbaden.

Münster, S., Bruschke, J., Kröber, C., Hoppe, S., Maiwald, F., Niebling, F., Pattee, A., Utescher, R., Zarriess, S., 2022. Multimodale ki zur unterstützung geschichtswissenschaftlicher quellenkritik – ein forschungsaufriss. *DHd2022: Kulturen des digitalen Gedächtnisses*, 179–182.

Razavian, A. S., Sullivan, J., Carlsson, S., Maki, A., 2016. Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications*, 4(3), 251–258. doi.org/10.3169/mta.4.251.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning feature matching with graph neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4938–4947.

Schmidt, S. C., Thiery, F., Trognitz, M., 2022. Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata. *Digital*, 2(3), 333–364.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. doi.org/10.48550/arXiv.1409.1556.

Utescher, R., Pattee, A., Maiwald, F., Bruschke, J., Hoppe, S., Münster, S., Niebling, F., Zarrieß, S., 2022. Exploring naming inventories for architectural elements for use in multi-modal machine learning applications. *Workshop on Computational Methods in the Humanities*.

Vrandečić, D., Krötzsch, M., 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10), 78–85. doi.org/10.1145/2629489.