Computer vision and architectural history at eye level

Mixed methods for linking research in the humanities and in information technology (ArchiMediaL)

Mager, Tino; Khademi, Seyran; Siebes, Ronald; van Gemert, Jan; de Boer, Victor; Löffler, Beate; Hein, Carola

DOI
10.1515/9783839469132-014

Publication date
2023

Document Version
Final published version

Published in
Mixing Methods

Computer Vision and Architectural History at Eye Level: Mixed Methods for Linking Research in the Humanities and in Information Technology (ArchiMediaL)

Tino Mager, Seyran Khademi, Ronald Siebes, Jan van Gemert, Victor de Boer, Beate Löffler, Carola Hein

Abstract *Information on the history of architecture is embedded in our daily surroundings, in vernacular and heritage buildings and in physical objects, photographs and plans. Historians study these tangible and intangible artefacts and the communities that built and used them. Thus valuable insights are gained into the past and the present as they also provide a foundation for designing the future. Given that our understanding of the past is limited by the inadequate availability of data, the article demonstrates that advanced computer tools can help gain more and well-linked data from the past. Computer vision can make a decisive contribution to the identification of image content in historical photographs. This application is particularly interesting for architectural history, where visual sources play an essential role in understanding the built environment of the past, yet lack of reliable metadata often hinders the use of materials. The automated recognition contributes to making a variety of image sources usable for research.*

Introduction

Architectural history is a discipline dedicated to the analysis of the built environment of the past. By examining historical buildings and sources of reference such as relics, texts and images, architectural history seeks, among other things, to uncover the conditions and characteristics of the structures of olden times. The buildings themselves, both surviving and lost, and the textual and visual sources that document their production, decay, and use are primary sources. Computer vision offers a potential to expand the range of classical approaches to extracting knowledge from

these sources.[1] While the automatic analysis of text and the search for key terms are well advanced, the automated recognition of image content is an area in which significant progress has only recently been made to such an extent that it appears applicable to visual archive material as well.[2]

Architectural production and also the documentation of the built environment have left a wealth of visual material since the mid-nineteenth century. In particular, photographs collected in public and private repositories have largely remained unexplored. Digitization is a necessary step to facilitate access to this enormous stock of visual material and digital cataloguing is essential to provide comprehensive and efficient information about the existence and accessibility of the material. Digital catalogues such as *Europeana*, the *German Digital Library* or the *Digital Public Library of America* provide access to millions of digitized documents; and architectural history repositories, such as the *Repository of Colonial Architecture and City Planning*[3], contain further collections of visual resources, including historical images of architecture and urban planning. Their digital availability is a decisive aid for research, as they can be viewed and analysed without much effort. Besides, they enable conclusions as to whether it is relevant to inspect the original.

Digitization and digital tools help architectural historians go beyond their traditional, usually limited visual material—archival documents, physical collections or books. As they dig into a big new set of imagery—electronic repositories, crowdsourcing or web-scale datasets—they need to refine their theories and methods. The handling of huge and unfamiliar datasets may throw up questions that go beyond the traditional hermeneutic reading of text and images. They must understand code as a cultural practice and learn to see qualitative data as the result of abstract 'technocratic' sorting that relies on established interpretation systems. Innovation in computer technology, both in crowdsourcing and in AI, creates opportunities and challenges for urban and architectural history, notably the recognition of visuals in vast archives. Crowdsourcing metadata for historical images is an issue closely related to those of communication, mediatization and urban future.

Metadata is required to navigate the visual contents of the repositories. They enable one to search, sort or filter a corpus that is non-verbal. Therefore, the names of

---

1 Beate Löffler, Carola Hein and Tino Mager. "Searching for Meiji-Tokyo: Heterogeneous Visual Media and the Turn to Global Urban History, Digitalization, and Deep Learning", *Global Urban History* (20.03.2018). https://globalurbanhistory.com/2018/03/20/searching-for-meiji-tokyo-heterogeneous-visual-media-and-the-turn-to-global-urban-history-digitalization-and-deep-learning/.

2 Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems* 2 (01.12.2012): 1097–1105.

3 TU Delft, University Library: Colonial architecture & town planning, http://colonialarchitecture.eu.

objects, places and persons related to the image contents as well as keywords, etc., are essential annotations that help find specific visual sources in repositories and relate collection holdings all around the globe. This metadata needs to be assigned to the images systematically and correctly. Defective, wrong or missing annotations mean that the image material may not be found at all and is lost for research unless someone corrects them. Unfortunately, however, historical image collections in particular often have incomplete metadata.[4] Researchers or collection experts can hardly take on the task of manual annotation as due to the abundance of the material they could only handle a tiny part of the media concerned. Even larger teams would not be able to maintain a relevant amount of material. Computer vision can make a significant contribution here. It can help humanities scholars to manoeuvre through the wealth of digitally available visual material by recognising its content, in our case historical buildings. However, artificial intelligence (AI)-based computer vision systems need precise training to be able to recognize and identify specific image content. Therefore, it is necessary to build high-quality datasets that are used for training the algorithms and for evaluating their performance. At this intersection between information technology and humanities research, crowdsourcing can be used as a connecting method to bring together knowledge in the field of architectural history with specific data demands relevant for AI research.[5] The article outlines the application of crowdsourcing in the research project *ArchiMediaL*, which is dedicated to the identification of buildings in historical photographs by using computer vision.[6]

Use of computers for the analysis of historical urban images

The research project ArchiMediaL explores the possibilities of using current information technologies to open up architectural and urban image repositories for research. It develops strategies for the automatic recognition of historic image content through computer vision. Recent advances made in data-driven computer vision have improved the ability of visual intelligent systems to infer complex semantic

4 Beate Löffler and Tino Mager. "Minor politics, major consequences—Epistemic challenges of metadata and the contribution of image recognition", *Digital Culture & Society* 6, 2 (2021): 221–238.

5 Johan Oomen and Lora Aroyo. "Crowdsourcing in the cultural heritage domain: opportunities and challenges", *Proceedings of the 5th International Conference on Communities and Technologies* (2011): 138–149.

6 Seyran Khademi, Tino Mager, Ronald Siebes, Carola Hein, Beate Löffler, Jan van Gemert, Victor de Boer and Dirk Schubert, *Research project ArchiMediaL – Enriching and linking historical architectural and urban image collections* (TU Delft, VU Amsterdam, TU Dortmund, HafenCity University Hamburg). https://archimedial.eu.

relationships. The performance of modern computer vision requires a large number of images that have already been annotated. It takes millions of annotated images with hundreds of thousands of object classes to teach a computer 'to see the world'. Such images are abundant in the digital age, but training AI to see the past is more complex. To teach computers to recognize architecture in historic photographs, they must first be shown what the world looked like before the advent of digital cameras, otherwise they will not be able to recognize and detect objects and semantics of the past.

Again, to teach computers about the past, a large number of images are needed for training so that they can do the desired visual task, for example object recognition. The brain-inspired computer systems, referred to as convolutional neural networks (CNNs),[7] extract effective features in the form of tensor representation by seeing.[8] These trained models are later used for inference on visual data. Ideally, CNN models learn general rules during the training process, so that when they apply those rules to data they have not seen before, they can produce accurate classifications. Such inductive learning is consistent with human intelligence where the learned skills are used in real-world scenarios which might not have been fully covered in the training period. The more correlated the training and the test scenarios are, the more effective the learning is, for both the human and the machine intelligence. Accordingly, once a CNN model is trained on high-quality natural contemporary images for visual information retrieval, it cannot effectively perform the task on illustrations, low-quality and blurred images, drawings or paintings that are often found in archival records. In short, to develop intelligent tools that can handle the latter, we need to train the CNN models for archival image collections or else we end up creating incompetent tools.[9]

While Machine Learning based computer vision is finding its applications in the humanities domain,[10] the application of computer vision for architectural history and targeting historic photographs have not been pursued so far.[11] For facilitating the opening up of large image sets, research must start in a fairly well-documented

---

7 Krizhevsky, Sutskever, Hinton, "ImageNet".

8 Yoshua Bengio, Aaron C. Courville and Pascal Vincent. "Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives", *CoRR*, abs/1206.5538 (24.06.2012): 1–30.

9 cf. Melvin Wevers and Thomas Smits, "The Visual Digital Turn: Using Neural Networks to Study Historical Images", *Digital Scholarship in the Humanities* 35, 1 (April 2020): 194–207, https://doi.org/10.1093/llc/fqy085.

10 Cf. Melvin Wevers and Thomas Smits. "The Visual Digital Turn: Using Neural Networks to Study Historical Images." *Digital Scholarship in the Humanities* 35, 1 (April 2020): 194–207, https://doi.org/10.1093/llc/fqy085.

11 A similarity can hitherto only be found in the Urban Panorama research project: North Carolina State University: Urban panorama. https://www.visualnarrative.ncsu.edu/projects/urban-panorama/.

area of architectural and urban history, since the performance and reliability of the algorithm can only be tested if the topic to which it is applied is known. To enable a high recognition rate, the objects that we study, in this case the built environment, should not have changed too much compared to the situation captured in the images, and the images must provide sufficient metadata to enable the results to be verified. To meet this requirement, a database of images of a well-studied location with a sufficient number of metadata is needed. We found it in the *Beeldbank* repository of the Amsterdam City Archives—the world's largest city archive. The Beeldbank contains several hundred thousand images taken in the streets of Amsterdam since the nineteenth century, among them many are images of facades, buildings and streets.[12]

In order to identify the buildings in the historic photographs with the help of computers, one can compare the image content with a large number of current photographs of Amsterdam that contain geo-information. Such a repository can be found in online mapping services such as *Mapillary*.[13] In Mapillary, a large part of Amsterdam's buildings is captured and provided with address data. If a building in a historical photo can be identified through similarity with a building in a Mapillary photo, the location of the building in the historical photo and thus the building itself are identified. The core of the project is, therefore, to use AI to automatically identify buildings in historical photographs that are geolocated in the Mapillary repository. The training of state-of-the-art AI models on available historical image data repositories can effectively help computers to become more intelligent and expert in the domain of historical data; in return, researchers and librarians can make use of these models to interact optimally with the visual archives.

This mutual interaction between computer scientists and humanity researchers breaks the classical pattern of computer science *serving* other fields without a real reciprocal conversation between the two parties. The two disciplines can effectively pursue research at eye-level, thereby producing new results in architectural history as a field of the humanities, and in computer science by making a step towards interpretable AI. However, training the AI models requires reliable data. In this case it means image pairs of the same building, once on a historical photo and once on a Mapillary photo. With a large number of such image pairs, the algorithm must be trained to be able to assign one of the many Mapillary images to a historical photo. The street address of the historical building can then be determined by the geolocation of the Mapillary photo. Having said this, these image pairs must be created by humans and they must be reliable. The involvement of many people is helpful for this purpose, as in this way a larger data set (>1000 image pairs) can be created relatively easily.

12 Beeldbank Stadsarchief Amsterdam. https://archief.amsterdam/beeldbank/.

13 Mapillary. http://mapillary.com, 2020 [accessed: 04.06.2020].

Crowdsourcing image pairs

High-quality data sets with images and reliable metadata are required to train and evaluate the AI systems. Crowdsourcing is a key factor in the creation of these. Virtually at the interface between social science research and information technology support, crowdsourcing serves to create the necessary data sets in a time- and cost-efficient way. One of the first Web2.0 crowdsourcing successes, where the value of content is created through the self-organised collaboration of an online community was launched on 9 March 2000 – *NuPedia*, the predecessor of *Wikipedia*.[14] More recently, various projects use crowdsourcing as a way to gather training data for Machine Learning methods. Especially large scale and complex learning approaches Deep Learning require large numbers of such instances to perform well. Crowdsourcing platforms such as Amazon *Mechanical Turk* and *CrowdFlower* enable researchers to collect training data efficiently. Applications of crowdsourcing for machine learning range from the medical domain (e.g. for recognizing tumors in images[15,16]), to Internet of Things[17] to Digital Humanities. Examples of the latter include geographical classification of location names in historical texts[18] or analysis of film colours.[19]

A first step involves the creation or compilation of training data: We needed annotations of historical photographs to match with geolocated photographs of the same building in their current form, so that computer could study the architectural

14 Larry Sanger. "The Early History of Nupedia and Wikipedia: A Memoir." *Open Sources 2.0: The Continuing Evolution*, ed. Chris DiBona, Mark Stone, and Danese Cooper (O'Reilly Media, Inc. 2005): 307–38.

15 Larry Sanger, "The Early History of Nupedia and Wikipedia: A Memoir", in: Chris DiBona, Mark Stone and Danese Cooper (eds.): *Open Sources 2.0: The Continuing Evolution*, ed. (O'Reilly Media, Inc. 2005): 307–38.

16 Shadi Albarqouni, Christoph Baur, Felix Achilles and Vasileios Belagiannis, "Aggnet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images", IEEE transactions on medical imaging 35.5 (2016): 1313–1321.

17 Qingchen Zhang, Laurence Tianruo Yang, Zhikui Chen, Peng Li and Fanyu Bu, "An Adaptive Dropout Deep Computation Model for Industrial IoT Big Data Learning with Crowdsourcing to Cloud Computing", IEEE Transactions on Industrial Informatics 15, 4 (April 2019): 2330–2337, doi: 10.1109/TII.2018.2791424.

18 Benjamin Adams and Grant McKenzie, "Crowdsourcing the Character of a Place: Character-Level Convolutional Networks for Multilingual Geographic Text Classification", Transactions in GIS 22.2 (2018): 394–408.

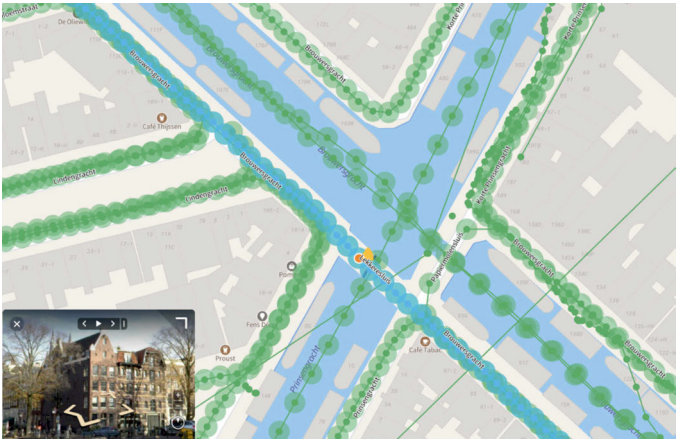
19 Barbara Flueckiger and Gaudenz Halter, "Building a Crowdsourcing Platform for the Analysis of Film Colors", Moving Image: The Journal of the Association of Moving Image Archivists 18.1 (2018): 80–83.

details. Some tasks are generic enough to be done via captchas,[20] but annotating historical street-view images requires some familiarity with the urban area depicted or some advanced training in planning-related fields. The same applies to the crowdsourcing task for the *ArchiMediaL* project. Due to limited means, it was necessary to find ways other than payment to motivate annotators to contribute. Therefore, the task needed to be uncomplicated and possibly combined with some gaming experience.[21]

The ArchiMediaL Annotator

In this regard, an annotation tool was created which showed the annotator a historical image and a 3D street-view navigator mostly positioned close to the expected location where the historical image had been taken, thus allowing the user to pan or zoom and match the historical and contemporary image in a game-like fashion (Figures 1 and 2).[22]

Fig. 1: Amsterdam, Brouwersgracht 160 on Mapillary. The green dots in the map indicate the positions of 360° images. Source: Mapillary (2016).

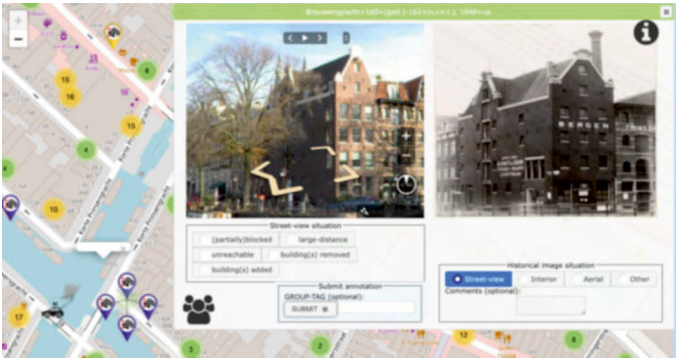


20 Luis von Ahn, Manuel Blum, Nicholas J. Hopper and John Langford, "Captcha: Using Hard AI Problems for Security", in: Eli Biham (ed.): *Advances in Cryptology—EUROCRYPT 2003* (Berlin, Heidelberg: Springer 2003), 294–311.

21 Benedikt Morschheuser, Juho Hamari and Jonna Koivisto, "Gamification in Crowdsourcing: A Review", 2016, 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016.

22 Clark C, Serious games, 1987.

Fig. 2: A historical picture of the Brouwersgracht 160 in Amsterdam from the 1940s and the corresponding street view scene in the submission form of the crowdsourcing tool. Source: ArchiMediaL, 2020.



Once the participant navigates to the approximate location, (s)he can use the rotating, panning and zooming features to approximate the historical image. The participant can choose which area to discover and annotate by navigating on a map and then click on blue markers in the desired neighbourhood. Each marker contains one or more historical images that were taken in proximity of the marker. (An orange marker indicates that this historical image is already annotated and currently under review. If the marker is green, it means that at least one successful annotation is already added (see Figure 2). In case the historical image is not a street-view image (e.g. an interior or an aerial photo), the user can skip this task by selecting the appropriate checkbox and submit the result. Otherwise, the navigation procedure as described above can start and when finished, specific checkboxes can be marked which describe the current street-view situation in comparison to the situation in the historical photograph. Users can indicate whether a perfect match is not possible, for example, because new buildings have been built or old ones taken down or because the street-view panoramic tool cannot reach the point where the photographer of the historical image took the shot. The tool also provides an assessment interface that allows administrators to manually review or reject the results. This task takes only a fraction of the time allotted per image in comparison to the annotation task itself. The administrator can also make changes and resubmit a result. In the experimental setup, project administrators used this tool to verify or deny crowd submissions.

System design

To increase transparency and reusability of the tool, it must be Free and Open Source Software (FOSS). This requires a mix of technology, software libraries and data that in combination fulfils the requirements. For example, although they have high-quality content, Google Street-view data, professional web front-ends, and GIS platforms such as ArcGIS etc. have restrictive licences or are too expensive, thus rendering them unsuitable for the project. In the end, it was possible to put together a mix that enabled the implementation of the annotation platform. It contains the following elements that are briefly descried:

Mapillary is a street-level imaging platform that automatically annotates maps by using computer vision. Mapillary brings together a global network of contributors who want to make the world accessible to everyone by visualizing it and creating richer maps. In order to use their widgets, users must register for a free API key that can be accessed from their well-documented JavaScript library.[23] *OpenStreetMap* is an editable map database created and maintained by volunteers and distributed under the Open Data Commons Open Database License.[24] *Leaflet* is a widely used open source JavaScript library used to create web mapping applications.[25] Leaflet allows developers without a GIS background to easily view tiled web maps hosted on a public server with optional tiled overlays. It can load feature data from GeoJSON files, apply a style to them, and create interactive layers, such as markers with pop-ups when clicked. *Google Maps* provided the latitude and longitude coordinates that were used to map the Amsterdam Beeldbank title and description metadata on these coordinates. *PostGIS*[26] is a spatial database extender for PostgreSQL[27] object-relational database. It adds support for geographical objects and allows one to execute location queries in SQL. To store the locations internally in PostGIS, that is the latitude and longitude coordinates from the previous step, a transformation must be applied.[28]

Additionally, the following JavaScript libraries were used: jQuery is a JavaScript library designed to simplify HTML DOM tree traversal as well as manipulation, event handling, CSS animation, and *Ajax*.[29] Leaflet.*MarkerCluster* is used for clustering

23 The Mapillary API, 2020 [accessed: 04.06.2020].

24 https://www.openstreetmap.org, [accessed: 02.07.2021].

25 Vladimir Agafonkin et al., "Leaflet", https://leafletjs.com/, 2020 [accessed: June 2, 2020]; Leaflet Wikipedia. https://en.wikipedia.org/wiki/Leaflet (software), 2020 [accessed: May 28, 2020].

26 Postgis. https://postgis.net/, 2020 [accessed: 05.06.2020].

27 Postgresql. https://www.postgresql.org/, 2020 [accessed: June 5, 2020].

28 Postgresql-lat-lon.https://postgis.net/docs/STMakePoint. html, 2020 [accessed: June 5, 2020].

29 jquery. https://jquery.com/, 2020 [accessed: June 5, 2020].

markers in *Leaflet*.[30] It uses a grid-based clustering method which makes it ideal for providing a fast solution to the many markers problem. Tabulator creates interactive tables for any HTML Tables, JavaScript Arrays, *AJAX* data sources or JSON formatted data.[31] *FontAwesome* is an icon toolkit featuring icon font ligatures, an SVG framework, official NPM packages for popular front-end libraries such as React, and facilitating access to a new CDN.[32] Most of the *ArchiMediaL* annotator icons have *FontAwsome* as the source. Node.js is a platform built on Chrome's JavaScript runtime for easily building fast and scalable network applications.[33] *Node.js* serves as the larger part of the *ArchiMediaL* back-end. *Puppeteer* is a *Node.js* library which provides a high-level API to control Chrome or Chromium over the *DevTools* Protocol.[34] *Puppeteer* runs headless by default, but can be configured to run full (non-headless) Chrome or Chromium.

Data collection

Users

We have experimented with various strategies to attract new commentators to the platform. The easiest and most successful one is that which gives students some time to experiment during lectures. The peak in new subscriptions was reached during one of the lectures at TU-Delft in the week beginning December 16, 2019 (see Figure 3).

30 Leaflet.markercluster. https://github.com/Leaflet/Leaflet.markercluster, 2020 [accessed: June 5, 2020].

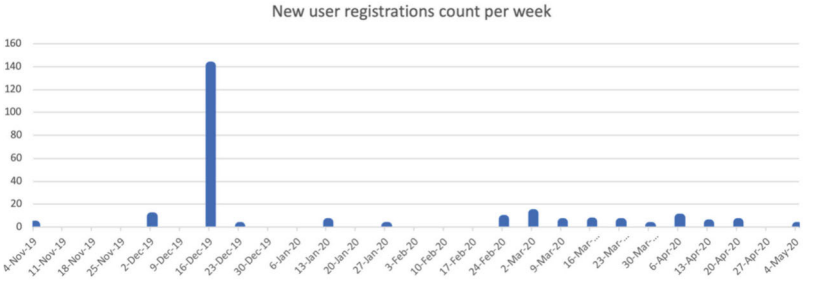
31 Tabulator. http://tabulator.info/, 2020 (accessed June 5, 2020).

32 Fontawesome. https://github.com/FortAwesome/Font-Awesome, 2020 (accessed June 5, 2020).

33 Node.js. https://nodejs.org/en/, 2020 (accessed June 5, 2020).

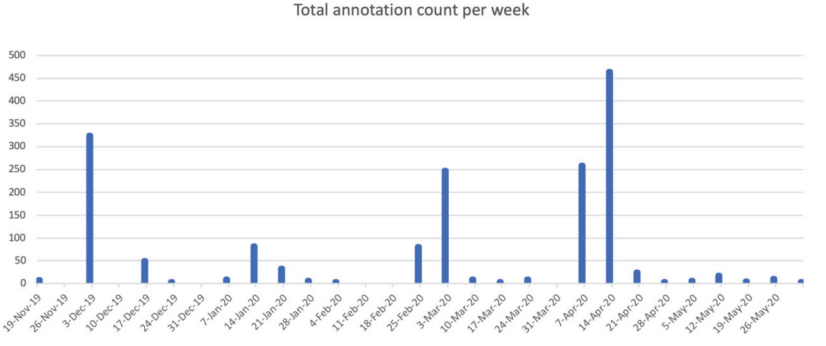
34 Puppeteer for node.js. https://github.com/puppeteer/puppeteer, 2020 (accessed June 5, 2020).

Fig. 3: New registrations per week.



Unfortunately, this stress test resulted in an unforeseen, unacceptable peak in the use of computing resources on the shared servers; and the host decided to temporarily shut down all processes. Since then, a considerable amount of time has actually been spent on making the processes more efficient. We hoped that the students would continue to play around with the annotator and help us with further annotations, but this did not happen. This explains why, although most new registrations were made during this event, most of the students did not submit an annotation either then or afterwards. Figure 4 shows the weekly number of new subscriptions on our platform. In the future, we plan to check the effectiveness of the measures we have taken to attract new subscribers.

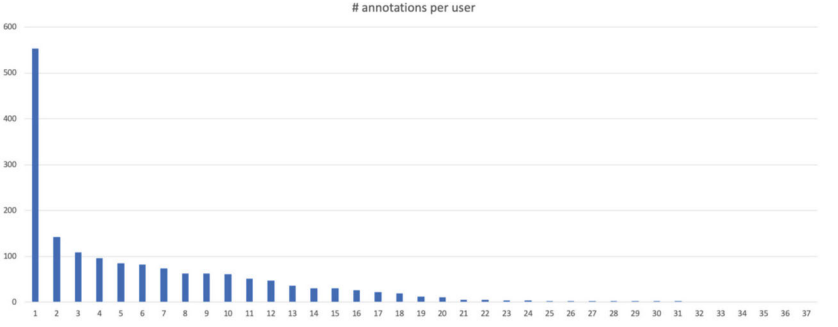
Fig. 4: Annotations per week



A highly distorted distribution, another interesting observation, can be found in Figure 5. Only a few users are responsible for the majority of the annotations. This is

a well-known phenomenon known as crowdsourcing participation inequality[35] best known in the case of contributions on Wikipedia.[36]

Fig. 5: Total amount of annotations per user.



Annotation statistics

Table 1 shows statistics of the annotations gathered during the initial experiment period (November 2019-May 2020). A total of 1,656 annotations were made, of which the majority (1,116 or 67%) were annotations where participants identified a historical image as a street view, with images of interiors being the second largest category (with 225 annotations). The 'other' category was the third largest. It included images, such as portraits and photographs of building equipment, paintings and other objects. The smallest category was that of aerial photographs (95).

Annotations submitted initially were categorized as 'pending review'. Thus, administrators could review and then accept or reject the annotations. In our experiment, the role of administrators was taken up by the authors. The table also shows the numbers of accepted versus rejected annotations. In total, 80% of annotations were accepted. This shows the quality of individual annotations. The percentage of accepted annotations was the lowest for street-view (75%) and highest for interiors (95%). One explanation for this is that street-view annotations are more elaborate as participants are asked to actually navigate to the correct location leading to greater risk of errors, whereas for the other categories a simple categorization is sufficient.

35 Stewart Osamuyimen, David Lubensky and Juan M. Huerta, "Crowdsourcing Participation Inequality: A SCOUT Model for the Enterprise Domain", Proceedings of the ACM SIGKDD Workshop on Human Computation. 2010.

36 Wikipedia. List of wikipedians by number of edits. https://en.wikipedia.org/wiki/Wikipedia:ListofW ikipediansbynumberofedits, 2020 (accessed: June 5, 2020).

Table 2 shows the result of further annotation details about the accepted street-view annotations by showing the additional categories identified by the participants. Here, we see that a small number of annotations concern street-view situations where the view of the building is partially hindered or changed. For the majority, this is however not the case (67.7%) and the feasibility of this annotation task is shown in 2/3rds of the cases in Amsterdam.

Table 1: Total of accepted, pending and rejected annotations per category.

	Street-view	Interior	Aerial	Other	Total
Accepted	835	243	84	165	1327
Rejected	278	12	11	25	326
Pending review	3	0	0	0	3
Total	1116	255	95	190	1656

Table 2: Statistics on street-view situations for accepted annotations.

Street-view situation	Total (percentage)
Blocked (partially)	103 (12.3%)
Large distance	50 (6.0%)
Unreachable	45 (5.4%)
Buildings removed	60 (7.2%)
Buildings added	79 (9.5%)
None	565 (67.7%)

Similarity Learning

Recent research literature shows that similarity learning is a powerful way to gain insight into data.[37] Considering the large variety of objects and therefore classes in most archival data and, in our case, the cross-time dataset for historical and current street view of Amsterdam, we use deep similarity learning for representation learning. We address the cross-domain image retrieval task, formulated as content-based image retrieval (CBIR), where the semantic similarity needs to be learned to find the most similar images in the gallery with reference to the query image. We use a *Siamese network* that uses two sister networks with shared learning parameters for the training process. The training tuples are images and their labels, in contrast to the classification where the training pair is the image and the label. The CNN network learns to project an image to a vector (latent) space such that similar images are placed closer, in terms of Euclidean distance, compared to dissimilar images.[38]

Once the network is trained for similarity learning, the CNN is used to map all the images in the dataset to a vector space. The distances between the vector representations in Euclidean sense determine the corresponding image pairs. In the retrieval task, the distance between the vector representations of the query image is computed with regard to the vector representations of the images in the gallery. The images are then ranked in ascending order with respect to their distances to the query.

Cross-domain Retrieval

In the context of cross-domain CBIR, representations of the objects in the gallery database can be potentially different from the query image. For example, the images may contain sketches, paintings or old photos as discussed in the beginning.

---

37 Sumit Chopra, Raia Hadsell and Yann Lecun, "Learning a similarity metric discriminatively, with application to face verification", *The IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005*, 1 (07, 2005): 539– 546; Gregory Koch, Richard Zemel and Ruslan Salakhutdinov, "Siamese Neural Networks for One-Shot Image Recognition", *ICML deep learning workshop* 2. 2015.

38 In other words, deep CNN is a mapping function f from an image of size w × h with r color channels to a k-dimensional representation space, $f : IN^{w\times h\times r} IR^k$ where distances in IR^k between similar image pairs are smaller than distances to dissimilar image pairs by a predefined margin m in the desired metric space, i.e., dPositive + m ≤ dNegative, where dPositive = d(f(I),f(I′)|y = 1), (1) dNegative = d(f(I), f(I′)|y = 0). (2) Similar image pairs have a label y = 1 and dissimilar image pairs the label y = 0. We consider a Euclidean metric space where d(f(I),f(I′)) outputs the Euclidean distance between two vectors of representations for an image pair {I, I′}.

For cross-domain CBIR tasks, a common failure mode, which comes into play while deploying the CNN models trained only to detect single-domain images, is that the network places similar-style images in a neighbourhood which depicts different objects as it has never seen one from the second domain (featuring historical images here). To resolve this domain-disparity issue, we propose to learn domain-invariant image representations that focus on semantics rather than on image style or colour. This leads to a specialized CNN model that learns indifferent image representations for archival and contemporary image domains but is discriminative at the content level.

In our work, a combination of attention and domain adaption is used to train robust CNN networks for an age-agnostic image retrieval task.[39] Here, similarity is learned by training a Siamese network to detect images with the same geo-tags in the contemporary image domain. It is commonly referred to as weakly supervised learning as the labels for training are not the same as the ones for testing. In our case, the test (evaluation) set contains a query from historical Amsterdam, but the gallery houses current street-views of Amsterdam. The results reveal a performance gap as the intra-domain trained model for street-view images is tested on cross-domain data, indicating a drop from 99% top 20 accuracy to 28%. The main conclusion is that full supervision is required to achieve reasonable performance for the cross-domain image retrieval task.

Conclusion

The research project started out by exploring automatic recognition of buildings in historic images by AI. Analogous to automatic facial recognition, buildings are to be recognized and identified. The input objects are historical images of buildings whose contents are localized by a specially designed and trained artificial neural network. The localization allows unique identification. The recognition can be realized for a specific area by providing the computer with many images of already identified, that is, localized buildings. This training of the network enables the computer to recognize buildings in historical images hitherto unknown to it. First, however, the training dataset must be created with hundreds of historical images of identified buildings.

Computer vision can help to identify image content in historical images of the built environment. As the case of Amsterdam exemplifies, a stock of +400k architectural images from the Beeldbank archives has not yet been clearly identified in

39 Ziqi Wang, Jiahui Li, Seyran Khademi and Jan van Gemert. "Attention-aware Age-agnosticVisual Place Recognition", *The IEEE International Conference on Computer Vision (ICCV) Workshops* (October 2019).

terms of geographical location, building name or type. Millions of similar images exist worldwide in online and offline repositories. The identification of the image content is not possible for architectural and urban historians simply because of the sheer quantity of the images. As a result, valuable visual source material for research in the humanities is lost. Computer vision seems to be helpful here. As buildings can be identified in images by comparing each of them with a geolocated image of the same building, street views from Mapillary are used as a reference system. However, there is no one-size-fits-all solution when it comes to learning representative features for various visual domains.[40] A CNN model pre-trained to follow standard benchmarks will not work on a non-standard image dataset with a different style of images and sizes.[41]

In our experimental setup mentioned, we showed that unsupervised or weakly supervised methods performed poorly when it came to recognizing historical image sets. Therefore, some form of supervised learning is inevitable for feature learning. The contribution of supervised learning to the cross-domain image retrieval is only revealed once annotated data is available for training. This type of crowdsourcing is a valuable way to achieve a sufficient number of exactly matching image pairs. The crowdsourcing stands at the intersection of deep learning and humanities research. Through advertising in university courses and social media, volunteers helped pair 1,656 images. A review and correction of these pairs have resulted in a set of 902 useful image pairs that can be used to train the neural network that is currently in progress. We hope to achieve a useful level of accuracy which will make automatic image content recognition a useful tool for architectural history research.

The findings of the ArchiMediaL project open up new perspectives for architectural history in diverse areas. Researchers, politicians and planners can explore 4D reconstructions of the past (e.g., in their respective websites, the HistStadt4D research project or various local Time Machine projects) to increase historical understanding, to enrich tourist experiences, or to facilitate design decisions. New research questions can be framed through the availability of such data. Using the ArchiMediaL tool can raise numerous questions. For example, scholars could examine bubbles where data are more or less available for raising questions such as: How

40 Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly and Neil Houlsby, "A large-scale study of representation learning with the visual task adaptation benchmark", 2019.arXiv:1910.04867.

41 Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel, "Imagenet-trained CNNs [PLS CHECK THE CAPITALIZATION] are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness", *CoRR, abs/1811.12231*, 2018.

does the availability of pictorial data from the past correlate with the architectural quality of the building stock or the socioeconomic composition of its citizens?

In the case of Amsterdam, many datasets with spatial information are available in digital form, including the ones based on the age of buildings, the number of breeding birds in green areas, climate information (heat, drought, flooding), post-war monumental wall art and land value, to name but a few. The crossing of this data with the visual sources localized within the work of the project allows framing of new research questions that investigate the connection between architectural or urban form and phenomena such as property value or gentrification. An expansion to other cities and areas will make it possible to formulate new findings on the basis of a greater number of correlations and thus make more general statements than those emerging from individual case studies.

The application of AI in historical research is not a mere information technology task. As in any mixed-methods approach, it requires meeting and communicating with different disciplines, and profound expertise in the humanities.[42] Interpretation of the past needs careful framing of available data to achieve meaningful findings. Such a step can only be made through interdisciplinary collaboration among humanities scholars, computer scientists, historians and designers. Moreover, this project has required people to contribute their knowledge both to create the training dataset and to eventually evaluate the performance of the algorithm. Crowdsourcing can offer an important opportunity for participation—important in terms of not only identifying past worlds, but also involving people in research. It would help integrate their points of view and ultimately awaken their interest in questions of urban history and urban development.

In Digital Humanities, the heterogeneity of data demand and data supply is a common challenge. Since research in the field of information technology requires high quality datasets, for example, to advance neural networks and the performance of computer vision, it is common to use datasets that are well-suited for technological search and to ignore those that fall short of the task. These are the datasets that humanities work with. In the course of addressing this challenge, we developed tools to enhance, even create the required datasets. In doing so, we were able to identify research questions that were useful for both scientific parties. Thus we refused to make IT research a mere supplier of a solution to a humanities problem and instead launched a joint research initiative involving researchers from all disciplines. In such an intellectual environment, the combination of humanities and IT research can help to make progress and gain new insights in both areas. It also offers the opportunity to formulate new research questions and to advance to more complex interdisciplinary research designs.

42 John W. Creswell and Vicki L. Plano Clark, Designing *and Conducting Mixed Methods Research*. 3rd. ed. (Los Angeles: Sage, 2018).

Bibliography

Abt, Clark C. Serious games. Lanham et al.: University Press of America, 1987.

Adams, Benjamin, and Grant McKenzie. "Crowdsourcing the Character of a Place: Character-Level Convolutional Networks for Multilingual Geographic Text Classification". In Transactions in GIS 22.2 (2018): 394–408.

Bengio, Yoshua; Aaron C. Courville, and Pascal Vincent. "Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives". In CoRR, abs/1206.5538 (24.06.2012): 1–30.

Chopra, Sumit; Raia Hadsell, and Yann Lecun. "Learning a similarity metric discriminatively, with application to face verification". In The IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005, 1 (07, 2005): 539–546.

Creswell, John W., and Vicki L. Plano Clark. Designing and Conducting Mixed Methods Research. 3rd. ed. Los Angeles: Sage, 2018.

Flueckiger, Barbara; Gaudenz Halter. "Building a Crowdsourcing Platform for the Analysis of Film Colors". In Moving Image: The Journal of the Association of Moving Image Archivists 18.1 (2018): 80–83.

Geirhos, Robert; Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. "Imagenet-trained CNNs are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness". In CoRR, abs/1811.12231, 2018, 22 p.

Khademi, Seyran, Tino Mager, Roland Siebes, Carola Hein, Beate Löffler, Jan van Gemert, Victor de Boer, and Dirk Schubert. Research project ArchiMediaL – Enriching and linking historical architectural and urban image collections (TU Delft, VU Amsterdam, TU Dortmund, HafenCity University Hamburg). https://archimedial.eu.

Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese Neural Networks for One-Shot Image Recognition". In ICML deep learning workshop 2. 2015.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In Advances in Neural Information Processing Systems 2 (01.12.2012): 1097–1105.

Löffler, Beate, Carola Hein, Tino Mager. "Searching for Meiji-Tokyo: Heterogeneous Visual Media and the Turn to Global Urban History, Digitalization, and Deep Learning". In Global Urban History (20.03.2018). https://globalurbanhistory.com/2018/03/20/searching-for-meiji-tokyo-heterogeneous-visual-media-and-the-turn-to-global-urban-history-digitalization-and-deep-learning/.

Löffler, Beate, and Tino Mager. 'Minor politics, major consequences—Epistemic challenges of metadata and the contribution of image recognition'. In Digital Culture & Society 6, 2 (2021): 221–238.

Mager, Tino, and Carola Hein. "Mathematics and/as Humanities: Linking Humanistic Historical to Quantitative Approaches". In The Mathematics of Urban Morphology, edited by Luca D'Acci. 523–528. Cham: Birkhäuser, 2019.

Morschheuser, Benedikt, Juho Hamari, Jonna Koivisto. "Gamification in Crowdsourcing: A Review". In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016.

Oomen, Johan, and Lora Aroyo. "Crowdsourcing in the cultural heritage domain: opportunities and challenges". In Proceedings of the 5th International Conference on Communities and Technologies (2011): 138–149.

Sanger, Larry. "The Early History of Nupedia and Wikipedia: A Memoir". In Open Sources 2.0: The Continuing Evolution, edited by Chris DiBona, Mark Stone, and Danese Cooper, 307–338. Beijing et al.: O'Reilly Media, Inc., 2005.

Albarqouni, Shadi, Christoph Baur, Felix Achilles, and Vasileios Belagiannis. "Aggnet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images". In IEEE transactions on medical imaging 35.5 (2016): 1313–1321.

Stewart, Osamuyimen, David Lubensky, and Juan M. Huerta. "Crowdsourcing Participation Inequality: A SCOUT Model for the Enterprise Domain". In Proceedings of the ACM SIGKDD Workshop on Human Computation. 2010.

von Ahn, Luis, Manuel Blum, Nicholas J. Hopper, and John Langford. "Captcha: Using Hard AI Problems for Security." In Advances in Cryptology—EUROCRYPT 2003, edited by Eli Biham, 294–311. Berlin, Heidelberg: Springer 2003.

Wang, Ziqi; Jiahui Li, Seyran Khademi, and Jan van Gemert. "Attention-aware Age-agnostic Visual Place Recognition." In The IEEE International Conference on Computer Vision (ICCV) Workshops. (Oct 2019).

Wevers, Melvin, and Thomas Smits. "The Visual Digital Turn: Using Neural Networks to Study Historical Images". In Digital Scholarship in the Humanities 35, 1 (April 2020): 194–207, https://doi.org/10.1093/llc/fqy085.

Wikipedia. List of wikipedians by number of edits. https://en.wikipedia.org/wiki/Wikipedia:ListofW ikipediansbynumberofedits, 2020 [accessed: 05.06.2020].

Zhai, Xiaohua; Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. "A large-scale study of representation learning with the visual task adaptation benchmark". 2019. arXiv:1910.04867

Zhang, Qingchen; Laurence Tianruo Yang, Zhikui Chen, Peng Li, and Fanyu Bu. "An Adaptive Dropout Deep Computation Model for Industrial IoT Big Data Learning With Crowdsourcing to Cloud Computing". In IEEE Transactions on Industrial Informatics 15, 4 (April 2019): 2330–2337, doi: 10.1109/TII.2018.2791424.

Repositories

TU Delft, University Library: Colonial architecture & town planning, http://colonialarchitecture.eu.
North Carolina State University: Urban panorama. https://www.visualnarrative.ncsu.edu/pro jects/urban-panorama/.
Beeldbank Stadsarchief Amsterdam. https://archief.amsterdam/beeldbank/.
Mapillary. http://mapillary.com, 2020 [accessed: 04.06.2020]