IMGS.AI. A MULTIMODAL SEARCH ENGINE FOR DIGITAL ART HISTORY

FABIAN OFFERT AND PETER BELL

ABSTRACT | We present a web application that facilitates multimodal search within institutional image collections using current-generation machine learning models like CLIP. Further, we discuss image retrieval as a combined computer vision/human-computer interaction problem, and propose that the standardization of feature extraction is one of the main problems that digital art history faces today.

KEYWORDS | Big image data, computer vision, feature extraction, machine learning, image retrieval

Introduction

Art history, as a discipline, is concerned with *multiple* images. A *singular* image can only become a historical entity in relation to its 'neighbors', to similar and dissimilar, related and unrelated original works, in both time and space. At the same time, sketches, copies, photographs, and other derivatives connect a work to its genesis, history, and reception. The significance of this comparative element has been already emphasized by Heinrich Wölfflin[1] and epitomized by Aby Warburg's[2] idiosyncratic method of tracing iconographic elements across history by physically arranging hundreds of images on special panels. Any singular image, in other words, leads to an *image corpus*.

Although today's digitization of art-historical collections has simplified the task of compiling image corpora enormously, it has also produced new challenges. One major challenge is the addressability of data through metadata. Institutional art-historical datasets commonly consist of images and related metadata, but web- and API-based search interfaces commonly target metadata exclusively. This is a result of the historical importance of periodization, attribution, and localization in art history. Many art-historical questions, however, that relate to "semantic" (e.g. iconographic) *and* "syntactic" (e.g. stylistic) aspects of a work are irreducible to metadata. Metadata schemas, especially those achievable with limited institutional resources, simply cannot accommodate the variety and complexity of the visual world in its entirety.

Such questions, then, can only be operationalized either as visual queries in the form of sample images that possess a unique combination of visual properties, or as fuzzy textual queries; free-form descriptions of visual characteristics. Where such questions are posed, or where metadata is unavailable or even unattainable (for instance for images of unknown provenance), the compilation of an image corpus thus becomes a computer vision (CV) task. This computer vision task then turns into a human-computer interaction (HCI) task, as computer vision tools and methods that perform visual and fuzzy textual search queries exist but are not robustly linked to art-historical resources.

imgs.ai, which has been in public beta since the fall of 2020, was the first publicly available digital art history application that implemented a multimodal approach to image retrieval based on CLIP and other pre-trained deep neural networks.[3] It addresses the combined CV-HCI challenge of image retrieval in digital art history by providing a Web-based interface, and machine-learning-based backend that allows the end user to both descriptively and visually search arbitrary image datasets.

Technologies of visual search

Within the field of digital art history, machine learning still has an experimental status.[4] Digital art history and computer vision share many of the same technical problems, yet few technical solutions from computer vision have "made it" into digital art history. While deep learning has been used productively in a number of academic experiments (see section 3), there exists, so far, no widely used system for, or unified approach to, multimodal search in digital art history. At the same time, existing systems suffer from severe limitations. Large-scale, commercial applications, like Google's image search, operate on proprietary algorithms and datasets, require a significant amount of resources to run, and cannot be easily adapted to art-historical content. On the academic end of the spectrum, experimental systems commonly suffer from complex setup procedures and limited reusability.

The main reason for this imbalance used to be the stark difference in subject matter. In computer vision, visual search systems used to target Internet-vernacular imagery exclusively. Research datasets like ImageNet[5] which consist of thousands of low-quality images of everyday objects exemplify this historical and disciplinary trajectory. In digital art history, however, no two corpora are alike. The variety of both form and content in art history surpasses even the most "diverse" benchmark datasets in computer vision. Solutions that worked within the default constraints of computer vision research thus used to fail (often spectacularly) in a digital art history context.[6]

With the introduction of models like CLIP, which are trained on very large image datasets that include art-historical data, these issues begin to become less prevalent. Consequently, the productive application and critical analysis of large, pre-trained models turns into a major field of inquiry for digital art historians. Specifically, the already established technique of feature extraction[7] – exploiting internal representations of deep neural networks, so-called *embeddings*, as compressed semantic descriptors for images – has not been used to its full potential since CLIP's release. imgs.ai allows the user to employ feature extraction in an interactive manner, and thus attempts to close this gap on both the CV and the HCI side, facilitating a human-in-the-loop approach to multimodal search.

Review of existing visual search tools in digital art history

Existing deep visual search tools in digital art history can be roughly divided into three different classes: toolkits, macro interfaces, and search interfaces. Toolkits are collections of scripts, standalone tools and libraries – or combinations of all three – that allow the user to integrate selected machine learning algorithms into their research projects. Macro interfaces present entire image datasets as interactive or static plots, where the spatial proximity of images is determined by similarity. Finally, search interfaces allow the user to upload an image, or select an image from a dataset, and receive a set in return that, again, is determined by a similarity relation to the input image. Macro interfaces and search interfaces can either be dataset-specific or dataset-agnostic. Toolkits are, by default, dataset-agnostic.

Macro interfaces have been part of digital art history from the very beginning with Lev Manovich's *ImagePlot*[8] and have been continuously improved within the framework of projects like Douglas Duhaime's *PixPlot*,[9] which is the most widely used macro interface today. Macro interfaces usually provide two core functionalities: feature extraction and image clustering, which are often supplemented by metadata integration. On one hand the prominence of macro interfaces is partly due to the pragmatic need to "get to know" large image datasets, to ask, what images do we even have? On the other hand, it is due to their direct link to more traditional forms of "distant viewing"[10] in art history, namely Aby Warburg's *Mnemosyne Atlas*[11] and its contemporary equivalents in the work of artists like Gerhard Richter[12] or Douglas Blau. Beyond supporting standalone tools, feature extraction and image clustering have become important methods in research endeavors like the *Sphere* project[13] or the work by Impett and Moretti[14] on Aby Warburg's *Pathosformeln*.

It is difficult to establish a boundary between "regular" libraries and packages and those which could be understood as toolkits for deep visual search. Almost all deep visual search implementations depend on Python libraries like *NumPy*,[15] *scikit-learn*,[16] *PyTorch*,[17] and others. More specific resources include the *Distant Viewing Toolkit*[18] and the resources published by the *Training the Archive Project*.[19] Hybrid systems like *imagegraph.cc*[20], developed by Leonardo Impett, promise to combine the feature richness of toolkits with the ease-of-use of macro interfaces.

Search interfaces are almost always dataset-specific. One of the earliest examples explicitly designed for digital art history was Benoît Seguin's *Replica Project*[21] which combined deep visual and metadata search. While imgs.ai was the first publicly available, dataset-agnostic multimodal

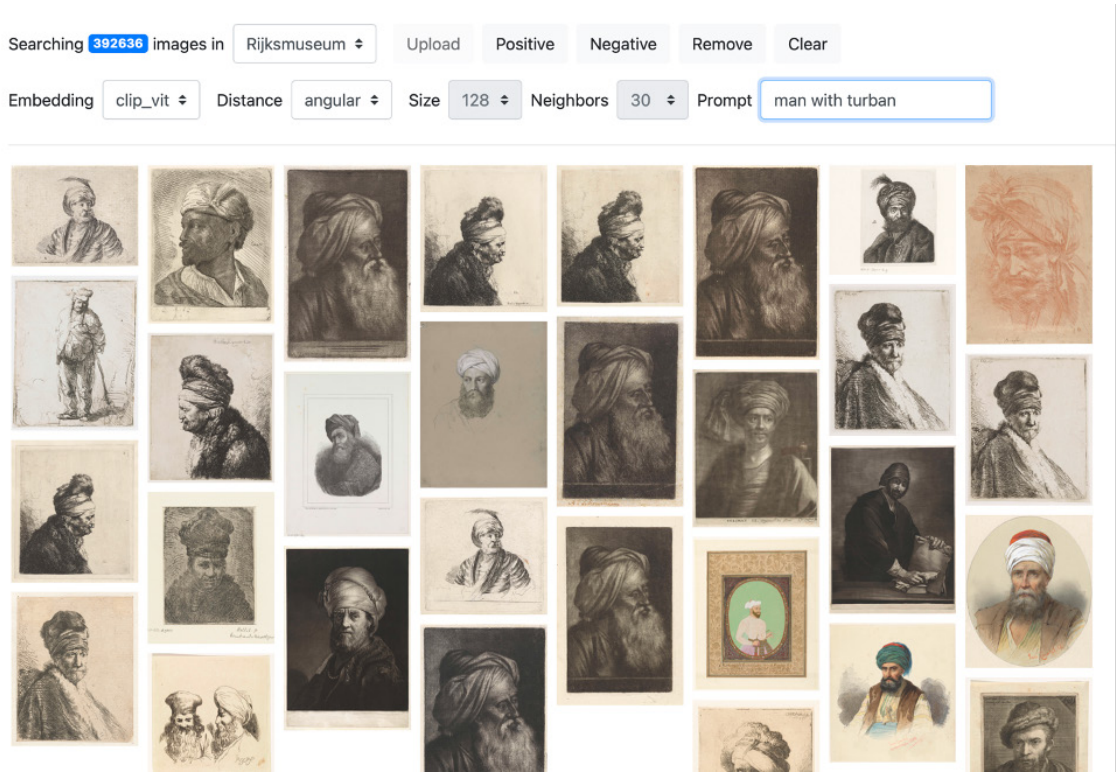

Figure 1. Application interface, showing results for textural search for "man with turban" in the Rijksmuseum collection. The toolbar allows switching datasets or embeddings, adding an image as a positive or negative example, or starting a new search.

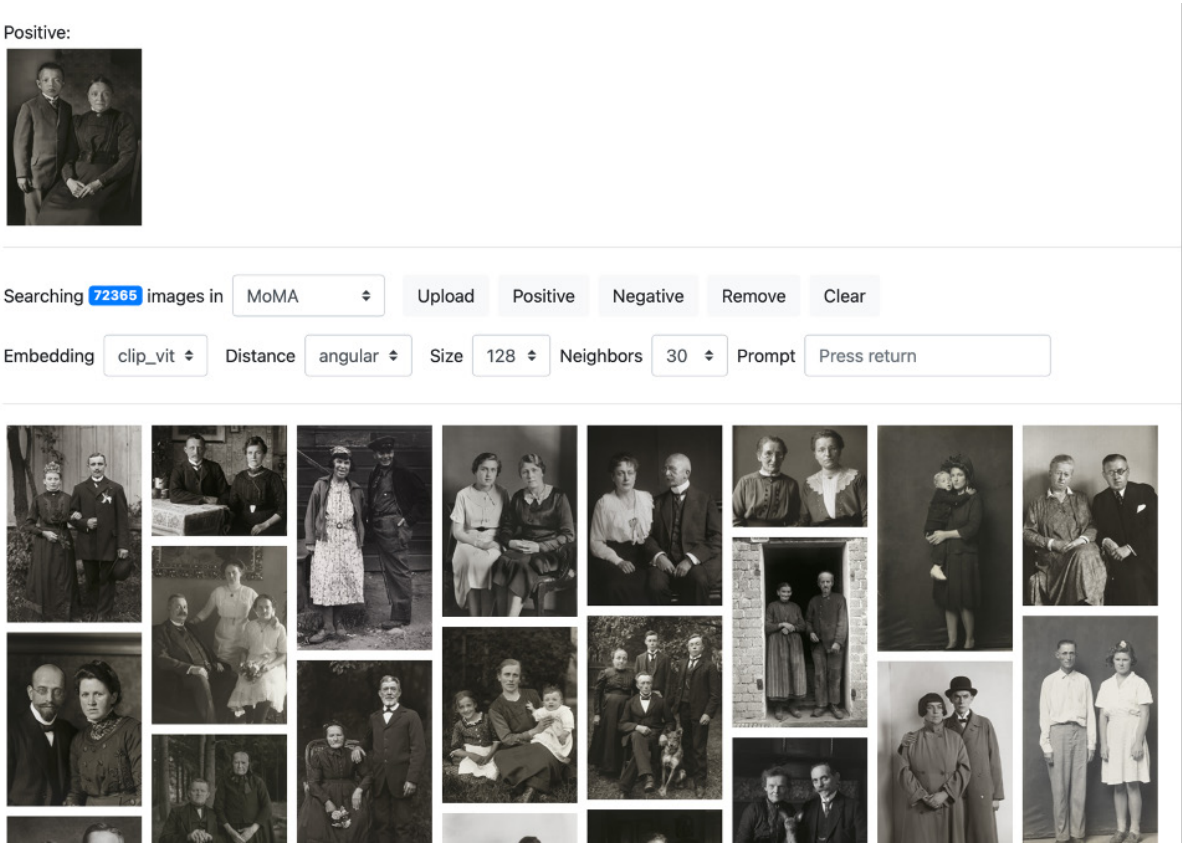

Figure 2. Application interface, showing results for the CLIP-based textural search "a photograph by August Sander" in the Museum of Modern Art, New York collection. The initial results were refined by selecting a portrait photograph featuring two people. The second round of results now features group portraits by August Sander exclusively. No metadata was involved in producing these results.

search engine for digital art history, a few other systems have since emerged. A recent example is the search interface developed by Florian Kräutli for the *Bilder der Schweiz* image repository.[22] Another is *imagemesh.ai*, intended as an exploration tool for scientific illustrations uploaded to arXiv[23] which was directly inspired by the application described in this paper. A recent dataset-agnostic example, developed by Stefanie Schneider, is iArt[24]. Finally, Microsoft's *MosAIc*[25] emphasizes cross-cultural similarity but is not transparent about the data used.

Design considerations

imgs.ai responds to what we see as five essential criteria of multimodal search in digital art history:

- Simplicity. A visual search system should be both simple to run and simple to use. "Simple to run" implies moderate compute requirements and an easy setup process, and "simple to use" implies a consistent interface that requires very little additional documentation.
- Speed. Art-historical work within large image corpora means sighting and sorting hundreds of images as part of an iterative process. Consequently, results should be produced in real-time.
- Non-locality. Image datasets use significant amounts of storage, and are often accessed remotely via APIs or web interfaces. A multimodal search system should allow for remote searches and should work on datasets without locally cached copies.
- Transparency. The criteria that a visual search system uses to produce results should be clear. They should not be hidden behind abstract concepts like "similarity" but named explicitly. Furthermore, techniques from explainable machine learning should be integrated to facilitate the interpretation of results.
- Interactivity. If speed, modularity, and simplicity are given, the user should enter into a hermeneutic dialogue with the machine, to explore the data while refining their own search intention. One of the main design considerations that supports this criterion is the concept of 're-search': results from a search (both descriptive and visual) can immediately form the basis of another search. This allows the intuitive and successive compilation of an image corpus based on continuous refinement.

All of these criteria have implications for both the frontend and backend design of imgs.ai.

The application's Web frontend is rendered server-side by the *Flask* Python library[26] and consists of two "light boxes" that are divided by a toolbar. The top light box contains the query images, and the bottom light box contains the search results. When first initialized, the query light box is empty, and the results light box contains a random selection from the dataset under investigation. A search process is initiated by selecting a similarity criterion, and either uploading or selecting one or multiple images as a query. The frontend, therefore, encourages the following "human in the loop" workflow:

1. The user selects a dataset.
2. The user selects a similarity criterion: a neural network and a distance metric.
3. The user either uploads an image, selects one or more images, or enters a text prompt to produce a query in the top light box. The prompt textbox is only available if CLIP is the selected neural network.The bottom light box displays images in the dataset "most similar" to the query.
4. The user selects additional images from the results light box, which are added to the query light box.
5. The refined results are displayed.

Selecting a result allows it to be added to the next query as a positive or negative example, right-clicking a result shows links to a full-resolution version of the image, and its source, for example, the specific institutional website it is hosted on.

Datasets can be changed mid-search which allows the user to refine a query in relation to one dataset, and then look for the specific visual attributes defined by this refined query in another. Following this workflow, the user enters into a dialogue with the machine that, in addition to surprises and insights, can also present disappointments and misunderstandings. These "negative" results are, however, another step in the refinement of the user's search. Nevertheless, negative results can also be explicitly marked. The user can not only define what should be searched for (positive), but also what should no longer appear in the result set (negative). By means of vector arithmetic, future results will then contain "less of" the kind of image marked. These negative examples, like the positive ones, do not only stand for themselves but also for further images with similar characteristics.

Depending on the dataset chosen, multiple similarity criteria are available. These always combine a pre-trained model used as a feature extractor to compute embeddings for all images in the dataset and a distance metric from which the suggested nearest neighbors are determined. Given any pre-indexed dataset, a search query simply amounts to a look-up operation, which usually processes in under a second, and some time to load the result images, usually directly from an institution's servers.

Figure 3. Diego Velásquez, *Las Meninas* (1656, © Museo nacional del Prado.)

The application described in this paper is set up to index datasets based on four kinds of feature extractors: VGG19, "poses", "raw", and CLIP. VGG19[27] is a neural network architecture that has been incredibly popular for all kinds of classification tasks. While the pre-trained model utilized has been trained on ImageNet, it works well for art historical material, with a focus on stylistic similarity. The pose feature extractor is built on top of a Keypoint R-CNN model with a ResNet-50-FPN backbone,[28] which is especially useful for datasets of figurative works, as shown for instance by Impett and Moretti.[29] The "raw" feature extractor simply treats a (resized) image's raw color data as its embedding, and thus allows searching for images that use similar pallets.

Finally, the CLIP extractor uses the pre-trained model of the same name released by OpenAI in 2021[30] and enables multimodal search based on images and text prompts. CLIP learns from images in context by projecting an image and its context into a common embedding space. The 'context' here could be an image caption, a so-called 'alt text' which describes the image in case it is not loaded properly and to accommodate people with screen readers, or simply a news article that the image illustrates.

The backend for imgs.ai paper consists of two separate applications: one that indexes image datasets (training backend), and one that communicates with the frontend described above and powers the actual search within a pre-indexed dataset (server backend). The training backend allows the user to select the feature extractors to be used on each dataset, and run them in a fast, parallel fashion. Subsequently, approximate nearest-neighbors relations for each image are pre-computed using a tree-based implementation provided by the *Annoy* Python library[31] developed for Spotify. The resulting model, which consists of the extracted nearest-neighbors relations as well as the full embeddings for each image, can then be used with the server backend. The server backend prioritizes the fast delivery of results. Both embeddings and nearest neighbors for each indexed dataset are stored as compressed HDF5 files on disk which allows the system to operate on a minimal amount of memory even given a high frequency of queries. The server backend delivers nearest-neighbors results to the frontend based on queries submitted, and extracts new embeddings/nearest-neighbors relations for newly uploaded images.

Figure 4. Robert Doisneau, *La Dame Indignée* (1948, © Robert Doisneau / Gamma Rapho) and Joel Meyerowitz, *Jeu de Paume, Paris, France* (1967, ©Joel Meyerowitz, Courtesy Howard Greenberg Gallery). Museum of Modern Art, New York.

Figure 5. Richard Hamilton, *Picasso's Meninas* from *Homage to Picasso* (1973, © with kind permission of VG Bildkunst). Museum of Modern Art, New York.

Figure 6. GRAD-CAM heatmap overlay for *Man in oosterse kleding* by Rembrandt van Rijn (1635), resulting from a search for "man with turban" in the Rijksmuseum collection. The overlay shows that the focus of the CLIP model lies on the head garment.

The potential of CLIP

Without taking any metadata into account, CLIP makes it possible to search for iconographic criteria. CLIP not only "knows" image objects like "man with turban", "three cows", etc., but also "understands" more abstract terms or concepts. For instance, it has a sense of crowds ("crucifixion with a huge crowd"), feelings ("a happy family"), activities, and genres. As it allows arbitrary user input, it is also possible to search for aspects of an image that would never appear in conventional metadata. This not only makes it possible to make surprising finds, but also to transcend the limitations of metadata-based search.

One surprising result facilitated by CLIP, for instance, is the list of search results for the prompt "Las Meninas", which references the famous 1656 Spanish Golden Age painting by Diego Velásquez. *Las Meninas* is certainly one of the most analyzed paintings in the history of art. Michel Foucault, famously, spends the whole introduction of *The Order of Things* on it,[32] W.J.T. Mitchell dedicates a whole chapter in his book on picture theory to it.[33] It is also the focus of countless analyses in less prominent scholarship. The painting is famous, in particular, for its play on representation. Such metapictorial aspects, however, are difficult to synthesize, and certainly not part of any approach to metadata. If we run a search for "Las Meninas" in the collection of the Museum of Modern Art, New York – an institution that does not only not have the famous painting in its collection (which is kept in the Prado in Madrid) but also focuses on modern/contemporary art, rather than works from the Spanish Golden Age – the results are surprisingly "accurate" and show the conceptual depth that CLIP allows the user to access. Among them are two photographic works, Joel Meyerowitz' *Untitled* from *The French Portfolio* (1980) and Robert Doisneau's *La Dame Indignée* (1948). Both are explicit plays on representation, and both clearly pick up on the same themes as *Las Meninas*,

especially the question of the gaze relation between people in, and people before the image. Another result is Richard Hamilton's *Picasso's Meninas* from *Homage to Picasso* (1973) which takes up the structure of the Velásquez original but fills it with figures from Picasso paintings. While both works have nothing in common but their compositional structure, CLIP is still able to "understand" their compositional similarity.

While the possibilities provided by CLIP are thus extensive, like many pre-trained models CLIP suffers from issues of opacity. While the proprietary nature of its training data cannot be mitigated on the application side, imgs.ai provides a mechanism to aid with the interpretation of results generated from text queries. By right-clicking a result suggested by CLIP, the user can generate a basic attribution heatmap using a GRAD-CAM[34] implementation provided by the *TorchRay* Python library[35], which shows the importance of certain image regions in relation to the query.

Conclusion: Standardizing feature extraction

imgs.ai is just one of what could be many solutions to the productive use of feature extraction with a focus on multimodal features in digital art history. As such, it points to a significant challenge at the exact interface of technical and academic work within digital art history in particular, and the digital humanities in general. Given the increasing footprint of machine learning models, it seems counterproductive to extract features – which, overall, have proven incredibly useful to digital art history – more than once. The standardization of extracted features – in terms of method, format, and infrastructure – is thus the next big challenge the digital humanities community has to solve. While some aspects of this process seem straightforward to implement, others, such as the question of the accessibility and sustainability of models used for extraction, provide additional challenges. Solving these challenges will require a close collaboration between academic researchers and software engineers within the digital art history community.

NOTES

1 Heinrich Wölfflin, *Kunstgeschichtliche Grundbegriffe. Das Problem der Stilentwicklung in der neueren Kunst* (München: Verlag Hugo Bruckmann, 1917), see also Matthias Bruhn and Gerhard Scholtz, *Der Vergleichende Blick* (Berlin: Dietrich Reimer Verlag, 2017)
2 Aby Warburg, "Mnemosyne Einleitung," in *Werke*, ed. Martin Treml, Sigrid Weigel, and Perdita Ladwig (Frankfurt am Main: Suhrkamp, 2010), see also Ernst H. Gombrich, *Aby Warburg* (Warburg Institute, University of London, 1970) and Georges Didi-Huberman, *The Surviving Image: Phantoms of Time and Time of Phantoms: Aby Warburg's History of Art* (Pennsylvania State University Press, 2017)
3 Alec Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," in International Conference on Machine Learning (ICML), 2021, 8748–63
4 See Leonardo Impett and Fabian Offert, "There Is a Digital Art History", forthcoming 2023
5 Jia Deng et al., "Imagenet: A Large-Scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 248–55, see also Olga Russakovsky et al., "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115, no. 3 (2015): 211–52
6 A few exceptions exist in specialized work at the intersection of digital art history and computer science, see for instance Peter Bell and Fabian Offert, "Reflections on Connoisseurship and Computer Vision," *Journal of Art Historiography*, no. 24 (2021), Eva Cetinic, Tomislav Lipic, and Sonja Grgic, "Fine-Tuning Convolutional Neural Networks for Fine Art Classification," *Expert Systems with Applications* 114 (2018): 107–18, Prathmesh Madhu et al., "Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-Grounded Style Transfer Learning," *arXiv preprint 2012.05616*, 2020, or Melvin Wevers and Thomas Smits, "The Visual Digital Turn: Using Neural Networks to Study Historical Images," *Digital Scholarship in the Humanities* 35, no. 1 (2020): 194–207.
7 e.g. Richard Zhang et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 586–95
8 Lev Manovich, "How to Compare One Million Images?," in *Understanding Digital Humanities* (Springer, 2012), 249–78, see: https://github.com/culturevis/imageplot
9 https://github.com/YaleDHLab/pix-plot
10 Taylor Arnold and Lauren Tilton, "Distant Viewing: Analyzing Large Visual Corpora," *Digital Scholarship in the Humanities*, 2019
11 Warburg 2010
12 see Benjamin Buchloh, "Gerhard Richter's 'Atlas': The Anomic Archive," *October*, 1999, 117–45
13 Florian Kräutli, Daan Lockhorst, and Matteo Valleriani, "Calculating Sameness: Identifying Early-Modern Image Reuse Outside the Black Box," *Digital Scholarship in the Humanities*, 2020
14 Leonardo Impett and Franco Moretti, "Totentanz. Operationalizing Aby Warburg's Pathosformeln," *New Left Review* 107 (2017), 68-97
15 https://numpy.org
16 https://scikit-learn.org/stable/
17 https://pytorch.org
18 Arnold and Tilton 2019, see: https://github.com/distant-viewing/dvt
19 Dominik Bönisch, "The Curator's Machine: Clustering of Museum Collection Data through Annotation of Hidden Connection Patterns between Artworks," *International Journal for Digital Art History*, no. 5 (2020): 5–20, see: https://github.com/DominikBoenisch/Training-the-Archive
20 http://imagegraph.cc
21 Isabella di Lenardo, Benoît Laurent Auguste Seguin, and Frédéric Kaplan, "Visual Patterns Discovery in Large Databases of Paintings" (DHLAB – Ecole Polytechnique Fédérale de Lausanne (EPFL), 2016), see: https://profile.benoitseguin.net/2016/12/19/replica-project-status-and-roadmap.html
22 https://bso.swissartresearch.net/resource/page:clipSearch
23 Kynan Tan, Anna Munster, and Adrian Mackenzie, "Images of the ArXiv: Reconfiguring Large Scientific Image Datasets," *Journal of Cultural Analytics* 3, no. 1 (2021), see: https://imagemesh.ai
24 Stefanie Schneider et al., "iART - Eine Suchmaschine zur Unterstützung von bildorientierten Forschungsprozessen," in *8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHd 2022, Potsdam, Germany*, ed. Michaela Geierhos, 2022, see: https://projects.tib.eu/iart/ueber-das-projekt/
25 Mark Hamilton et al., "MosAlc: Finding Artistic Connections across Culture with Conditional Image Retrieval," in *NeurIPS 2020 Competition and Demonstration Track*, 2021, 133–55
26 https://flask.palletsprojects.com/
27 Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint 1409.1556*, 2015

28 Kaiming He et al., "Mask R-CNN," *arXiv preprint 1703.06870*, 2017
29 Impett and Moretti 2017
30 Radford et al. 2021
31 https://github.com/spotify/annoy
32 Michel Foucault, *The Order of Things* (Routledge, 2005)
33 W. J. Thomas Mitchell, *Picture Theory: Essays on Verbal and Visual Representation* (University of Chicago Press, 1995)
34 Ramprasaath R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 618–26
35 https://github.com/facebookresearch/TorchRay

BIBLIOGRAPHY

Arnold, Taylor, and Lauren Tilton. "Distant Viewing: Analyzing Large Visual Corpora." *Digital Scholarship in the Humanities*, 2019.

Bell, Peter, and Fabian Offert. "Reflections on Connoisseurship and Computer Vision." *Journal of Art Historiography*, no. 24 (2021).

Bönisch, Dominik. "The Curator's Machine: Clustering of Museum Collection Data through Annotation of Hidden Connection Patterns between Artworks." *International Journal for Digital Art History*, no. 5 (2020): 5–20.

Bruhn, Matthias, and Gerhard Scholtz. *Der Vergleichende Blick*. Berlin: Dietrich Reimer Verlag, 2017.

Buchloh, Benjamin. "Gerhard Richter's 'Atlas': The Anomic Archive." *October*, 1999, 117–45.

Cetinic, Eva, Tomislav Lipic, and Sonja Grgic. "Fine-Tuning Convolutional Neural Networks for Fine Art Classification." *Expert Systems with Applications* 114 (2018): 107–18.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A Large-Scale Hierarchical Image Database." In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–55, 2009.

Didi-Huberman, Georges. *The Surviving Image: Phantoms of Time and Time of Phantoms: Aby Warburg's History of Art*. Pennsylvania State University Press, 2017.

Foucault, Michel. *The Order of Things*. Routledge, 2005.

Gombrich, Ernst H. *Aby Warburg*. Warburg Institute, University of London, 1970.

Hamilton, Mark, Stephanie Fu, Mindren Lu, Johnny Bui, Darius Bopp, Zhenbang Chen, Felix Tran, et al. "MosAIc: Finding Artistic Connections across Culture with Conditional Image Retrieval." In *NeurIPS 2020 Competition and Demonstration Track*, 133–55, 2021.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. "Mask R-CNN." *ArXiv Preprint* 1703.06870, 2017.

Impett, Leonardo, and Franco Moretti. "Totentanz. Operationalizing Aby Warburg's Pathosformeln." *New Left Review* 107 (2017).

Impett, Leonardo, and Fabian Offert. "There Is a Digital Art History". *Visual Resources* 38 no.2 (2024).

Kräutli, Florian, Daan Lockhorst, and Matteo Valleriani. "Calculating Sameness: Identifying Early-Modern Image Reuse Outside the Black Box." *Digital Scholarship in the Humanities*, 2020.

Lenardo, Isabella di, Benoît Laurent Auguste Seguin, and Frédéric Kaplan. "Visual Patterns Discovery in Large Databases of Paintings." *DHLAB – Ecole Polytechnique Fédérale de Lausanne (EPFL)*, 2016.

Madhu, Prathmesh, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas Maier, and Vincent Christlein. "Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-Grounded Style Transfer Learning." *ArXiv* Preprint 2012.05616, 2020.

Manovich, Lev. "How to Compare One Million Images?" In *Understanding Digital Humanities*, 249–78. Springer, 2012.

Mitchell, W. J. Thomas. *Picture Theory: Essays on Verbal and Visual Representation*. University of Chicago Press, 1995.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. "Learning Transferable Visual Models from Natural Language Supervision." In *International Conference on Machine Learning* (ICML), 8748–63, 2021.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. "Imagenet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115, no. 3 (2015): 211–52.

Schneider, Stefanie, Matthias Springstein, Javad Rahnama, Hubertus Kohle, Ralph Ewerth, and Eyke Hüllermeier. "iART - Eine Suchmaschine zur Unterstützung von bildorientierten Forschungsprozessen." In *8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum*, DHd 2022, Potsdam, Germany, edited by Michaela Geierhos, 2022.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.

"Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." In *Proceedings of the IEEE International Conference on Computer Vision*, 618–26, 2017.

Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *ArXiv* Preprint 1409.1556, 2015.

Tan, Kynan, Anna Munster, and Adrian Mackenzie. "Images of the ArXiv: Reconfiguring Large Scientific Image Datasets." *Journal of Cultural Analytics* 3, no. 1 (2021).

Warburg, Aby. "Mnemosyne Einleitung." In *Werke,* edited by Martin Treml, Sigrid Weigel, and Perdita Ladwig. Frankfurt am Main: Suhrkamp, 2010.

Wevers, Melvin, and Thomas Smits. "The Visual Digital Turn: Using Neural Networks to Study Historical Images." *Digital Scholarship in the Humanities* 35, no. 1 (2020): 194–207.

Wölfflin, Heinrich. *Kunstgeschichtliche Grundbegriffe. Das Problem der Stilentwicklung in der neueren Kuns*t. München: Verlag Hugo Bruckmann, 1917.

Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 586–95.

FABIAN OFFERT is Assistant Professor for the History and Theory of the Digital Humanities in the Department of Germanic and Slavic Studies at the University of California, Santa Barbara. His research and teaching focuses on the visual digital humanities, with a special interest in the epistemology and aesthetics of computer vision and machine learning. At UCSB, he is affiliated with the Media Arts and Technology program, the Comparative Literature program, and the Center for Responsible Machine Learning. He is also principal investigator of the UCHRI multi campus research group "Critical Machine Learning Studies" (2021-23), and the international research project "AI Forensics" (2022-25), funded by the VW foundation

Correspondence e-mail: offert@ucsb.edu

PETER BELL studied Art History, Economics, and Visual Arts at Marburg University. He was Research Associate in the DFG SFB 600 research cluster at Trier University, (PhD 2011), Postdoctoral Researcher at Heidelberg University, Research Associate at the Prometheus Image Archive at Cologne University, as well as group leader at the Heidelberg Academy of Sciences and Humanities. From 2017-2021 he was Assistant Professor in Digital Humanities at the University of Erlangen-Nürnberg (FAU) and is now professor of Art History and Digital Humanities at Philipps University Marburg. He was also Principal Investigator of the DFG SPP 2172 "The Digital Image" and is speaker of the Digital Art History working group.

Correspondence e-mail: peter.bell@uni-marburg.de