VIII.

Digitalna humanistika

Overview of Computer Vision-Based Tools in Search and Analysis of Historical Content

Thomas MANDL
Institute of Information Science and Language Technology
University of Hildesheim, Germany
mandl@uni-hildeheim.de

Abstract

Thanks/Due to progress in image analysis, researchers in the Digital Humanities (DH) can now use sophisticated computer vision systems for their work. Deep learning models have led to the implementation of powerful search systems, which are also suitable for historical and artistic content. However, current systems often lack controllability for users and therefore lead to an unsatisfying user experience. This paper explores research in this domain and analyses the potential of several current search and analysis tools. Tools based on deep learning computer vision yield sometimes unexpected but highly promising results; however, the controllability for users needs to be improved.

Keywords: Computer Vision, cultural heritage, Digital Humanities (DH), image analysis

1. Introduction

Research on historical documents has seen a considerable development in the last years. The availability of extensive digital collections on the internet has led to the adoption of new research paradigms in the humanities (Bullin & Henrich, 2020). The digitization in the humanities is currently characterized by the increased use of digital tools, which, when combined with conventional qualitative-hermeneutic methodologies, give rise to mixed methods

approaches (e. g., Heuwing, Mandl & Womser-Hacker, 2016). In many research areas, the need to combine qualitative, analytic-interpretative oriented methods with digital and quantitative methods becomes apparent.

Digital editions (Sahle, 2016) allow more interactivity and facilitate access. Optical character recognition makes content searchable and is even applied to handwritten text. Thus, less explored material such as books written in Glagolitic script gain attention (Tomić, 2018).

Text mining technology can be used to automatically analyse multiple texts at once. Knowledge derived from text mining includes distribution patterns and the frequency of words in texts and their components. Based on such knowledge, the user can explore and determine how topics are dealt with in many text documents, which attitudes on specific topics are expressed, and how issues evolve over time (Mandl, 2015). Research infrastructures such as DARIAH.eu provide a variety of methods for digital text analysis.

Moreover, the visual dimension of cultural heritage is also valued in digital collections and research (Donig, 2023). Cultural heritage collections of visual material in digital formats have grown tremendously in size. The digital library Europeana.eu alone contains millions of art-related objects. The analysis of such collections can be approached from various perspectives and with a large variety of objectives (Stork, 2024). These include art historical interests, book design questions, and historical research. Traditional systems are based on manual indexing and metadata. This article will focus on automatic approaches for analysing and searching images.

First, a brief summary of image processing in Computer Vision is provided. Then, approaches of processing images within Digital Humanities are sketched out. At the core of this article, several recent systems for image search are presented. Their potential and their weaknesses are discussed.

2. Fundamentals of Image Processing

Image processing can be considered a part of signal processing or computer graphics. Currently, it also is related to Artificial Intelligence because recent methods have led to great advances. Basically, digital images can be seen as collections of pixels. This is true for raster images as they are produced by

digital cameras or scanners. Vector images are another class of images that are produced by graphic programs. This paper will focus on raster images, as they are mostly used in digital collections. In raster images, each pixel is associated with colour values. Operations such as blurring and sharpening can be carried out by slightly modifying these pixel values in relation to the values of the neighbouring pixels.

A digital perspective focusing on the matrix of numbers neglects the complexity of human image processing. Humans process images at higher levels and focus on the boundaries between regions and objects. The lack of systems capable of processing images in the same way humans do is often referred to as the semantic gap (Barz & Denzler, 2021). However, during the last decade computers have managed to mimic some of the human visual recognition capabilities.

Very successful are Convolutional Neural Networks (CNN), which combine many neurons into complex architectures. A basic CNN is composed of recurring sets of two layers: a convolution layer and a pooling layer. The convolution in a CNN combines pixels in a close neighbourhood similar to a blurring filter. The filter can be thought of as a calculation operation, which leads to new pixel values at the next level. Several filters are applied simultaneously. Pooling is a mechanism that considers a neighbouring group of pixels and only lets the strongest value pass through. Pooling reduces the amount of data. The image is transferred through several layers of convolution and pooling. The remaining values represent the content of the initial image and can be understood as its fingerprint. This representation cannot be read by humans; however, similar images will receive similar output values. These output vectors of numerical values derived from the initial pixel values are called embeddings. The embeddings are fed into a supervised learning algorithm to solve tasks such as classification. An example could be the identification of indoor vs. outdoor images. The system also feeds learning errors back into the CNN, and modifies the values of the filters accordingly, so that the system will be able to perform the task more effectively and minimize errors (Skansi, 2018).

CNNs and similar models do not require pre-defined features for developing a learning algorithm. The features are extracted from the images automatically. Features that are useful for a classification task are strengthened

by the learning algorithm. Such feature learning is part of the deep learning paradigm (Skansi, 2018). Such systems for computer vision can carry out many complex tasks including medical image analysis or plant identification.

These systems work differently from human vision. Since CNNs start from very small features and analyse neighbouring pixels, they primarily perceive texture. Based on CNNs and other models, many pre-trained systems are available for research off the shelf, for example the Yolo system or the DETR for object identification.

Multimodal models integrate text and visual information, bridging a gap between these modalities. People search images mostly based on text, and often descriptions such as captions in the form of text are available. Images and text appear together in multimodal document formats. Joint processing of image and text can exploit these semantic connections. Deep learning methods process both modalities and learn the relations between them through numerous examples in a training set. This process leads to a common representation space in which similar embeddings stand for the word elephant and for pictures of elephants.

The quality of such multimodal models depends on the data and the training procedure. The Contrastive Language-Image Pre-training (CLIP) system has achieved very high performance for the first time with 400 million pairs of related texts and images collected from the web. CLIP was trained to select the most suitable image from over 30,000 texts (Radford et al., 2021). More advanced system such as VisualBERT (Li et al., 2019) intend to learn which

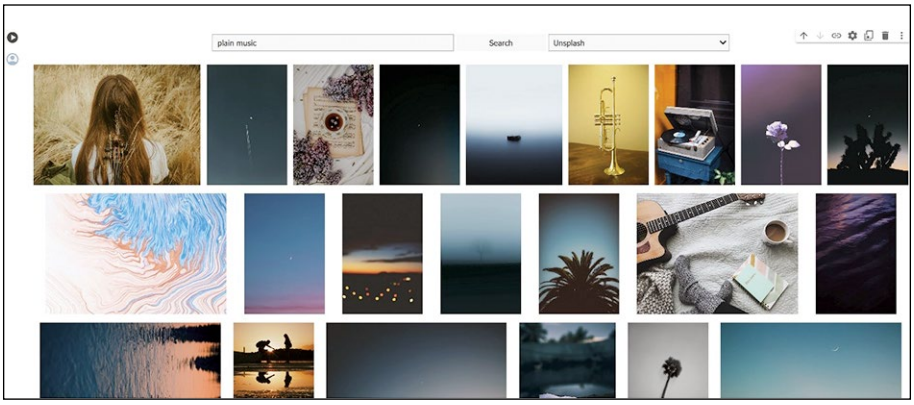

Figure 1. *Result for the query "plain music" in an open collection*

parts of the patterns of an image relate to which parts of the text. Multimodal models bring the semantics of language to the visual domain and can be used for more than just describing the obvious content of an image. This bridge between text and image can be used for several purposes, for example disambiguation of text by using images (Diem, Im & Mandl, 2023). The potential of multimodal systems can be seen in a search within a collection using CLIP. Even a search for "plain music" yields results without any metadata because textual and visual semantics are interwoven (see Figure 1).

3. Image Analysis in Cultural Heritage

The *Iconic Turn* led to a stronger orientation of cultural studies towards visual material. Research with images and visual material has established itself within the humanities beyond classic image sciences such as art history. Nevertheless, the development of appropriate tools and methods for Distant Viewing, which stands for the automatic analysis of large amounts of objects and visual data (also considering architecture and movies) with AI algorithms is still an emerging research field. Although the first conference on computers and arts dates back to as early as 1968 (Pratschke, 2018), automatic processing of mass data is not widely established. According to Münster and Terras (2020), the work on image and object analysis in DH can be broadly categorized into the following areas: image analysis (patterns in large scale collections), perception-based techniques, spatial modelling, and visualization. This overview emphasizes image analysis systems as well as the visualization of functions and results in these systems. More specifically, image analysis can be divided into these areas:

- Comparison of images,
- Comparison of scenes, objects, or details (semantic),
- Detailed analysis of deviations between variants,
- Comparison of technical features (in the painting or construction process),
- Representations for searching large scale collections to find structural similarities (Bell & Ommer, 2018).

There are several attempts to facilitate visual access to large amounts of pictures by visualization techniques such as miniaturization in interfaces. A concept developed by Manovich (2013) enables insights by plotting a large number of images in small scale. This enables the observation of colour patterns by providing numerous thumbnails in a meaningful way. This idea was applied to the cover images of journals, allowing viewers to see several decades of the editions (Manovich 2013).

The classification of artists is a typical application. Good performance has been achieved for the author identification of graphic novels, which also shows that a concept such as author style can be captured (Dunst & Hartel, 2018). An error analysis of misclassified items for a portrait set can reveal insights about the collection under consideration (Diem & Mandl, 2023).

A satisfying performance has been achieved for the classification of printing technology for images in historical children's books (Im, Kim & Mandl, 2022). The identification of objects within images or illustrations is also seen as a highly relevant processing technology for art history (Impett & Offert, 2023). Although many models have been developed, they do not always reach a high performance for historical content. Since such models are trained to recognize objects in realistic and modern photos, they may not always find objects in artworks (Kim, Im & Mandl, 2024). Very different results can be expected even within, for example, various books of the same genre (Mitera et al., 2021).

However, one needs to consider that concepts in DH are not always clearly defined but fuzzy and cannot be identified by recognizing one object type. This is especially true for high level concepts such as beauty (Cetinic, Lipic & Grgic, 2019). Although attempts to define aesthetic concepts have been made, many issues in human perception are still not fully understood. A useful overview for the analysis of aesthetic concepts is provided by Brachmann and Redies (2018).

The possibilities to process images based on several levels range from simple colour similarity analysis to matching based on object detection, as well as similarity based on classification outcomes. These classification systems can be trained on genre, artist, or aesthetic concepts. This enables the retrieval of images based on matching paradigms at different complexity. The

methods mentioned above can also be applied to movies (e. g., Schmidt & Kurek, 2022).

Innovations from digital disciplines need to be adopted by DH, and the specific demands and special requirements of scholarly topics in the humanities need to stir new developments in the digital domain. Given the multiplicity of options arising from innovation in computer vision, there is a great need for interactive functions of systems (Bell & Ommer, 2018).

4. Image-Based Search and Analysis Tools in Cultural Heritage

General Web search engines can also be used to search for images and to find similar ones. Specific image services such as GettyImages or Pinterest enable access to huge collections of curated or user generated visual content. Many systems have been proposed specifically for analysing historical images either for art history or within illustrations in books. Most of them employ search within metadata such as, for example, the image search within Europeana.eu.

A system that allows exploration based on similarity from a given image is the online search of the Bavarian State Library (Bayerische Staatsbibliothek München, https://www.bsb-muenchen.de/sammlungen/bilder/). The system works on traditional image processing systems without machine learning or deep learning methods. The algorithm extracts a descriptor, which relies on edge and colour values from individual sections of an image. These sectional values are assembled into a histogram similar to the MPG-7 standard (Brantl et al., 2017). An advantage of the system is this robust and fast processing and the large image base of the online database.

A search system developed at the University of Oxford allows similarity search for a part of an image. It can be helpful to find reused images (see Figure 2).

Deep representation models have been used in a similarity matching tool that works on a collection of visual material within children's and youth literature (Helm et al., 2021). Children's books typically contain more images than adult books. As a consequence, they are of special interest for an analysis of images. Although non-fiction books for children enjoy rather

great popularity at present as well as their modern (inter)medial and (inter) modal forms, the resonance within the research area of children's literature with respect to this genre is quite limited. Illustrated books have played a significant role in knowledge dissemination. The declining production costs for printed images have led to a growing exposure of more and more people to rich visual resources.

A system has been developed in order to analyse illustrations in 19th-century books by applying modern deep learning computer vision methods (Im, 2024). The system finds exact copies but also more fuzzy levels of similarity. Furthermore, the system implemented uses object detection to au-

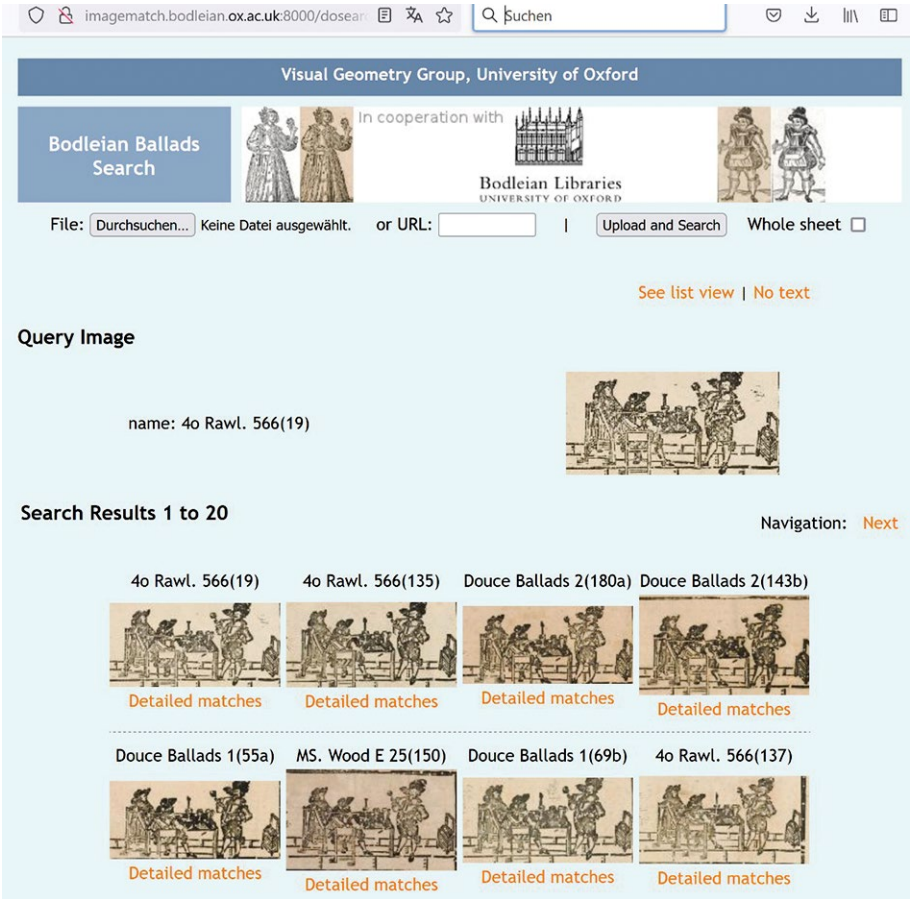

Figure 2. *Search for part of an image in the ImageMatch system at http://ballads.bodleian.ox.ac.uk/*

tomatically enrich the metadata available for searching the images. It allows not only a search for the objects, which were recognized, but also the specification of their location within a 3x3 grid over each image. The user can search for an image with two humans in the middle and a bird on the top left. The composition of the image apart from visual similarity, for example based on colours, becomes explorable (Im, 2024). With further work, this functionality could be extended to high level scenes such as teaching or family scenes.

Similar goals are pursued by other search systems based on deep machine learning technology. The iArt system uses deep embedding and, among others, the CLIP representation to map between text and image (see Figure 3). IArt allows the search for both concrete and high level concepts (Springstein et al., 2021). It comprises a powerful clustering system that orders the objects into meaningful groups. Furthermore, it allows setting parameters for the search.

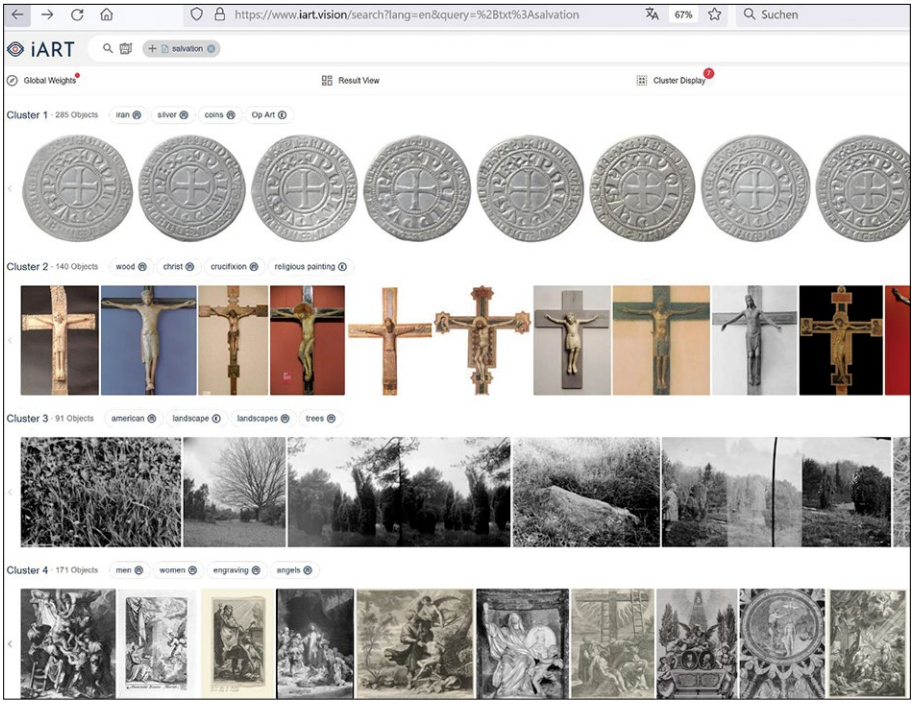

Figure 3. *Search result for the high-level concept salvation in iArt.vision*

The system Imgs.ai is also based on deep multimodal models and includes CLIP (Impett & Offert, 2023). Due to the multimodal capabilities of the background models, even abstract search terms such as *rhythm* or *energy* lead to potentially relevant results. Furthermore, Imgs.ai offers options for refinement of the result set and the provision of positive and negative examples. This leads to an iterative interaction process that increases controllability through experimentation (see Figure 4).

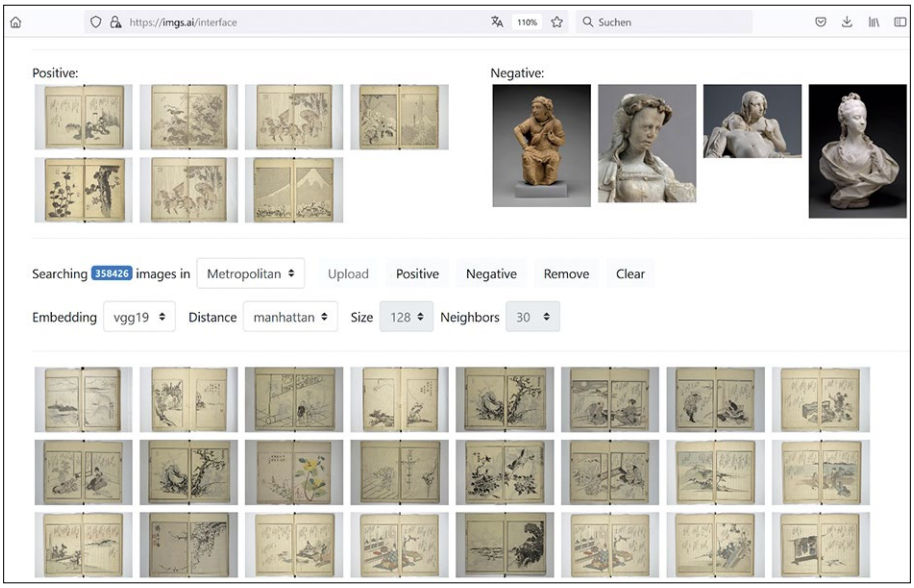

Figure 4. *Search result in Imgs.ai after providing positive and negative examples*

Similar to iArt, the system Imgs.ai enables the user to change some matching parameters such as the embeddings and distance functions. However, these configuration options are not aligned with the users' language and require technological knowledge. Furthermore, not even an expert in image embeddings would be able to predict how the selection of a model would affect the result.

Controllability and predictability are essential criteria for good user interfaces (ISO, 2019). Usability and a positive user experience require that users feel in control and can, to some extent, predict the system's output based on their inputs. This is a common problem for advanced AI systems in general and for the systems shown above in particular. It is necessary to provide

more control for the user. One approach for portraits allows the selection of specific embeddings using terminology from the domain (Diem, 2023).

Similar work is also conducted for improving control for users of generative models. While language prompts offer little control over results of language generation models such as ChatGPT or image creation models such as StableDiffiusion, advanced systems provide greater control over the visual result (e. g., Zhang, Rao & Agrawala, 2023).

5. Conclusion

The availability of large collections of visual material as well as modern computer vision systems have opened new opportunities for analysis in DH. Standards such as the Open Archives Initiative and the IIIF standard have led to tools for mass download of images and digitized books. Nevertheless, there is still a lack of standards for storing deep learning representations and making them accessible to researchers. That might significantly lower the barrier for entry into, for example, image analysis.

For the analysis of visual content, object identification, similarity analysis, and self-trained classification are the most often used tasks. If they are applied in interesting combinations, and if they are geared towards the expectations of scholars in cultural studies, they can reveal relevant insights. However, AI powered tools need to enable much more control over the search process for researchers and other users in order to produce results that can really satisfy complex tasks. Otherwise, users might not accept these tools.

Furthermore, similar to other AI applications, image search tools need to incorporate explainability in order to show how results were obtained. Explainable AI models for images, which are currently developed (Kamakshi & Krishnan, 2023), may also support users in understanding systems and their output more efficiently.

Algorithm aversion has often been observed in other domains when results were not satisfactory. One way to alleviate this danger lies in offering configuration options (Dietvorst, Simmons & Massey, 2018), which need to build on users' domain expertise and terminology. Systems providing

scene description (Im, 2024) or transparent selection of embeddings (Diem, 2023) offer the potential to better configure and steer deep learning image systems towards user intentions (Heuwing et al., 2016).

Acknowledgements

This research was partially funded by the Fritz Thyssen Foundation within the project *Distant Viewing* as well as the Volkswagen Foundation within the project *PorträtApp*.

References

BARZ, B.; DENZLER, J. (2021). Content-Based Image Retrieval and the Semantic Gap in the Deep Learning Era. In: *Pattern Recognition. ICPR International Workshops and Challenges. ICPR.* Springer, Cham. https://doi.org/10.1007/978-3-030-68790-8_20

BELL, P.; OMMER, B. (2018). Computer Vision und Kunstgeschichte–Dialog zweier Bildwissenschaften. In: *Computing Art Reader: Einführung in die digitale Kunstgeschichte* (pp. 1, 61-75). https://doi.org/10.11588/arthistoricum.413.c5769

BRACHMANN, A.; REDIES, C. (2017). Computational and experimental approaches to visual aesthetics. In: *Frontiers in computational neuroscience*, 11, 102. https://doi.org/10.3389/fncom.2017.00102

BRANTL, M.; CEYNOWA, K.; MEIERS, T.; WOLF, T. (2017). Visuelle Suche in historischen Wer-ken. *Datenbank-Spektrum*, 17(1): p 53-60. https://doi.org/10.1007/s13222-017-0250-0

BULLIN, M.; HENRICH, A. (2020). Die inhaltsbasierte Bildsuche und Bilderschließung: Ansätze und Problemfelder. In: *Bilddaten in den Digitalen Geisteswissenschaften*. https://doi.org/10.13173/9783447114608.001

CETINIC, E.; LIPIC, T.; GRGIC, S. (2019). A deep learning perspective on beauty, sentiment, and remembrance of art. In: *IEEE Access*, 7: 73694-73710. https://doi.org/10.1109/ACCESS.2019.2921101

DIEM, S. (2023). Entwicklung und Evaluierung einer flexiblen Bildähnlichkeitssuche für Expert*innen im Bereich der Porträtforschung. In: *Proceedings des 17. Internationalen Symposiums für Informationswissenschaft (ISI 2023)* Chur, Schweiz, 7.–9. November 2023. Doctoral Consortium.

DIEM, S.; MANDL, T. (2023). Automatic Classification of Portraits: Application of Transformer and CNN Based Models for an Art Historic Dataset. In: *Proceedings of the LWDA Workshops*. Marburg, Germany, 09.-11. October. http://ceur-ws.org

DIEM, S.; IM, CJ.; MANDL, T. (2023). University of Hildesheim at SemEval-2023 Task 1: Combining Pre-trained Multimodal and Generative Models for Image Disambiguation. In: *The 61st Annual Meeting of The Association for Computational Linguistics* (ACL). https://doi.org/10.18653/v1/2023.semeval-1.18

DIETVORST, B.J.; SIMMONS, J.P.; MASSEY, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (Even slightly) modify them, In: *Management Science, Institute for Operations Research and the Management Sciences*, Bd. 64, Nr. 3: 1155–1170. https://doi.org/10.1287/mnsc.2016.2643.

DONIG, S. (2023). Der Digital Turn in den Geisteswissenschaften und seine Implikationen für Gedächtniseinrichtungen. In: *Bibliothek Forschung und Praxis*. https://doi.org/10.1515/bfp-2023-0046

DUNST, A.; HARTEL, R. (2018). Auf dem Weg zur Visuellen Stilometrie: Automatische Genre- und Autorunterscheidung in graphischen Narrativen. In: *Kritik der digitalen Vernunft. 5. Tagung „Digital Humanities im deutschsprachigen Raum"* http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf

HELM, W.; SCHMIDELER, S.; IM, C.; MANDL, T.; KOLLMANN, S.; MÜLLER L. (2021). Wie sich die Bilder ähneln: Vom Zufallsfund zur systematischen Forschung im Bereich der automatisierten Bildähnlichkeitssuche. In: *ZfdG – Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.26298/melusina.8f8w-y749-wsdb

HEUWING, B.; MANDL, T.; WOMSER-HACKER, C. (2016). User-Centered Design and Evaluation of Text Analysis Tools in a Digital History Project. In: *Proceedings of the Association for Information Science and Technology (ASIS&T)*. Vol. 53, (1): 1–10. https://doi.org/10.1002/pra2.2016.14505301078

IM, C. (2024). Deep Neural Networks for Illustration from 19th-Century Children's and Young Adults Literature: Development and Evaluation of an Illustration Information System. University of Hildesheim, Dissertation. To appear.

IM, C.; KIM, Y.; MANDL, T. (2022). Deep learning for historical books: classification of printing technology for digitized images. In: *Multimedia Tools and Applications*, 81(4): 5867-5888. https://doi.org/10.1007/s11042-021-11754-7

IMPETT, L.; OFFERT, F. (2023). *There is a Digital Art History?* arXiv preprint arXiv:2308.07464.

ISO – International Organization for Standardization (2019). ISO 9241-210:2019. *Ergonomics of human-system interaction. Part 210: Human-cen-*

tred design for interactive systems. https://www.iso.org/standard/77520.html

KAMAKSHI, V.; KRISHNAN, N. C. (2023). Explainable image classification: the journey so far and the road ahead. In: *AI*, 4(3): 620–651. https://doi.org/10.3390/ai4030033

KIM, Y.; MANDL, T.; IM, C.; SCHMIDELER, S.; HELM, W. (2020). Applying Computer Vision Systems to Historical Book Illustrations: Challenges and First Results. In: *Post-Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN)* Riga. Ceur_ws. (pp. 255-260). http://ceur-ws.org/Vol-2865/poster7.pdf

KIM, Y; IM, C; MANDL, T. (2024). Object Detection in Historical Images: Transfer Learning and Pseudo Labelling. In: *ACM Journal on Computing and Cultural Heritage (JOCCH)*. ACM DL.

LI, L. H.; YATSKAR, M.; YIN, D.; HSIEH, C. J.; CHANG, K. W. (2019). VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint*. https://arxiv.org/abs/1908.03557

MANDL, T. (2015). Text Mining. In: *Encyclopedia of Information Science and Technology* (pp. 1923-1930). Idea Group Reference: Hershey et al.

MANOVICH, L. (2013). Museums without Walls, Art History without Names: Visualization Methods for Humanities and Media Studies. In: *Oxford Handbook of Sound and Image in Digital Media* (pp. 253-278). Oxford.

MITERA, H.; IM, C.; MANDL, T.; WOMSER-HACKER, C. (2021). Objekterkennung in historischen Bilderbüchern: Eine Evaluierung des Potenzials von Computer Vision Algorithmen. In: *BildWissen – KinderBuch: Historische Sachliteratur für Kinder und Jugendliche und ihre digitale Analyse* (pp. 137-150). J.B. Metzler. https://doi.org/10.1007/978-3-476-05758-7_9

MÜNSTER, S.; TERRAS, M. (2020). The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures. In: *Digital Scholarship in the Humanities*, 35(2), 366-389. https://doi.org/10.1093/llc/fqz022

PRATSCHKE, M. (2018). Geschichte und Kritik digitaler Kunst- und Bildgeschichte, In: *Computing Art Reader: Einführung in die digitale Kunstgeschichte* (pp. 20-37). Heidelberg: arthistoricum.net. https://doi.org/10.11588/arthistoricum.413.c5767

RADFORD, A.; KIM, J. W.; HALLACY, C.; RAMESH, A.; GOH, G., AGARWAL, S., ... SUTSKEVER, I. (2021). Learning transferable visual models from natural language supervision. In: *International Conference on*

Machine Learning (pp. 8748-8763) PMLR. [accessed: 2024–06–15]. Available at: http://proceedings.mlr.press/v139/radford21a

SAHLE, P. (2016). What is a scholarly digital edition?. In: *Digital scholarly editing: Theories and practices*, 1: 19-39.

SCHMIDT, T.; KUREK, S. (2022). Der Einsatz von Computer Vision-Methoden für Filme-Eine Fallanalyse für die Kriminalfilm-Reihe Tatort. In: *8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHd 2022, Potsdam, Germany*, March 7 - 11, 2022. https://doi.org/10.5281/zenodo.6310.528167

SKANSI, S. (2018). *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer. https://doi.org/10.1007/978-3-319-73004-2

SPRINGSTEIN, M.; SCHNEIDER, S.; RAHNAMA, J.; HÜLLERMEIER, E.; KOHLE, H.; EWERTH, R. (2021). iART: A Search Engine for Art-Historical Images to Support Research in the Humanities. In: *Proceedings 29th ACM International Conference on Multimedia.* (pp. 2801-2803). https://doi.org/10.1145/3474085.3478564

STORK, D. (2024). Computer vision, ML, and AI in the study of fine art. In: *Communications of the ACM*, 67(5): 68–75. https://doi.org/10.1145/3633454

TOMIĆ, M. (2018). Project Digitization, Bibliographic Description and

Research of Texts Written in Glagolitic, Croatian Cyrillic and Latin Scripts Until the End of the19th Century in the Zadar and Šibenik Area (Written Heritage). In: *Cataloging & Classification Quarterly*, https://doi.org/10.1080/01639374.2018.1491438

ZHANG, L.; RAO, A.; AGRAWALA, M. (2023). Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (pp. 3836-3847).