Analysis of large image Collections: Potential bias and requirements for future information infrastructure

Thomas Mandl, Chanjong Im, Sebastian Diem (University of Hildesheim, Germany)

Abstract

Modern image analysis for historical print products can be analysed quantitatively due to large digitalization efforts in the last decades. This contribution shows progress in image analysis and provides an overview on the state of the art of application of such AI methods to historic images. Finally, it is shown that several kinds of bias may occur and require care during analysis.

Introduction: Digitized Collections

Nowadays, many digital collections are being provided by libraries and other institutions. More and more books, journals and newspapers are digitized and accessible online. This development makes also images in historical print available for mass analysis. This opens new potentials for scholars for quantitative research on cultural trends as they are expressed in images. Within in the humanities there is an increased use of digital tools, which, together with the traditional qualitative-hermeneutic methods form the so-called “mixed methods“. In many research areas related to images, the need to combine qualitative, analytic-interpretative oriented methods with digital methods becomes apparent and has great potential.

However, the support for mass analysis is still limited (Helm et al. 2019). Research infrastructures like DARIAH contain a variety of methods of digital text analysis. So far these research environments do not yet reflect the growing importance of visual information. Standards like 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 e.g. image analysis.

For further progress, an intensive exchange between experts from different subject areas is required – i.e. art history, books studies, history, literature research, media studies, as well as computer science, information science and image processing.

Image processing and Analysis

In the last years, considerable progress has been made in image processing, especially through approaches of so-called Deep Learning. These data driven methods have performed well for many tasks and have often replaced traditional image processing based e.g. on color and shape analysis.

These algorithms also learn aspects of the pictures need to be analysed for the best results. Such feature learning is typical for deep learning. A prototypical system is a Convolutional Neural Network (CNN) which combines many simple neurons as processors into complex architectures. A basic CNN is composed of recurring sets of two layers: a convolution layer and a pooling layer. The CNN combines pixels locally and by working through many layers, more complex features can be extracted (Skansi 2018). Based on the features, diverse classification tasks can be learned by the neural networks. Based on CNNs and other complex

models, many pre-trained systems are available for research off the shelf, e.g. the Yolo system for object identification (Redmon et al. 2016).

The availability of pre-trained networks has driven research in image analysis, but for DH no intermediary results and processed data are available for facilitate further research (e.g. representation of images in vectors for further processing). Infrastructure currently being under development can improve the situation, but the community needs to discuss and express their needs. Of particular interest is the national initiative in Germany for culture (NFDI4culture, Altenhöner et al. 2019).

Automatic Analysis of large image collections

Since the Iconic Turn, research with images and visual material has itself established within the Humanties beyond the classic image sciences. 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, a first conference on computers and arts dates back to as early as 1968 (Pratschke 2018), the automatic processing of mass data is still limited.

The work on image and object analysis in DH can be very coarsely categorized in the following classes:

- Visualization approaches
- Detailed analysis of small sets of objects and images
- Search systems, often based on similarity
- Classification systems for large amounts of objects and images
- Analysis systems for identifying trends or other patterns

An alternative form of classifying research could be from basic operations like OCR (e.g. Neudecker et al. 2019) and layout analysis (e.g. Lehenmeier et al. 2020) to processing high level concepts like art period (Saleh & Elgammal 2015), beauty (Cetinic et al. 2019) or extremism (Fredheim et al. 2014). A useful overview for analysis of aesthetic concepts is provided by Brachmann & Redies (2018). However, in this short overview, we provide examples for each topic based on the categories mentioned above. This also means that domains like art, literature studies and others are not separated in the following sections.

Visualization: There are several attempts to facilitate the visual access to large amounts of pictures through miniaturisation and better user interfaces. Well known are the system developed by (Manovich 2013) which allow some insight e.g. into color patterns by plotting many thumbnails in meaningful ways. Another more recent approach is the neural neighbor system[1] (Resig 2013) which uses the output of a deep neural network for clustering and visualizing similarity within a 2-dimensional space. Images close to each other are supposed to have similar features. These features have been learned and might not be always interpretable. The identification of objects within images or illustrations can be seen as a subset of this task (Crowley & Zisserman 2014). However, one needs to consider that concepts in DH are not always clearly defined but fuzzy.

Innovations from digital disciplines need to be adopted by DH and the specific demands and special requirements of DH problems need to stir new developments in the digital domain. The workshop can be a step towards intensifying this exchange and cooperation.

Bias regarding Analysis of Image Collections

[1] https://dhlab.yale.edu/neural-neighbors/

Image collections and their analysis exhibit several kinds of bias. Most often, collections have a skewed distribution. One important issue for machine learning methods is the amount of training data in the classes to be trained. Unbalanced datasets lead to low performance in particular for the less represented classes.

In a classification experiment for printing technology within a large collection of children books, we experienced that wood cut which was less represented in the dataset was less well classified (Im et al. 2022). This may have several consequences. If the visual traditions are related to printing technology within a collection, the classification quality for some technologies may be low and lead to biased results.

In the collection of children books, we also observed a skewed distribution for the publishing houses. When creating a classifier for publishing houses based only on images, the less well represented publishing houses are very often confused with the dominating classes.

When applying pre-trained object detection we observe two forms of skewed distributions. The pre-trained models were trained on modern photos. They perform well for objects which exist in the modern world. The application of object detection systems like Yolo reveals another bias. The objects found in typical children books are limited to several dozens of objects and mostly people are found (Mitera et al. 2020).

Face detection and face recognition have become established fields in computer vision. The advancements of neural networks led to robust frameworks. A variety of applications are based on large pre-trained models consisting of millions of realistic images. In the experiments reported here, we use a state-of-the-art face detection model called OpenFace to analyse the applicability of those models on printed portraits from the early modern period until the 19th century. From a collection of 27.000 well-curated digitalized portraits we selected a subset and found that the older the portrait is, the less well the face is detected.

Even the amount of images in digitized collections may vary greatly. We show how collection standards can lead to large differences which require careful interpretation when conducting automatic analysis.

References

Altenhöner, Reinhard et al. (2019). Fokusthemen und Aufgabenbereiche für eine Forschungsdateninfrastruktur zu materiellen und immateriellen Kulturgütern. Living Document der NFDI-Initiative NFDI4Culture. Working paper Open Access. DOI 10.5281/zenodo.2763575

Arnold, Taylor B. & Tilton, Lauren (2019). Distant viewing: Analyzing large visual corpora, Digital Scholarship in the Humanities 34, Issue Supplement_1, December 2019, i3–i16 (https://www.distantviewing.org/pdf/distant-viewing.pdf).

Cetinic, E., Lipic, T., & Grgic, S. (2019). A deep learning perspective on beauty, sentiment, and remembrance of art. IEEE Access, 7, 73694-73710.

Crowley, E. & Zisserman, A. (2014). The State of the Art: Object Retrieval in Paintings using Discriminative Regions. Proc. British Machine Vision Conference. BMVA Press.

Bell, P., & Ommer, B. (2017). Kunst messen, Pixel zählen? – Die Zusammenarbeit zwischen Kunstgeschichte und Computer Vision oszilliert zwischen quantitativen und hermeneutischen Methoden. In Messen und Verstehen in der Wissenschaft. JB Metzler, Wiesbaden. pp. 225-236.

Brachmann, A., & Redies, C. (2017). Computational and experimental approaches to visual aesthetics. Frontiers in computational neuroscience, 11, 102. https://doi.org/10.3389/fncom.2017.00102

Fredheim, R., Howanitz, G., & Makhortykh, M. (2014). Scraping the monumental: Stepan Bandera through the lens of quantitative memory studies. Digital Icons: Studies in Russian, Eurasian and Central European New Media, 12, 25-53.

Helm, Wiebke; Mandl, Thomas; Putjenter, Sigrun; Schmideler, Sebastian; Zellhöfer, David (2019): Distant Viewing Forschung mit digitalisierten Kinderbüchern: Voraussetzungen, Ansätze und Herausforderungen. In: B.I.T.online – Zeitschrift für Bibliothek, Information und Technologie. Heft 2, S. 127-134. https://www.b-i-t-online.de/heft/2019-02-index.php

Im, Chanjong; Kim, Yongho; Mandl, Thomas (2022): Deep Learning for Historical books: Classification of Printing Technology for Digitized Images in: Multimedia Tools and Applications (MTAP) 81, pp. 5867–5888. DOI: 10.1007/s11042-021-11754-7

Lehenmeier, C., Burghardt, M., & Mischka, B. (2020). Layout Detection and Table Recognition–Recent Challenges in Digitizing Historical Documents and Handwritten Tabular Data. In International Conference on Theory and Practice of Digital Libraries. pp. 229-242. Springer, Cham.

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K. M., Hartmann, V., & Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. In Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. pp. 53-58.

Manovich, Lev (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. Ed. Carol Vernallis, Amy Herzog and John Richardson, Oxford, pp. 253-278.

Mitera, Hannah; Im, Chanjong; Mandl, Thomas; Womser-Hacker, Christa (2021): Objekterkennung in historischen Bilderbüchern: Eine Evaluierung des Potenzials von Computer Vision Algorithmen. In: Helm, Wiebke; Schmideler, Sebastian (2021): BildWissen – KinderBuch: Historische Sachliteratur für Kinder und Jugendliche und ihre digitale Analyse [Studien zu Kinder- und Jugendliteratur und -medien] J.B. Metzler. S. 137-150. https://doi.org/10.1007/978-3-476-05758-7_9

Pratschke, Margarete (2018). Geschichte und Kritik digitaler Kunst- und Bildgeschichte, in: Kuroczyński, Piotr, Bell, Peter und Dieckmann, Lisa (Hrsg.). Computing Art Reader: Einführung in die digitale Kunstgeschichte, Heidelberg: arthistoricum.net(Computing in Art and Architecture, Band 1), pp. 20-37. https://doi.org/10.11588/arthistoricum.413.c5767

Redmon, J.; Divvala, S.; Girshick, R. & Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788.

Saleh, B. & Elgammal, A. (2016). Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. International Journal for Digital Art History, (2).

Skansi, S. (2018). Introduction to deep learning: From Logical Calculus to Artificial Intelligence. Springer.