Three perspectives on a collaborative attempt to use computer vision techniques to automatically classify historical newspaper images

Martijn Kleppe[1][0000-0001-7697-5726] Thomas Smits[2][0000-0001-8579-824X] Willem Jan Faber[3]

[1+3] National Library of the Netherlands, The Hague, the Netherlands
[2] Utrecht University, Drift 6, Utrecht, The Netherlands
Martijn.Kleppe@KB.nl

Abstract. In the last couple of years, scholars in the Humanities have started to explore the possibilities of the large-scale analysis of images. This development can be linked to the increasing availability of large visual datasets, the increase in computing power, and the development of new techniques, such as convolutional neural networks. However, there are no one-size-fits all researchers that are able to gather the right data, apply the new techniques, and analyze the results in meaningful ways. In this paper we present the collaboration of a Humanities researcher, a Research Software Engineer and Digital Scholarship Advisor to explore how new computer vision techniques can be used to automatically classify images extracted from a large collection of digitized historical newspapers. We will present the outcomes of our research and share the lessons we learned from our collaboration. First we will discuss the experiences of the Humanities researcher. Second we will discuss the lessons we learned from a technical perspective. Third, we will elaborate on the institutional perspective of the National Library of the Netherlands (KB) as a data provider but also as full partner of the research project. We will end with a reflection on the broader strategic role of heritage institutes as research partners to stimulate, collaborate and to preserve results of research projects in a sustainable manner.

Keywords: Computer Vision, Distant viewing, Digitized newspapers

1 Introduction

Although the Digital Humanities have traditionally focussed on the large-scale analysis of texts (Nicholson, 2013), recent years have seen an upsurge in research that focuses on images. This move to the visual can be explained by the increasing availability of visual datasets (Russakovsky et al., 2015) and the techniques necessary to analyse them. Examples of this kind of research include the work of Seguin (Seguin et al., 2017) who focuses on automatic visual pattern detection across iconographic collections and the work of King and Leonard (2017) on colometrics, facial detection and neural network-based visual similarity. The International Digital Humanities conferences also displayed a growing interest for non-textual sources (Weingart, 2016), reflected in the workshops on computer vision organised by the Special Interest Group Audiovisual Material in Digital Humanities in 2017 (Kleppe et al., 2017) and 2018 (Tilton et al., 2018).

In the Netherlands we see a similar tendency. First, datasets of digitised visual sources are becoming more available. The National Library of the Netherlands (KB) offers access to a large collection of digitised newspapers on their portal www.delpher.nl, allowing full-text searches through all data and drill down the results by applying filters such as period, region or type of article. Furthermore, researchers can get access to all digital sources through the library's Dataservices and APIs and experimental datasets at the KB Lab, such as the KBK-1M Dataset (Kleppe et al., 2016). To stimulate the use of these datasets, understand the needs of researchers, and improve the library's services, the KB has set up the researcher-in-residence program (Wilms, 2017; Boekestein, 2017). This allows researchers to work part time at the Research Department of the KB for six months, together with one of KB's Research Software Engineers. During their project they are also assisted by a Digital Scholarship Advisor and several metadata and collection specialists.

In 2017, two researcher-in-residence projects were carried out to explore the possibilities of applying new computer vision techniques to analyse digitised historical newspapers. Melvin Wevers explored visual similarity search on newspaper advertisements (Wevers and Lonij, 2017). In this paper, we will focus on the second project by Thomas Smits, on classifying newspaper images. We will first describe the Humanities research question, followed by our technical approach and the project's results. The final part of the paper is a reflection on the collaboration between the Humanities Researcher and the Research Software Engineer. We will also reflect on the role of the KB as a data provider, but also full research partner.

2 Humanities research question: Fin de siècle visual news culture

The visual representation of news events is generally connected to the technological progress of photography (Gervais and Morel, 2017). The so-called half-tone revolution of the early 1880s, enabling the massive reproduction of photographs in print media, is seen as forming the basis for our current visual news culture. Several historians of nineteenth-century media have challenged this narrative (Gitelman and Pingree, 2003). Hill and Schwartz (2015) propose a contingent history of 'news pictures' as a separate 'class of images', which not solely focuses on photographic technology, but on the discourse surrounding them (p. 3). In relation to this recent theoretical development, several studies have demonstrated that photography was not the first medium used to visually represent the news. From the early 1840s, illustrated newspapers disseminated news pictures on a massive scale and developed a discourse of objectivity, based on eyewitness accounts, which would be adapted and used for photographs later in the century (Barnhurst and Nerone, 2000; Gervais, 2010; Park, 1999).

Although the visual representation of the news did not start with photography, the pre-eminence of this medium is clear in the twentieth century (Gervais and Morel, 2017; Kester and Kleppe, 2015). It follows that the turning point between the use of illustrations and photographs as the preferred medium to represent the news is a critical moment in the history of modern visual news culture. Most commonly, researchers have presented this point as a watershed, located at the publication of the first photograph of a news event in a newspaper (Kester and Kleppe, 2015). However, case studies from a media archaeological perspective, suggest a relatively long transitional

period in which illustrations and photographs coexisted and competed as authentic, objective visual representations of the news (Keller, 2013; Steinsieck, 2006). It remains unclear when photography exactly achieved its pre-eminence and why this happened.

The earlier reliance on case studies to describe the transitional phase is understandable, as, in pre-digital times, a 'distant reading' (Moretti, 2015) of the large number of images published in newspapers was all but impossible. Using several computer vision techniques, our project aspired to shed more light on this important debate by analysing pictures of the news in Dutch newspapers from a distance ('distant viewing') and on a large scale. Our main research questions were: When did Dutch newspapers start to use illustrations? And when did they switch to using photographs as the primary visual medium? More generally, we hoped to explore how these techniques could be used to analyse large collections of visual historical material.

3 Technical approach: convolutional neural networks

As most DH research, our project faced two main challenges: data collection and data analysis. Within Delpher, users can select facets to drill down to specific results. Upon selecting 'Illustration with caption' they will only get articles that contain an image. However, the results will not only contain photographs, but also cartoons, drawings, weather reports and even graphic displays of chess problems. Since this would not suffice to answer our main research question, we had to find new ways to classify the images found in newspapers.

Concerning data collection part, our project could build on the PhoCon project of Elliott & Kleppe (Kleppe et al., 2016), which created a database containing images extracted from Delpher's newspapers. However, the result of this project, the KBK-1M(illion) database only contained images from the period 1923-1930. Furthermore, we found that not all images in the period of our research (1860-1923) were correctly classified as 'captioned illustration' by the OCR company. Therefore, new code was needed to harvest all the images from digitised newspapers. We found that in the XML files (ALTO) the code-line 'imageblock' denotes images. Around 1900, Dutch newspapers contained many small images, like the often-recurring illustrations used at the beginning of a specific section, or small images that accompanied advertisements in newspapers. Because we were mainly interested in images of the news, we decided to only include images that could be related to newspaper articles (via the XML file), exclude images of advertisements, and discard all the images with a file size smaller than 30KB. We ended up with 313K images for the period 1923-1930.

We classified these images using a three-step pipeline. First of all, we used Adam Geitgey's facial recognition API, built using the Dlib's facial recognition library, to recognize faces on the images (Geitgey, 2017). In the second step, using the 'Tensorflow for poets' method, we applied an Inception-V3 convolutional neural network to recognize nine different categories (buildings, cartoons, chess, crowds, logos, maps, schematics, sheet music, and, weather reports).[1] Although the creators of this method recognize that it will be outperformed by a full training run, it is surprisingly effective (see below for performance) and does not require GPU hardware. We used training sets of around forty images for every category. For the final classification step, we

[1] https://codelabs.developers.google.com/codelabs/tensorflow-for-poets/#0

asked Leonardo Impett, a digital art historian at the Bibliotheca Hertziana, to build a convolutional neural network that could recognize if images were either drawings or photographs. His network focuses on the lower-layers of the network and a support vector machine (SVM) divides the images into photographs and illustrations.

The four-step classification pipeline resulted in the CHRONIC (Classified Historical Newspaper Images) database, which contains metadata for all the 313K newspaper images we extracted (Smits and Faber, 2018a). Based on this database, we created CHRONReader: a tool which allows users to search for images containing faces, one of the nine categories and being either illustrations or photographs (Smits and Faber, 2018b).

4 Results

Using computer vision we were able to analyse the images of Dutch newspapers on a large scale, or view them from a distance, and, as a result provide an answer to the main research questions: When did Dutch newspapers start to use illustrations? And when did they switch to using photographs as the primary visual medium? Fig. 1 depicts the publication of illustrations and photographs in Dutch newspapers between 1860 and 1930. The number of images in Dutch newspapers, both illustrations and photographs, increased noticeably in the early 1900s and peaked at the start of the 1920s. The number of photographs overtook the number of illustrations for the first time in 1927. This completed a development from nineteenth-century publications filled with letters, to pages filled with both images and text: the form of the newspaper we still know today.

On the one hand, the application of CNNs thus confirms the conclusions of earlier work, mentioned above, based on case studies. At the same time, vast digitized archives and new techniques like CNNs contribute to the construction of a exciting new overview of visual (news) culture, which allows for the analysis of trends and changes over an extended period of time. As Fig. 1 shows, the visual representation of the news took off in the earlier 1920s. Although earlier research noted and analysed the introduction of so-called ‘photo-pages’ in the 1920s using a limited set of sources (Broersma, 2014; Kester and Kleppe, 2015) the birds-eye view of the use of images in the entire Dutch press provides us with a new perspective on the magnitude of this watershed moment.

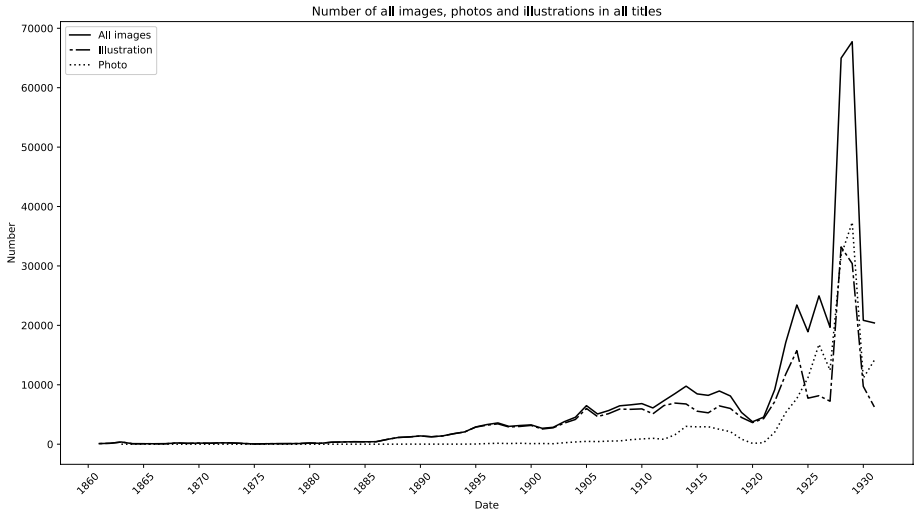

Fig. 1. Number of all images, photo's and illustrations in all digitized newspapers, 1860-1930

Next to the ability to view large collections of images from a distance, new computer vision techniques also provide direct access to visual content without having to refer to textual descriptions. In this sense the technique can be compared to OCR-technology, in that it provides users of digital archives with bottom-up access to sources (Nicholson, 2013).

For the KB, this project offered several results. First we gained more knowledge about the user needs of researchers who want to study the visual aspects of digital sources. Second, we got to know our data and metadata better. For example, resulting from the set-up of the metadata and the way this is created and stored at the KB, the creation of a dataset containing images and their captions turned out to be more complicated than expected. Third, due to the collaborative nature of the researcher-in-residence program, our research software engineer gained more knowledge about computer vision. Since libraries have been focused on texts for centuries, we nowadays mainly focus on Natural Language Processing techniques to analyze digital material. However, as we have learned from this and the PhoCon project, digital datasets also contain millions of images. Since libraries continuously want to improve access to their digital collections, they should focus on both textual and visual material. However, as we have learned from this project, this is far from an easy task. The assistance by Leonardo Impett to build a convolutional neural network to divide the images into photographs and illustration was e.g. fundamental for the end result of the project. Fourth, since the KB now has the knowledge on applying computer vision, we are taking steps to apply it on a large scale. On Delpher.nl users can select 'Illustrations with captions' but as we have described before, they then retrieve all sorts of images. The results of the CHRONIC project allows the KB to classify all images in historical newspapers to eventually implement an advanced selection option within Delpher to allow users also to select photographs, cartoons or even chess problems. However, scaling up the results

of this research project to the full collection will present several challenges in terms of computing power and infrastructure

5 Reviewing collaboration

For this project, we set up a team consisting of a Humanities researcher (Thomas Smits), a research software engineer (Willem Jan Faber) and digital scholarship advisor (Martijn Kleppe). The team met on a weekly basis to discuss the projects' progress, while the individual team members also regularly had bilateral meetings or were helped by KB's in house metadata and collection specialists. Since the Humanities researcher was researcher-in-residence, he was seconded for six months to the KB and was present in the KB for two days a week, which was very stimulating for the projects progress. He could easily get access to KB's in house experts who normally can only be contacted through KB's front office. In this way, he was able to get more easy access to (meta)data and more specialised knowledge about the data structure. Furthermore, the collaboration with the research software engineer allowed him to explore not only the data but also new techniques. Trained as a traditional historian, Smits was not used to working with innovative, and highly complex, digital methods of analysis, such as neural networks. Due to the intensive nature of the collaboration within the researcher-in-residence program, he eventually was able to understand the techniques applied and extrapolate them to the results of the project.

For the KB, this is a pivotal project showing the added value of close collaboration with a researcher. Although the KB participates in many research projects, its main role is acting as data provider, allowing researchers to use the large datasets of the KB. However, the library can do more to take full advantage of the knowledge created in these projects and implement the results of the research to its collections. Given the collaborative nature of the researcher-in-residence program, both aspects are covered. Since Smits is a domain expert in the field of historical visual culture, he helped the KB to understand their data better and together with the research software engineer he created a training set to build the algorithm that classified the images. If the KB manages to apply this algorithm to all images in the KB dataset and implement the filter option in Delpher, the results of this collaboration will be beneficial to all visitors of www.delpher.nl.

6 Conclusion

The project was a success for all parties involved. The Humanities researcher was able to answer his main research question and presented the results at several conferences (Smits, 2017; Smits en Wevers, 2018a, 2018b) and published an article in *Digital Scholarship in the Humanities* (Wevers and Smits, 2019). Furthermore, the Humanities researchers and the research software engineer created a dataset, tool and code that are all freely available through KB's Lab. The research software engineer gained a lot of knowledge about the possibilities of computer vision techniques to further open up the libraries digital collection. Finally, the digital scholarship advisor is currently exploring the possibilities to implement the results of the project within Delpher so that it can benefit a large audience (Delpher.nl has two million visits per year).

This last conclusion is an example of the potential of applying research results to library services in order to open up digital collections to a wider audience. Earlier, Peter Leonard (2016) made a plea for this for this when he stated he wanted to 'put TDM in the mainstream.' Alex Humphreys (2018) made a similar plea for 'Applied Digital Humanities' and (Kleppe, 2018) also referred to the potential of 'Libraries as incubators for DH Research Results'. It demonstrates the crucial role institutes, such as libraries, can play within research projects. When these institutes go beyond the role of data provider, they are not only a full partner by bringing and gaining knowledge, but they can also act as the ideal valorisation vehicle of research projects. By taking up an active role in adopting relevant research results in their own services, they can preserve these results in a sustainable manner and bring the affordances of DH research to the wider public.

References

Barnhurst, K., Nerone, J., 2000. Civic Picturing vs. Realist Photojournalism. The Regime of Illustrated News, 1856-1901. Design Issues 16, 59–79.

Broersma, M., 2014. Vormgeving tussen woord en beeld. De visuele infrastructuur van Nederlandse dagbladen, 1900 – 2000. Tijdschrift voor Mediageschiedenis 7, 5–32.

Geitgey, A., 2017. Face_recognition: The world's simplest facial recognition api for Python and the command line: https://github.com/ageitgey/face_recognition

Gervais, T., 2010. Witness to War: The Uses of Photography in the Illustrated Press, 1855-1904. Journal of Visual Culture 9, 370–384.

Gervais, T., Morel, G., 2017. The Making of Visual News: A History of Photography in the Press. Bloomsbury Academic, London.

Gitelman, L., Pingree, G. (Eds.), 2003. New Media, 1740-1915. MIT Press, Cambridge.

Hill, J., Schwartz, V., 2015. Getting the picture: the visual culture of the news. Bloomsbury Academic, London.

Humphreys, A., 2018. The Case for Applied Digital Humanities in Scholarly Communications. Presented at the SSP Annual Meeting, Chicago.

Keller, U., 2013. The iconic turn in American political culture: speech performance for the gilded-age picture press. Word and Image 29, 1–39. https://doi.org/10.1080/02666286.2012.729794

Kester, B., Kleppe, M., 2015. Persfotografie. Acceptatie, professionalisering en innovatie, in: Bardoel, J., Wijfjes, H. (Eds.), Journalistieke Cultuur in Nederland. Amsterdam University Press, Amsterdam, pp. 53–76.

King, L., Leonard, P., 2017. Processing Pixels: Towards Visual Culture Computation. Presented at the ADHO 2017.

Kleppe, M., 2018. Keynote: Bringing Digital Humanities to the wider public: libraries as incubator for DH research results. Presented at the Language Technologies & Digital Humanities Conferences, Ljubljana, Slovenia. https://doi.org/10.5281/zenodo.2532678

Kleppe, M., Elliott, D., Faber, W.J., 2016. *K*oninklijke Bibliotheek Kranten – 1 Miljoen (KBK-1M). KB Lab: The Hague. http://lab.kb.nl/dataset/kbk-1m // https://doi.org/10.17026/dans-xar-hqvg

Kleppe, M., Lincoln, M., Wevers, M., Williams, M., Seguin, B., Smits, T., 2017. Computer Vision in Digital Humanities, in: Confernce. Presented at the DH2017, ADHO, Montreal, pp. 833–836.

Moretti, F., 2015. Distant reading. Verso, London.

Nicholson, B., 2013. The Digital Turn. Media History 19, 59–73.

Park, D., 1999. Picturing the War: Visual Genres in Civil War News. The Communication Review 3, 287–321.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115, 211–252.

Seguin, B., di Leonardo, I., Kaplan, F., 2017. Tracking Transmission of Details in Paintings. Presented at the Digital Humanities 2017, Montreal.

Smits, T., 2017. Illustrations to Photographs: using computer vision to analyze news pictures in Dutch newspapers, 1860-1940. Presented at the Digital Humanities 2017, Montreal.

Smits, T., Faber, W.J., 2018. CHRONIC (Classified Historical Newspaper Images). KB Lab: The Hague. http://lab.kb.nl/dataset/chronic-classified-historical-newspaper-images

Smits, T., Faber, W.J., 2018. CHRONReader. KB Lab: The Hague. http://lab.kb.nl/tool/chronreader

Smits, T., Wevers, M., 2018a. Seeing History: The Visual Side of the Digital Turn. Presented at the DH2018, Mexico City.

Smits, T., Wevers, M., 2018b. Seeing History: The Visual Side of the Digital Turn. Presented at the DHBenelux 2018, Amsterdam.

Steinsieck, A., 2006. Ein imperialistischer Medienkrieg. Kriegsberichterstatter im Südafrikanischen Krieg (1899–1902), in: Daniel, U. (Ed.), Augenzeugen. Kriegsberichterstattung vom 18. zum 21. Jahrhundert. Vandenhoeck & Ruprecht, Göttingen, pp. 87–112.

Tilton, L., Arnold, T., Smits, T., Wevers, M., Williams, M., Torresani, L., Bell, J., Latsis, D., 2018. Computer Vision in DH. Presented at the DH2018, Mexico City.

Wevers, M., Lonij, J., 2017. SIAMESET. KB Lab: The Hague. http://lab.kb.nl/dataset/siameset

Wevers, M., Smits, T., 2019. The Visual Digital Turn. Using Neural Networks to Study Historical Images. Digital Scholarship in the Humanities (accepted).