The visual digital turn: Using neural networks to study historical images

Melvin Wevers
DHLab, KNAW Humanities Cluster, Amsterdam, The Netherlands

Thomas Smits
Department of Cultural Studies, Radboud University, Nijmegen, The Netherlands

Abstract

Digital humanities research has focused primarily on the analysis of texts. This emphasis stems from the availability of technology to study digitized text. Optical character recognition allows researchers to use keywords to search and analyze digitized texts. However, archives of digitized sources also contain large numbers of images. This article shows how convolutional neural networks (CNNs) can be used to categorize and analyze digitized historical visual sources. We present three different approaches to using CNNs for gaining a deeper understanding of visual trends in an archive of digitized Dutch newspapers. These include detecting medium-specific features (separating photographs from illustrations), querying images based on abstract visual aspects (clustering visually similar advertisements), and training a neural network based on visual categories developed by domain experts. We argue that CNNs allow researchers to explore the visual side of the digital turn. They allow archivists and researchers to classify and spot trends in large collections of digitized visual sources in radically new ways.

Correspondence:
Melvin Wevers, Oudezijds Achterburgwal 185, 1012 DK Amsterdam, The Netherlands.
E-mail:
melvin.wevers@dh.huc.knaw.nl

1 Introduction

In a 2013 article, Nicholson (2013) discussed the first stage of the Digital Turn: a 'formal body of scholarship driven by digital methodologies'. True to his predictions, six years later, the bulk of digital humanities scholarship has coalesced around one of the defining features of digital archives: the ability to use keywords to search through and analyze large numbers of digitized texts (Kestemont *et al.*, 2014; Smith *et al.*, 2015; Edelstein *et al.*, 2017). Nicholson (2013) argues that this focus can be explained by the results of the 'practical revolution' of optical character recognition (OCR) technology. Textual analysis, in all shapes and forms, has come to dominate the field. In the first paragraph of *Debates in the Digital Humanities*, Klein and Gold (2016) define digital humanities as being centered on 'digital archives, quantitative analysis, and tool-building projects'. Although the authors mention visualizations of image collections, not a single chapter is devoted to the large-scale visual analysis of images. In contrast, a whole section of the book is devoted to 'text-analysis at scale'. Similarly, Champion (2017) notes that another definition, the one put forth by the University of Oxford on its Web site, is 'text based and desk based'. In the same year that Nicholson published his article on the Digital Turn, Meeks (2013) raised the following question on his blog: 'Is Digital Humanities Too Text-Heavy?'

doi:10.1093/llc/fqy085 Advance Access published on 18 January 2019

In the introduction to the *Routledge Companion to Media Studies and Digital Humanities*, Sayers (2018) notes that 'digital humanities frequently deems text its primary medium for both composition and analysis'. Sayers presents the move 'beyond text' to a 'constellation of modalities, including listening, seeing, scanning, touching . . .' as one of the most important contributions of media studies to the field of digital humanities. Text-based querying indeed only provides us with a limited, and necessarily mainly textual, view of digital sources. As a field, digital humanities has grossly neglected visual content, causing a lopsided representation of all sorts of digitized archives. Therefore, we call for a turn toward the visual in digital humanities research.

To be fair, until recently, the computational analysis of visual material was limited by practical considerations. Despite its promising introduction, only one article in the *Routledge Companion* deals with the large-scale computational analysis of images. Kuhn (2018) mainly discusses adding metadata to visual material by manual tagging. While she takes note of the advances in the field of computer vision, Kuhn presents training an algorithm to identify objects as 'quite tricky'. This assessment most likely stems from the dearth of accessible techniques that allowed the same revolutionary large-scale and automated analysis of images as had been made possible for text by OCR technology. However, in the past five years, scholars have taken the first steps in the field of visual big data and started to use computational methods to study visual material (Smith, 2013; Ordelman *et al.*, 2014). Exemplary studies (Manovich, 2012, 2015; King and Leonard, 2017) analyze images either by looking at metadata or based on basic features, such as size, color, or saturation.

Convolutional neural networks (CNNs), a relatively recent development in computer vision, enable researchers to take the analysis of large numbers of images a step further. They can be used to explore the content (what is represented) and the style (how is it represented) of images. In the wake of this development, scholars have started to note how these techniques can benefit humanities research. For example, in digital art history, scholars applied these methods to discover patterns in large databases of paintings (di Lenardo *et al.*, 2016; Impett and Moretti, 2017). In their project 'Distant Viewing', Taylor Arnold and Lauren Tilton analyze television series using computer vision.[1] This article builds upon this work and demonstrates how CNNs can be used to explore and analyze the content of large numbers of digitized historical images. They open up a part of the digital archive for large-scale analysis, which, until now, has been left uncovered: the millions of images in digitized books, newspapers, periodicals, and historical documents. As a result, they allow us to explore the visual side of the digital turn in historical research. Using these techniques, we can explore visual material in archives using nontextual search methods. Scholars can, for example, find visual material related to a particular topic, or, they can indentify transitions in the use of a particular medium, such as illustrations and photographs.

This article presents three different approaches that use CNNs to gain a deeper understanding of visual trends in an archive of digitized Dutch newspapers. These include detecting medium-specific features, querying images based on abstract visual aspects, and training a neural network based on visual categories developed by domain experts. More specifically, we show how CNNs can separate photographs from illustrations, how they can be used to cluster visually similar advertisements, and how a retrained network can identify buildings, cartoons, chess problems, crowds, faces, logos, maps, and weather reports on historical images taken from digitized newspapers. Before we turn to these three approaches, we briefly discuss the field of computer vision, CNNs, and the data sets we used for our approaches. In our concluding remarks, we discuss the challenges of working with CNNs and offer several possibilities for future research and inquiry.

2 Computer Vision and Convolutional Neural Networks

Computer vision is an academic field that is concerned with using computation to gain a high-level understanding of images. It relies on mathematical techniques to detect the shape and appearance of

objects in imagery. Since the late 1980s, several methods have been developed to detect shapes in images that represent specific objects. Early computer vision techniques looked for specific, predetermined combinations of shapes.

An early computer vision task involved the recognition of handwritten digits, made available as the MNIST data set. In 1993, the CNN LeNet-1 was one of the first to recognize these digits with considerable accuracy and speed (Lecun *et al.*, 1995). As computing power increased, more extensive neural networks could be trained, improving the speed, accuracy, and complexity of recognition tasks. Concomitant with increasing computing power, the growing availability of digital visual material proved to be a catalyst for the advancement of computer vision techniques.

The development of computer vision tasks relies on benchmarked data sets. Because the images in these data sets have been manually tagged, researchers can compare an algorithm's performance with that of a human. In other words, they can check how well an algorithm performs in detecting, for example, cats in images that have been tagged as containing cats. The ImageNET Large Scale Visual Recognition Challenge is the most influential benchmark tests for object classification. The goal of the annual ImageNet challenge is to detect objects in a set of 1.2 million images extracted from the Internet and recognize around 1,000 different object categories. The precision (cats are tagged as cats) and recall (all cats on an image are identified) of computer vision algorithms that competed in the challenge have improved drastically in recent years. This improvement can be attributed to the introduction of CNNs in 2012 (Krizhevsky *et al.*, 2012). In 2014, this development continued when Google's winning algorithm had an error rate of 6.8%: a reduction of almost 50% compared with the previous year (Szegedy et al., 2014).

So, neural networks can be applied to computer vision tasks, but how do they work? First of all, algorithms do not *look* at an image as humans do; they process images as matrices of pixel values (Fig. 1). For an algorithm, each image consists of a grid of numbers between 0 and 255, indicating a pixel's intensity. Together these pixels display a specific conceptual structure, such as a cat or Abraham Lincoln. Computer vision algorithms are designed to detect particular patterns in these pixel values, indicative of elements in these conceptual structures.

The primary challenge for computer vision algorithms is variation. An image can show, for example, a cat in different positions: it can be walking, sitting down, or climbing up the stairs. Furthermore, not all parts of the cat might be visible on the image and cats can have different sizes and colors. A CNN can recognize all these different 'possibilities' of a cat, by looking at multiple lower-level features that uniquely predict the presence of a cat in all its variations. A computer vision algorithm 'sees' these low-level features as spatial relationships between a group of pixels in the larger grid of pixels.

The algorithm extracts these low-level features from images using a mathematical process, known as a convolution. A convolution changes the value of a particular pixel based on the values of the pixels that surround it. As a result, certain convolutions bring to the fore particular kinds of spatial relationships between pixels. Think of, for example, contrast, edges, curves, straight, or diagonal lines in images (for more on convolutions, see Karn, 2016). A CNN's structure, often described as its architecture, consists of multiple layers of convolutions that highlight particular low-level features. By combining several of these layers, the CNN learns that particular combinations of low-level features point to specific high-level features: for example, the particular shape of a cat's face or its pointy ears. During training, a neural network learns the optimal configuration of convolutions; that is the configuration that performs best in predicting the correct label for an annotated image.

This article describes three approaches that use CNNs for detecting several forms of visual similarity in historical images. We have predominantly relied on Inception-V3, a model trained on the ImageNet set, made available by Google as part of their TensorFlow library.[2] The first approach classifies images according to medium-specific characteristics. The CNN trained for this purpose focuses on the low-level features of images to detect whether an image is an illustration or a photograph. The second

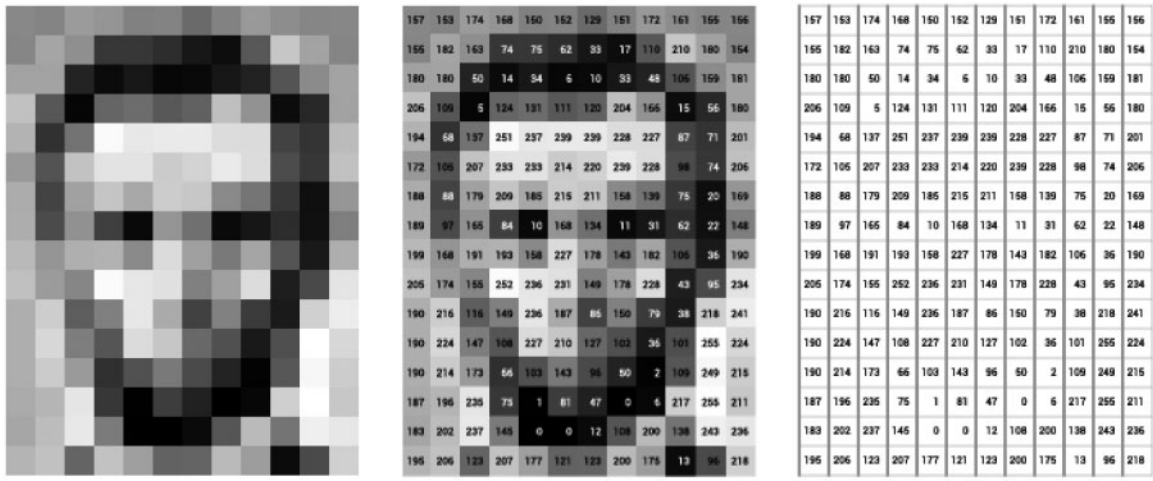

Fig. 1 Image of Abraham Lincoln as a matrix of pixel values

approach clusters images based on high-level features in the penultimate layer of the neural network. In this layer, images are represented as a series of 2,048 abstract features, rather than being classified in one of the thousand classes of ImageNet. This allows us to group images based on the presence of specific visual aspects rather than their category. The third approach demonstrates how retraining the final classification layer of a neural network using self-defined categories can be used to analyze the content of historical images. It demonstrates how a CNN can be retrained to look for images that share an explicit visual similarity, such as weather reports that look almost the same every day.

3 Data Sets: Images in Digitized Newspapers

This article focuses on two types of images that appeared in Dutch newspapers: images that were part of newspaper articles and images in advertisements.[3] We extracted the images from the digitized newspaper collection Delpher, to which we had access during our researcher-in-residence projects at the National Library of the Netherlands (KB).[4] In addition to large amounts of textual information, the archive also contains a wide range of visual content.

The first data set, CHRONIC (Classified Historical Newspaper Images), contains 452,543 images for the period 1860–1930. Because we were mainly interested in images of the news, we decided to only include images related to newspaper articles, thereby excluding the images of the advertisement sections and discarding relatively small images by filtering out files with sizes smaller than 30 kB. CHRONIC holds a wide variety of visual material, such as illustrations and photographs of news events, (political) cartoons, images that accompanied *feuilletons*, but also large numbers of chess problems and weather reports.

The second data set (SIAMESET) consists of 426,777 advertisements published in two influential Dutch newspapers, the *Algemeen Handelsblad* and *NRC Handelsblad*, between 1945 and 1995. Whereas the images of the CHRONIC data set are separated from the textual content, the advertisements of SIAMESET consist of both visual and textual content. In fact, many advertisements are mostly text-based. Before we could train a CNN to look for similarities within images, we, therefore, had to create a data set of advertisements that contained a high degree of visual content. We filtered the initial set, which consisted of 1.6 million advertisements, in two steps. First, we removed images with a width or height smaller than 500 pixels and advertisements with dimensions that resembled classified ads (height of >5,000 pixels and width of

<900 pixels). Second, we removed advertisements with a character proportion higher than 0.0005.[5] This resulted in a data set of 426,777 advertisements for the period 1945–1995.

4 Approach 1: Detecting Medium-Specific Characteristics

The visual representation of news events is generally connected to the technological progress of photography (Gervais and Morel, 2017). The so-called halftone revolution of the early 1880s, enabling the massive reproduction of photographs in print media, is seen as having shaped our current visual culture of the news. In contrast, from a media-archeological perspective, Hill and Schwartz (2015) propose a contingent history of 'news pictures' as a separate 'class of images', which focuses on not only photographic technology but also the discourse surrounding them. Concerning this recent theoretical development, several studies have demonstrated that photography was not the first medium 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 (Park, 1999; Barnhurst and Nerone, 2000; Keller, 2001; Gervais, 2010). Other studies suggest a relatively long transitional period in which illustrations and photographs coexisted and competed as authentic, objective visual representations of the news (Steinsieck, 2006; Keller, 2013). Can we use a CNN to study the transition between illustrations and photographs to visualize the news? More generally, how can we apply computational methods to shed new light on these kinds of long-standing questions of visual culture studies?

The earlier reliance on case studies to describe the transitional phase is unsurprising, as, in pre-digital times, a 'distant reading' of a large number of images published in newspapers was all but impossible. The recent rapid development of CNNs has made such a study possible. As part of the object recognition task, the neural network also picks up on particular features of images on a more elementary level. We exploited this and used a CNN trained especially for classifying an image as either a photograph or an illustration.[6] This shows how CNNs can be used to recognize medium-specific characteristics: in this case the distinctive patterns between groups of pixels in engraved illustrations and halftones.[7] Our neural network achieved an F_1 score—a harmonic mean of the precision and recall—of 0.9 over the entire period, meaning that it accurately predicted the type of 90% of the images. This means that we can reliably use it to divide the 452,543 images in the CHRONIC data set. This information can then be used to study the historical use of illustrations and photographs in Dutch newspapers.

The classification of CHRONIC's images enables the analysis of the visual culture of the news on a large scale. Figure 2 shows the presence of illustrations and photographs in Dutch newspapers between 1860 and 1930. Although the halftone technique was introduced in the early 1880s, Dutch newspapers already began printing illustrations in the same decade. 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 1925. 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 earlier, based on case studies. At the same time, vast digitized archives and new techniques like CNNs contribute to the construction of a radically new, overview of visual (news) culture, which allows for the analysis of trends and changes over extended periods. As Fig. 2 shows, the visual representation of the news took off in the earlier 1920s. Although earlier work noted and analyzed the introduction of so-called 'photo-pages' in the 1920s using a limited set of sources (Broersma, 2014; Kester and Kleppe, 2015), the bird's-eye view of the entire Dutch press provides us with a perspective on the magnitude of this watershed moment.

The application of computer vision techniques also leads to novel insights into the visual representations of news events in the early twentieth century.

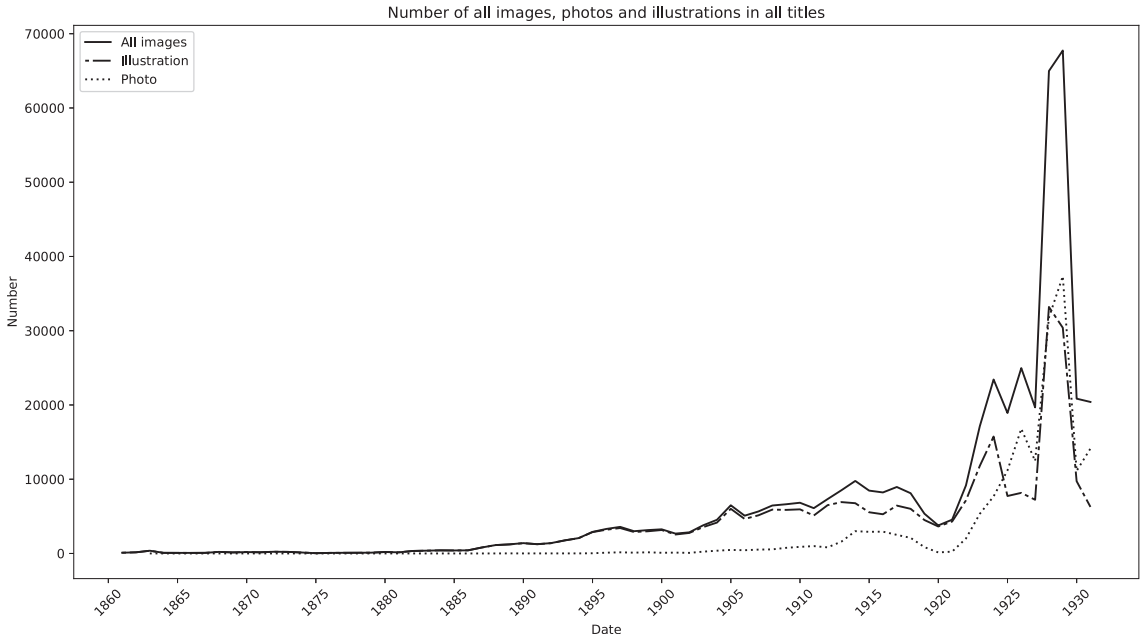

Fig. 2 The number of images, photos, and illustrations in digitized Dutch newspapers, 1860–1930

For example, the classification results show that photographs varied considerably in terms of quality. Many newspapers published illustrated supplements, mostly on the weekend, for which they used high-quality photographs. However, on weekdays, newspapers used not only considerably fewer photographs but also lower-quality ones. The photograph of a soccer team in the *Bataviaasch Nieuwsblad*, a newspaper published in the former Dutch colonial city of Bandoeng, is a good example (Fig. 3).[8] Some newspapers heavily retouched their photographs to improve the quality of such images, outlining objects and persons. The depiction of Austro-Hungarian infantrymen in the Carpathian Mountains in 1915 is a clear example of these practices (Fig. 4).[9] The heavy retouching makes it difficult for both humans and computers to distinguish photographs from illustrations.

5 Approach 2: Abstract Visual Aspects (SIAMESE)

After the Second World War (1940–1945), the Netherlands transformed into a modern consumer society, expressed by an increase in purchase power, the spread of household technologies and branded consumer goods, and a more efficient production and distribution system (de la Bruhèze and Oldenziel, 2009; Schot *et al.*, 2010). Scholars have also described the period as one of Americanization, which had its effect on advertisements and the Dutch consumer society at large (Roholl, 1992; Schreurs, 2001). These transitions in the Dutch consumer landscape make the selected period of particular interest for the study of visual trends in advertisements.

Until recently, it was difficult and time-consuming to study such trends on a large scale. Researchers had to manually browse through archives to spot visual trends, which they related to larger cultural–historical phenomena. In his seminal work *Advertising the American Dream*, Roland Marchand (Marchand, 1985) describes advertisements as social tableaux, in which depictions of consumers suggest the larger relationship to social structures. Particular examples include the use of the office window and family circles as visual clichés. Marchand applies a highly contextualized close-reading of specific ads to support his theories.

Fig. 3 A soccer team in Bandoeng

Fig. 4 Austro-Hungarian infantrymen in the Carpathian Mountains in 1915

In this second approach, we show how the penultimate layer of a CNN can be used to cluster images based on their visual similarity. As a result, we can discover visual trends in advertisements that can either support existing theories or generate new hypotheses on the marketing of consumer goods in mass media. In the 1930s, Dutch advertisers traveled to the USA to learn the tricks of the trade from their American colleagues. Being able to extract visually similar advertisements from an archive can help the researcher to detect transitions in the style of advertising. These styles could then be compared with archetypical American advertisements or advertisements in a different country.

The final layer of a CNN is used to group the detected objects into one of the thousand categories in ImageNet. We discovered that the classification in the final layer of the trained Inception-V3 network does not perform well on historical advertisements. The main reason being that the images in ImageNET were sourced from contemporary online sources. Therefore, neural networks trained on ImageNet can accurately identify objects in their contemporary form on pictures that have been made using modern high-definition cameras. However, historical images often feature objects that no longer exist or at least not in the same form. Moreover, historical images were not always photographed and when they were, they have different medium-specific features. For example, the color schemes in cameras from the interwar period were markedly different from modern ones. Furthermore, the visual content of advertisements regularly consisted of cartoonish drawings or highly stylized graphic designs. The classification layer of existing neural networks did not perform well in recognizing objects represented in this visual aesthetic.

Because retraining a neural network requires large annotated data sets and extensive computational power, we looked for different ways to use existing neural networks to identify visual similarity. We turned to the penultimate layer of the CNN to identify similar visual trends in advertisements. In the penultimate layer, after numerous convolutions and other transformations, an image is represented as a 2,048-dimensional vector, or in layman's terms, a list of 2,048 numbers. The CNN uses these aspects to predict what is represented on the image. The numeric expressions of the vector can also be used to cluster images. Within it, we can look for points that are close to each other—also called nearest neighbors—represented by smaller Euclidean distances.[10] Simply put, similar sets of numbers in two pictures suggest some form of visual similarity.

To make the data set more accessible to a broader audience and allow explorative searches, we created SIAMESE: a Web interface for querying the advertisements based on their abstract visual aspects.[11] When a user first visits SIAMESE, the system randomly selects an image as the source image. The Web interface presents users with the ten most

Fig. 5 Cropped output from SIAMESE displaying similar car ads

similar images to a source image and a timeline consisting of the most similar image in every year between 1948 and 1995. The first option allows users to detect whether the source image was part of an identifiable visual style, whereas the latter shows the development of a visual trend over time. SIAMESE can, for example, find images that contain objects that look like cars in particular periods, but also other objects that share car-like features (Fig. 5). On the one hand, this enables users to trace how the visual style of automobile adverts changed over time, while also uncovering visual similarities between ads for different products.

SIAMESE can also group advertisements based on a particular layout. In Fig. 6, we see advertisements for products ranging from toothpaste to coffee. These advertisements are grouped because they display a similar visual style: a large image in the upper part of the advertisements and a text box with an occasional smaller image in the bottom part. By browsing through the system, researchers can quickly uncover particular visual styles of advertisements and their prominence in particular periods.

SIAMESE performs well when advertisements contain clearly identifiable objects. This is the case in adverts for fashion, beverages, and automobiles. In contrast, other products that are heavily advertised in Dutch newspapers, such as mortgages or banks, rarely pop up as results, as they do not contain specific recurring objects. When they are clustered together, it is because adverts share similarities in their layout style. Interestingly, advertisements that contain similar objects but widely diverging layouts are often not clustered together. A more fine-grained approach to clustering based on visual similarity is required to solve this challenge.

The grouping of advertisements on specific visual aspects also confronts researchers with the difficulty of defining categories for image classification. For example, SIAMESE can detect bottle-like shapes. These shapes, however, could signify alcoholic beverages, soft drinks, or bottles of milk. These all fall into the category bottles, but depending on the research question, users might want to classify them accordingly. This goes to show that constructing a classificatory system should be done using the input of domain experts, and perhaps related to the research question at hand. Moreover, differences between the objects represented in images might not always be explicitly expressed visually, which makes classifications based on both textual and visual input all the more pressing when working with images. Future work will look into the combination of text and image similarity to counter some of these difficulties.

6 Approach 3: Building Your Own Classifiers

We created the CHRONIC data set to study the transition between illustrations and photographs in the history of the visual representation of the

Fig. 6 A set of advertisements featuring a particular visual style

news. Sifting through the images, we quickly discovered that Dutch newspapers contained a large number of frequently recurring images, such as chess problems and weather reports, and that many illustrations were not visual depictions of news events, but rather (political) cartoons. To create a more consistent subset of news images, we trained a new classification layer on top of the existing Inception-V3 model to be able to recognize nine relevant categories: buildings, cartoons, chess, crowds, logos, maps, schematics, sheet music, and weather reports.[12] Somewhat similar to using full-text search, this technique allows researchers to access the content of visual material stored in digital archives using faceted search.

These categories divide the images of the set by looking at different forms of visual similarity. The categories for weather reports and chess problems, for example, classify images that are similar in a straightforward manner (Fig. 7). In other words: they all look almost the same. The building and crowd categories do not select images on the basis of direct visual similarity, but instead look for visual concepts: human-made structures and crowds of people in this case. In contrast, the cartoon, map, and schematics categories more clearly identify images that share a similar visual style (Fig. 8).

We tested the performance of these different classifiers by calculating their F_1 scores on the basis of a manually tagged random sample of 500 images (Fig. 9).[13] Not surprisingly, with scores of 0.95 and 0.94, the images with the most explicit visual style, the weather reports and chess problems, performed especially well. Images with conceptual similarity but greater visual irregularity, such as schematics, maps, and crowds, showed decreased scores (0.85, 0.8, and 0.72, respectively). Although the concept of a crowd is quite clear, namely, that of a group of people, its visual representation can be quite varied. Images of crowds often share some but not all visual aspects.

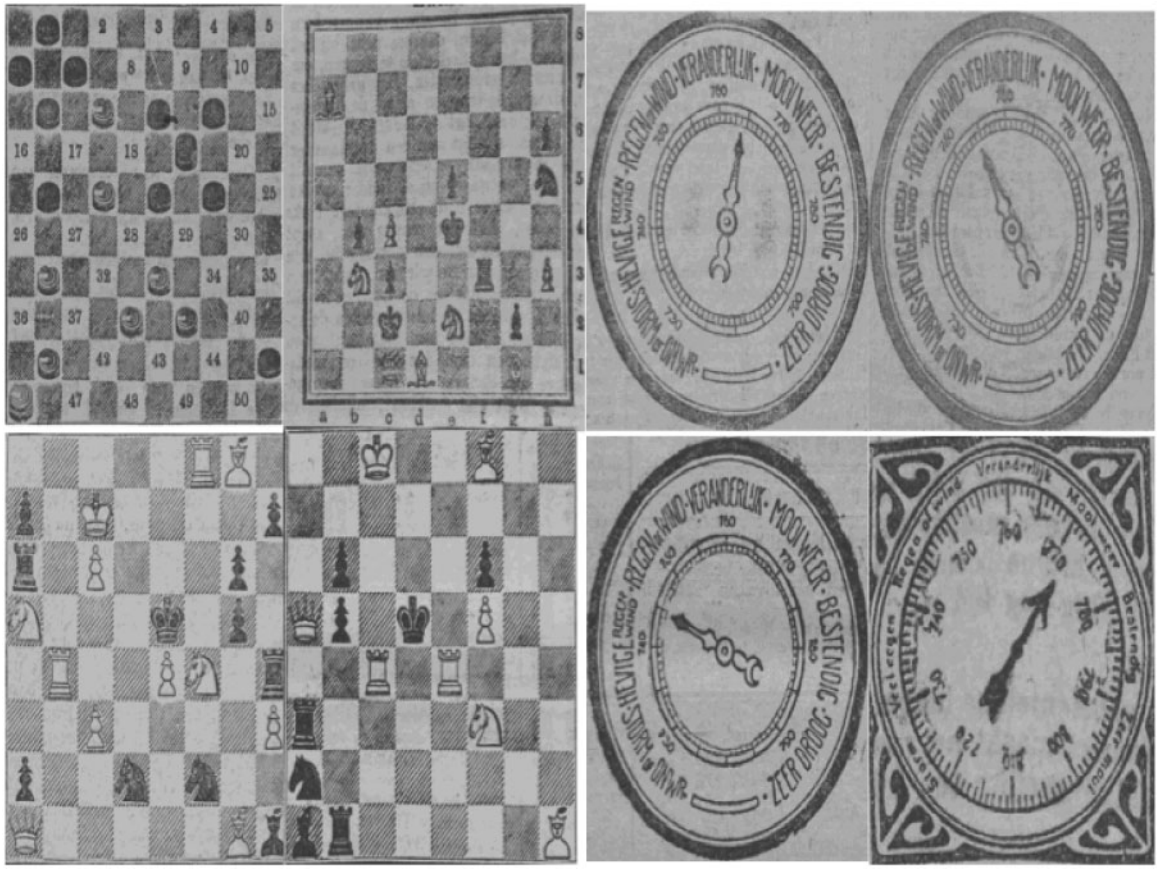

Fig. 7 Images used to train the categories: weather report and chess problem

Recognizing such concepts is still a challenge for computer vision algorithms.

Training a new classification layer on top of an existing model allowed us to look for specific kinds of images in a data set quickly. Because this method makes use of a pre-trained model, we only have to introduce a limited number of new positive and negative examples to construct an additional, functioning classification layer. In some cases, such as the weather reports, twenty-five positive and twenty-five negative images were enough to achieve relatively high F_1 scores. This method allows us to access the content of visual material stored in digital archives directly. We can look for, and chart, the visual representation of objects specific to a research question, for example, the depiction of mass demonstrations in the early twentieth-century press. In a way, the technique is similar to OCR technology; it allows us to access the content of images directly, without referring to textual content, such as titles and captions. To allow for the querying of CHRONIC's images using a combination of text search and visual categories, we build a Web application, CHRONReader.[14]

7 Conclusion

CNNs offer new and exciting possibilities for collection specialists, users of digital archives, and humanities researchers. First of all, as the first and third approaches show, we can use them to classify large numbers of visual sources semi-automatically. This enables archivists and researchers to look at collections in radically new ways. CNNs allow us to open

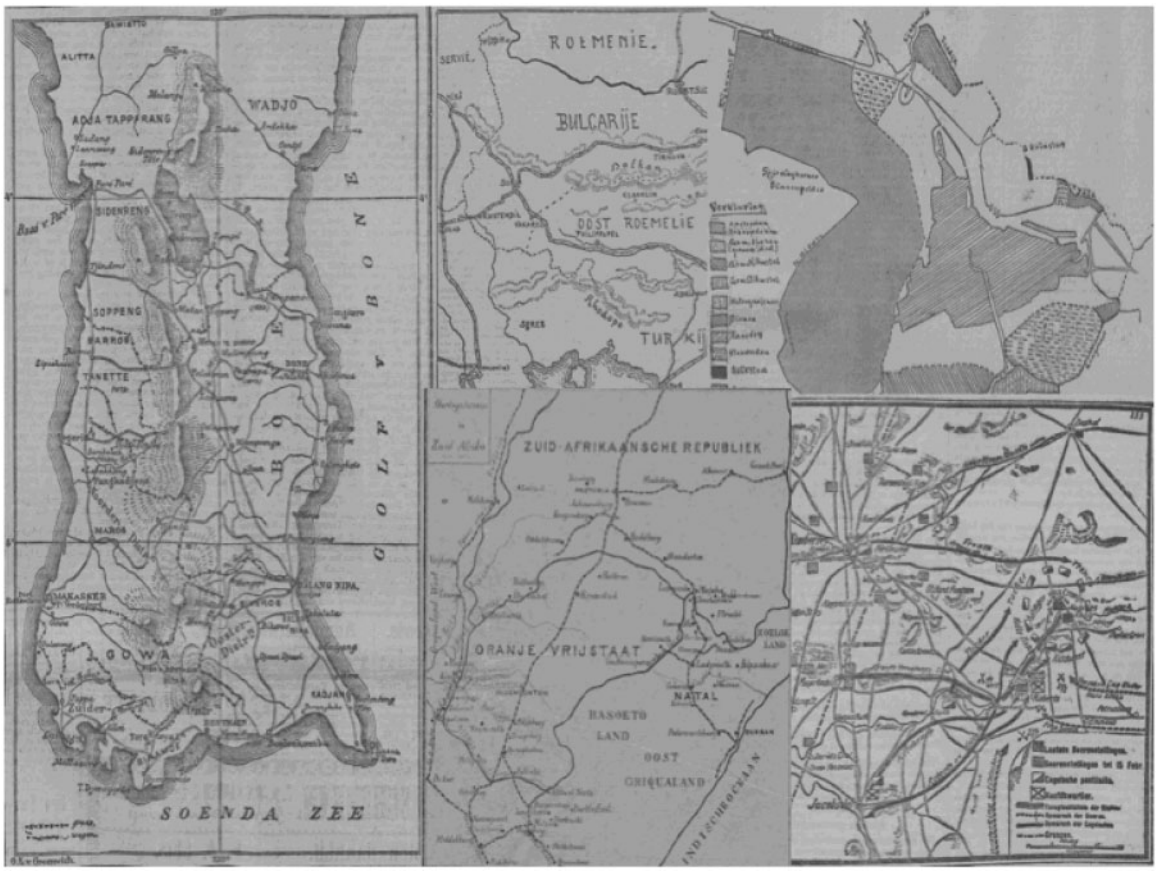

Fig. 8 Images used to train the category: map

up the visual side of historical archives. Instead of relying on titles and captions, the often-flawed, or at least often-inconclusive, textual descriptions of images, we can use CNNs to describe the visual content of historical sources. As CHRONReader shows, these new techniques allow users of digital archives to query the visual content of digital archives, without having to rely on textual elements. In addition, as our second approach shows, CNNs also open up a new world of intuitive and serendipitous exploration of the visual side of digital archives. They enable researchers to follow visual trends and styles and track visual similarity over time.

Scholars can use CNNs to explore and analyze large collections of visual sources without having to browse through archives manually. Just as OCR technology allowed them to read text from a distance, computer vision techniques offer us a macroscopic perspective on thousands of images. Our first approach demonstrates that this new possibility can be used to provide new answers to some long-standing questions in visual culture studies. More broadly, they open up new perspectives on a central concern of the cultural turn, namely, its reliance on close-reading small numbers of sources, often from short periods. Computer vision techniques offer the possibility to track and analyze specific kinds of visual representations over long periods in the archive. In addition, the use of these techniques and the results they yield can lead to new research questions and areas of inquiry.

Until now we discussed the benefits of computer science techniques for humanities research. However, in our opinion, central insights from the

Category	F1-score (1860 - 1900)	F1-score (1900-1922)	F1-score (1860-1922)
Photo	-	0,81	0,9
Faces	0,79	0,58	0,57
Buildings	0,89	0,65	0,45
Cartoon	0,67	0,7	0,67
Chess	0,99	0,95	-
Crowds	0,74	0,68	0,72
Logos	0,78	0,51	0,72
Maps	0,67	0,81	0,8
Schematics	0,82	0,81	0,85
Sheet music	-	-	-
Weather	0,67	0,95	0,94

Fig. 9 F_1 scores based on three sets of 500 random images from 1860 to 1900, 1900 to 1922, 1860 to 1922

humanities could likewise be a boon to the development of more accurate and more sophisticated computer vision techniques. As classification algorithms have been trained on manually tagged sets, structural biases in their classification schemes will be reproduced in the results produced by computer vision techniques. In collaboration with humanities scholars, computer scientists could critically engage with these biases and rethink the way we annotate data sets and measure algorithmic accuracy.

In addition, a better grasp of what we talk about when we talk about visual similarity could result in more sophisticated computer vision algorithms. As our results show, visual similarity is not always conceptual similarity. The direct visual similarity of weather reports is different from the medium-specific similarity of photographs and illustrations, the stylistic similarity in certain advertisements, and the conceptual similarity of images containing crowds. Together, humanities scholars and computer scientists should work toward a layered understanding of visual similarity.

Applying new techniques indubitably raises many questions and introduces novel tasks. The most significant challenge we identified during our research involves the artificial separation in textual and visual discourse analysis. In his seminal 'There are no visual media', W.J.T. Mitchell (2005) takes issue with the notion of 'pure' media, consisting of a single form of discourse. Media always consist of more than one form of discourse, and the meaning of these mixed-media forms can only be studied if we look at the relations between the different forms of discourse. With its focus on text, digital humanities has until now certainly not contributed to a better understanding of media in the general sense. Although CNNs offer us the possibility to study and analyze the previously neglected visual side of the digital archive, we also run the risk of widening the gap between the study of visual and textual discourse. If we want to improve the methodological and conceptual rigor of digital humanities research, we have to look for techniques that are able to process and examine multiple forms of discourse in conjunction.

In this paper, we have highlighted three ways in which neural networks can benefit historical research of visual material. By making our data sets of images available, we hope we can stimulate further research in this domain. Moreover, as research on computer vision is advancing with tremendous speed, we think the time is right for humanists and computer scientists to team up and start applying these techniques in conjunction with text analysis. Only then can we truly speak of a visual digital turn.

Acknowledgments

This research was supported by the National Library of the Netherlands and executed during the authors' stay as researchers-in-residence at the library. The authors would like to thank Martijn Kleppe,

Willem-Jan Faber, Juliette Lonij, and Leonardo Impett for their assistance.

References

Barnhurst, K. and Nerone, J. (2000). Civic picturing vs. realist photojournalism. The regime of illustrated news, 1856-1901. *Design Issues*, **16**(1): 59–79.

Broersma, M. (2014). Vormgeving tussen woord en beeld De visuele infrastructuur van Nederlandse dagbladen, 1900–2000. *Tijdschrift voor Mediageschiedenis*, **7**(1): 5–32.

de la Bruhèze, A. A. and Oldenziel, R. (eds) (2009). *Manufacturing Technology, Manufacturing Consumers: The Making of Dutch Consumer Society*. Amsterdam: Aksant.

Champion, E. M. (2017). Digital humanities is text heavy, visualization light, and simulation poor. *Digital Scholarship in the Humanities*, **32(supplement 1)**: 25–32.

Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., and Coleman, N. (2017). Historical research in a digital age: reflections from the mapping the republic of letters project. *The American Historical Review*, **122**(2): 400–24.

Gervais, T. (2010). Witness to war: the uses of photography in the illustrated press, 1855–1904. *Journal of Visual Culture*, **9**(3): 370–84.

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

Gold, M. K. and Klein, L. F. (eds) (2016). *Debates in the Digital Humanities*. Minneapolis: University Of Minnesota Press. https://www.upress.umn.edu/book-division/books/debates-in-the-digital-humanities-2016 (accessed 31 August 2018).

Hill, J. and Schwartz, V. (eds) (2015). *Getting the Picture: The Visual Culture of the News*. London: Bloomsbury Publishing.

Impett, L. and Moretti, F. (2017). *Totentanz. Operationalizing Aby Warburg's Pathosformeln*. Stanford, CA: Stanford Literary Lab. https://litlab.stanford.edu/LiteraryLabPamphlet16.pdf.

Karn, U. (2016). An intuitive explanation of convolutional neural networks. *The Data Science Blog*. https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/ (accessed 24 May 2018).

Keller, U. (2001). *The Ultimate Spectacle: A Visual History of the Crimean War*. Amsterdam: Amsterdam University Press.

Keller, U. (2013). The iconic turn in American political culture: speech performance for the gilded-age picture press. *Word and Image*, **29**(1): 1–39.

Kestemont, M., Karsdorp, F., and Düring, M. (2014). Mining the twentieth century's history from the time magazine corpus. In Proceedings of LaTeCH 2014, Association for Computational Linguistics, Göteborg.

Kester, B. and Kleppe, M. (2015). Persfotografie. Acceptatie, professionalisering en inovatie. In Bardoel, J. and Wijfjes, H. (eds), *Jouralistieke Cultuur in Nederland*. Amsterdam: Amsterdam University Press, pp. 53–76.

King, L. and Leonard, P. (2017). Processing pixels: towards visual culture computation. Presented at the *ADHO 2017*, Montreal.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 1097–105.

Kuhn, V. (2018). Images on the move. Analytics for a mixed media approach. In Sayers, J. (ed.), *The Routledge Companion to Media Studies and Digital Humanities*. New York, NY: Routledge, pp. 300–9.

Lecun, Y., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U. A., Sackinger, E., Simard, P., and Vapnik, V. (1995). *Comparison of learning algorithms for handwritten digit recognition*. In International Conference on Artificial Neural Networks, Paris, pp. 53–60.

di Lenardo, I., Seguin, B., and Kaplan, F. (2016). Visual patterns discovery in large databases of paintings. Presented at the *ADHO 2016*, Krakow (accessed 5 March 2018).

Manovich, L. (2012). How to compare one million images?, In Berry, D. (ed.), *Understanding Digital Humanities*. London: Palgrave Macmillan, pp. 249–78.

Manovich, L. (2015). Data science and digital art history. *International Journal for Digital Art History*, **1**(1): 12–37.

Marchand, R. (1985), *Advertising the American Dream: Making Way for Modernity, 1920-1940*. Berkeley, CA: University of California Press.

Meeks, E. (2013). Is digital humanities too text-heavy?. Digital Humanities Specialist. https://dhs.stanford.edu/spatial-humanities/is-digital-humanities-too-text-heavy/ (accessed 11 April 2018).

Mitchell, W. J. T. (2005). There are no visual media. *Journal of Visual Culture*, **4**(2): 257–66.

Nicholson, B. (2013). The digital turn: exploring the methodological possibilities of digital newspaper archives. *Media History*, **19**(1): 59–73.

Ordelman, R., Kleppe, M., Kemman, M., and de Jong, F. (2014). How to integrate audiovisual material in digital humanities research. Presented at the *ADHO 2014*, Lausanne.

Park, D. (1999). Picturing the war: visual genres in civil war news. *The Communication Review*, **3**(4): 287–321.

Roholl, M. (1992). Uncle Sam: an example for all?. In Loeber, H. (ed.), *Dutch-American Relations, 1945-1969: A Partnership: Illusions and Facts*. Assen: Van Gorcum, pp. 105–52.

Sayers, J. (2018). Introduction: studying media through new media. In Sayers, J. (ed.), *The Routledge Companion to Media Studies and Digital Humanities*. New York, NY: Routledge, pp. 1–7.

Schot, J., Rip, A., and Lintsen, H. (eds) (2010). *Technology and the Making of the Netherlands: The Age of Contested Modernization, 1890-1970*. Cambridge: MIT Press.

Schreurs, W. (2001). *Geschiedenis van de Reclame in Nederland*. Utrecht: Het Spectrum.

Smith, D. A., Cordell, R., and Mullen, A. (2015). Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, **27**(3): E1–15.

Smith, J. R. (2013). Riding the multimedia big data wave. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY: ACM, pp. 1–2.

Steinsieck, A. (2006). Ein imperialistischer medienkrieg. kriegsbenchterstattung im sudafrikanischen krieg (1899-1902). In Daniel, U. (ed.), *Augenzeugen. Kriegsberichterstattung Vom 18. Zum 21. Jahrhundert*. Göttingen: Vandenhoeck & Ruprecht.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going deeper with convolutions. ArXiv, 1409.4842 [Cs]. http://arxiv.org/abs/1409.4842 (accessed 24 May 2018).

Notes

1 http://distantviewing.org

2 Inception is a specific type of CNN trained by Google (Szegedy et al., 2014).

3 Both sets have been made available by the National Library of the Netherlands. After signing a license, the data sets can be downloaded. http://lab.kb.nl/dataset/siameset; http://lab.kb.nl/dataset/chronic

4 During our researcher-in-residence projects, the authors were assisted by scientific programmers Juliette Lonij and Willem-Jan Faber.

5 This character proportion was calculated by dividing the size of the ad by the number of characters in the advertisements.

6 We would like to thank Leonardo Impett for training this particular CNN.

7 This approach builds upon the 'Illustrated Image Analytics' project of Paul Fyfe. https://ncna.dh.chass.ncsu.edu/imageanalytics/

8 'De Stedenwedstrijden. Bandoeng', Bataviaasch Nieuwsblad (21 April 1919).

9 'In de Karphaten', Rotterdamsch Nieuwsblad (8 April 1915).

10 For the calculation of these distance metrics, we used Annoy, an approximate nearest neighbor algorithm, which was developed by Spotify. https://github.com/spotify/annoy. The authors would like to thank Peter Leonard for his insights and pointers on working with Annoy.

11 http://kbresearch.nl/siamese/

12 For more on retraining a classification layer, see https://www.tensorflow.org/tutorials/image_retraining

13 Images with multiple possible classifications, for example, those showing a protest in a city (crowds/buildings), were only assigned the category with the highest similarity, according to the algorithm. This might explain the relatively low score of the building category (0.45).

14 http://lab.kb.nl/tool/chronreader