CA

Journal of **Cultural Analytics**

Image Analytics and the Nineteenth-Century Illustrated Newspaper

Paul Fyfe and Qian Ge

10.25.18

Peer-Reviewed By: Natalie Houston and Alison Chapman

Clusters: Image

Article DOI: 10.22148/16.026

Dataverse DOI: 10.7910/DVN/U5HTNJ

Journal ISSN: 2371-4549

Cite: Paul Fyfe and Qian Ge, "Image Analytics and the Nineteenth-Century Illustrated Newspaper," Journal of Cultural Analytics. *October 25, 2018. 10.22148/16.026*

The nineteenth-century British periodical press took textual production to a scale that, for many commentators then and now, summoned the sublime. It was a "flood" which was "too vast to be dealt with as a whole," in the words of the *British Quarterly Review* in 1859. It remained "a vast wilderness … its extent unknown, its ramifications unfathomed" for subsequent researchers, according to Michael Wolff in a 1971 issue of the *Victorian Periodicals Newsletter*.[1] By the numbers for newspapers alone, stamped titles increased almost five-fold from 550 in 1846 to 2,440 in 1906.[2] Simon Eliot estimates the number of copies those

[1] "Cheap Literature," *British Quarterly Review*, April 1859, 316; Michael Wolff, "Charting the Golden Stream: Thoughts on a Directory of Victorian Periodicals," *Victorian Periodicals Newsletter*, no. 13 (1971): 23.

[2] Graham Law and Robert L Patten, "The Serial Revolution," in *The Cambridge History of the Book in Britain: Volume VI, 1830-1914*, ed. David McKitterick, vol. VI (Cambridge: Cambridge University Press, 2009), 156-57.

newspapers rising from 16 million in 1801 to over 78 million by 1849.[3] By the end of the era, the *Daily Mail* claimed to circulate a million copies of each issue on its own. The periodical archive has sprawled even more with its gradual digitization which, while representing only a fraction of nineteenth-century print, still astonishes its researchers. "We are now on the brink of a further, exponential expansion … as vast new quantities of hitherto inaccessible records and texts become available for digital searching," Patrick Leary claimed in 2004.[4] In many ways, nineteenth-century newspapers and periodicals set the very terms for not simply "digital searching" but quantitative methods, having only ever existed as "numbers" (a term for individual issues) in a state of impossible profusion, seeming to welcome computational approaches to the unruly digital archive.

The internal complexity and varying digitization quality of historical newspapers and periodicals make their computational analysis a formidable challenge. Unlike the corpus of nineteenth-century fiction, which has become such a signal example for distant reading and corpus linguistics, periodicals contain multitudes of heterogeneous content, variously arranged by columns and pages, and often serialized across different issues.[5] Several researchers and teams are creating methods which adapt to that complexity, including the text-matching algorithms of Viral Texts project, the image analysis work of Elizabeth Lorang et. al to identify distinctive textual features in newspapers, the work of Ted Underwood et al. to segment genres within periodicals in HathiTrust, and the semantic mapping of page content in the nineteenth-century serials edition (ncse), and Matthew Philpotts's P-MApp tool to measure fractal complexity.[6] While these projects all take ingenious approaches to the periodical archive, they have largely focused on text or the textual page. Arguably, methods in humanities computing have until recently privileged text over the study of other modalities, including research into visual materials which relies on text-based searching or underlying metadata.[7] But as Lorang herself asks, "What can we do with the millions of images

[3] *Some Patterns and Trends in British Publishing, 1800-1919* (London: Bibliographical Society, 1994), 117.

[4] Patrick Leary, "Googling the Victorians," *Journal of Victorian Culture* 10, no. 1 (2005): 15.

[5] For a discussion of the difficulties and opportunities, see James Mussell, *The Nineteenth-Century Press in the Digital Age* (New York: Palgrave Macmillan, 2012).

[6] Ryan Cordell and David Smith, "The Viral Texts Project: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines," 2012; Elizabeth Lorang et al., "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections," *D-Lib Magazine* 21, no. 7/8 (August 2015); Ted Underwood, "Understanding Genre in a Collection of a Million Volumes, Interim Report," December 29, 2014; James Mussell and Suzanne Paylor, "Mapping the 'Mighty Maze': Nineteenth-Century Serials Edition," *19: Interdisciplinary Studies in the Long Nineteenth Century* 1 (2005); Matthew Philpotts, "Dimension: Fractal Forms and Periodical Texture," *Victorian Periodicals Review* 48, no. 3 (2015): 403-27.

[7] Lev Manovich, "Media Visualization: Visual Techniques for Exploring Large Media Collections,"

that represent the digitized cultural record?"[8] In context of nineteenth-century periodicals, this question responds not only to a contemporary possibility but a historical exigency, as the periodical archive enfolds another vastness which we have yet to approach: its illustrations.

With the advent of end-grain wood-block engraving techniques in the early 1800s, periodical publishers could generate pre-photographic images at the same industrialized scale that awed commentators then and now. *The Penny Magazine*, founded in 1832 and among the first titles to exploit the technique, claimed circulations of 200,000 copies for its issues, and sold stereotyped copies of many of its image blocks for further reprinting abroad. The dramatic expansion of periodical illustration would transform Victorian visual culture. The *Illustrated London News*, founded in 1842 and quickly developing success on a similar scale, valued its illustrations as having "an impetus and rapidity almost coequal with the gigantic power of steam," promising (to its own advantage) that "there is now no staying the advance of this art into all the departments of our social system."[9] As Patricia Anderson argues, the subsequent diffusion of the "new inexpensive printed image thus became the first medium of regular, ongoing, mass communication."[10] Produced by a variety of periodicals on the same scales which transformed textual genres into a vast wilderness, illustrations turned nineteenth-century visual culture into a mass phenomenon. As these historical materials become digitized and more available for study, we confront again their extent and ponder methods of researching a massive collection of visual materials too extensive to see.

The challenges of searching, sorting, comparing, and analyzing a sprawling world of digital images have spurred contemporary developments in computer vision and image processing techniques. If the nineteenth-century's textual vastness demands a more capacious approach to reading, so its proliferating printed images might similarly invite a "distant seeing," or at least consideration of how computer vision techniques might be adapted for studying historical illustrations.[11] Julia Thomas has argued that "illustrations dominated the cultural landscape of the nineteenth century" and yet remain paradoxically "invisible" in scholarship

in *Media Studies Futures*, ed. Kelly Gates (Blackwell, 2012).

[8] Elizabeth Lorang, "Text and/as Image: Expanding Possibilities for Description, Discovery, and Analysis in Digital Collections" (Collections as Data: Stewardship and Use Models to Enhance Access, Library of Congress, September 27, 2016).

[9] "Our Address," *Illustrated London News* 1, no. 1 (May 14, 1842): 1.

[10] Patricia Anderson, *The Printed Image and the Transformation of Popular Culture, 1790-1860.* (Oxford: Clarendon; Oxford University Press, 1994), 3.

[11] See the proposal for "distant viewing" by Taylor Arnold and Lauren Tilton, "Distant Viewing: Analyzing Large Visual Corpora," 2017.

on the period.[12] With a few exceptions, nineteenth-century periodical illustration has likewise remained a blind spot in computer vision research, even though it made possible the very conditions of mass viewing which image analytics rises to investigate.

The Illustrated Image Analytics project at NC State University has been experimenting with how computer vision research might be adapted to the study of historical illustrations, using a collection of Victorian newspapers. Our corpus includes newspapers which employed several modalities of image production, including wood engraving as well as photo-process techniques, all of which present unique challenges to the interpreter. In this article, we describe the progress of our experiments and the methodological reflections they generate. Our article, like our early research, pursues exploratory data analysis rather than a set of hypotheses, as we attempted to determine just how computer vision and image processing techniques might be adapted for large-scale interpretation of historical materials. Ultimately, we propose an approach related to Lev Manovich's methods of sorting a digital image collection at various scales using low-level image features, as opposed to machine-learned attempts to identify image content. We also reflect on how our case studies might suggest new avenues for illustration studies and periodicals research, providing historical insights as well as adjusting our narratives about the past.

Images, Extraction, and Visual Evidence

In 2014, our university library signed a license with Gale Cengage allowing "content mining rights" for its collection of *British Nineteenth-Century Newspapers*. This phrase means to compass the fullest range of analytical techniques which researchers might employ, including beyond text. This collection includes three illustrated newspapers: the *Penny Illustrated Paper* (1861-1913), *The Illustrated Police News* (1864-1938), and *The Graphic* (1869-1932). The collection only contains issues published before 1901 with some minor gaps (1869 for the *Graphic*; 1875 and 1891 for the *Illustrated Police News*).[13] These three titles hardly represent the extent and variety of illustrated nineteenth-century periodicals which

[12] Julia Thomas, *Nineteenth Century Illustration and the Digital: Studies in Word and Image* (Palgrave Macmillan, 2017), 17.

[13] For a more discussion of what the collection includes and lacks, see Bob Nicholson, "Counting Culture; or, How to Read Victorian Newspapers from a Distance," *Journal of Victorian Culture* 17, no. 2 (2012): 243-44.

have other, important stories to tell, such as the early visual experiments of the *Illuminated Magazine* (1843-44) or the flourishing of illustration in literary periodicals signaled by *Once a Week* (1859-1880).[14] That said, these three newspapers do index some important differences in editorial approach to illustration. *The Graphic* was launched as a direct competitor to the *Illustrated London News*, reporting news while attempting to further ennoble wood-engraved illustrations as a form of graphic art.[15] *The Illustrated Police News* (*IPN*) was a sensational Sunday paper, publishing crime news with somewhat lower-grade images but which are often more visually complex, blending multiple styles and textual elements in their illustrations. *The Penny Illustrated Paper* was among several attempts to copy the *Illustrated London News*'s format while undercutting its price. And all three titles emerged after the abolition of the Stamp Act (1855) and amid a flourishing of Victorian illustration more commonly associated with books.[16]

Our three titles result primarily from what data we could access—data that has also been shaped by the history of how British Library newspapers were digitized.[17] Newspaper pages were scanned from microfilm, most of it newly shot for the digitization process (including the *Graphic* and *IPN*). However, the *Penny Illustrated Magazine* was among the few titles scanned from legacy microfilm made in the 1950s, which resulted in low-quality digital images unsuited for our computational processes. Even with the higher quality scans of the *Graphic* and *IPN*, some pages are unevenly lit or poorly exposed, appearing far lighter or darker than the clarity of their printed versions. Seeking additional illustrated newspapers and potentially better quality images, we contacted the HathiTrust Research Center about access to their collections. Unfortunately, their online PDFs introduce image noise, or aliasing due to under-sampling which makes them far less suitable for analysis. Access to the original scanned image files, many of them made by Google, is complicated by licensing and large file sizes. Our own collection is similarly restricted by the terms of Gale's license with NCSU Libraries, in that we cannot freely share the source data. In many ways, our project exemplifies the problems of access to historical content ostensibly in the public domain. But as our initial goals were methodological, simply figuring out what approach

[14] Brian Maidment, "The Illuminated Magazine and the Triumph of Wood Engraving," in *The Lure of Illustration in the Nineteenth Century: Picture and Press*, ed. Laurel Brake and Marysa Demoor (Basingstoke [England] ; New York: Palgrave Macmillan, 2009), 17-39; Linda K. Hughes, "Inventing Poetry and Pictorialism in Once a Week: A Magazine of Visual Effects," *Victorian Poetry* 48, no. 1 (May 9, 2010): 41-72.

[15] W.L. Thomas, "The Making of the 'Graphic,' " *Universal Review* 2 (December 1888): 80-93.

[16] See for example Simon Cooke and Paul Goldman, eds., *Reading Victorian Illustration, 1855-1875: Spoils of the Lumber Room* (Farnham, Surrey, England: Ashgate, 2012).

[17] For a fuller discussion, see Paul Fyfe, "An Archaeology of Victorian Newspapers," *Victorian Periodicals Review* 49, no. 4 (December 29, 2016): 546-77.

might even yield interesting results, we chose to make do with what we had in hopes these experiments suggest ways of approaching other collections.

Our first task was to identify and extract distinct illustrations from the data set. Gale's data for the collection comprise XML files with OCR text and metadata for each newspaper issue, along with high-resolution page facsimile images in JPG and TIFF formats. Page counts for each newspaper in our collection range into the tens of thousands; each page may contain several illustrations, a full page picture, or none at all. Gale's product does offer page zoning for articles and illustrations by pixel location, which it uses to segment and highlight different items browsed or found through the database's web interface. But we chose to identify and extract illustrations from page facsimile images using open-source code developed for OpenCV (https://github.com/acdha/image-mining/). The scanned newspaper page is first converted into a black and white image using a threshold. The Canny edge detector was applied to find a set of contours.[18] Finally, the contours which are too small or too large were filtered out to remove text and borders of the newspaper page. The illustrations were obtained by finding the bounding boxes of the remaining contours on the image. This results in significant numbers of image files for each newspaper: roughly 70,000 for *The Graphic*, 8,600 for the *IPN*, and 70,000 for the *Penny Illustrated Paper*.

Even at this first step—declaring what is an image—we introduce an interpretive bias, as in any attempt to model and analyze historical images computationally. Defining the boundaries of an illustration already supposes a visual ontology which nineteenth-century sketch artists, engravers, and readers did not necessarily observe. For example, the *IPN* often combines several different modalities into a single tableau: portraits of the criminals, a narrative scene of the crime, textual labels:

[18] J. Canny, "A Computational Approach to Edge Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986.

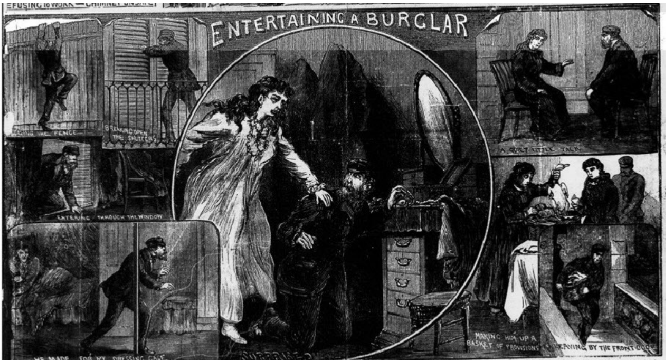

ENTERTAINING A BURGLAR
BREAKING OPEN THE SHUTTERS
ENTERING THROUGH THE WINDOW
A QUIET LITTLE TALK
SURPRISED
MAKING HIM UP A BASKET OF PROVISIONS
LEAVING BY THE FRONT DOOR

Figure 1. “Entertaining a Burglar.” IPN January 6, 1883, 1.

Is this one illustration or many? As Brian Maidment suggests, wood-engraved illustrations mixed a historical legacy of representational codes, including from comic and satirical prints, diagrams for “useful knowledge,” gothic broadsides, and emerging social documentary modes. At a more fundamental level, can an illustration even be separated from text? Scholars of nineteenth-century illustration have argued that these images only become meaningful in their relations to text. As Linda Hughes claims, an illustration “is not self-sufficient; the image changes (by contextualizing) the text, but the text conditions what is seeable in the image.”[19] For Maidment, mass circulation does not liberate the image but helps cement the “crucial interdependence of text and image.”[20] Lorraine Janzen Kooistra suggests that nineteenth-century illustrated periodicals and gift books were distinguished by their “hybridity” of text, image, and crafted materials.[21] Julia Thomas claims that the “alternation between the visual and textual” defines the very experience of pictorial values in the Victorian era.[22] To extract an illustration from its context on the page, from its location within an issue, and from the discursive conditions of historical seeing all risk evacuating the image’s significance. There is an interpretive violence in “chunking” the visual experience of illustrated periodicals into discrete objects, then applying computational measures which subordinate complex visual experiences to quantitative parameters.[23] These processes are operational necessities for our project, but may also

[19] Hughes, ”Inventing Poetry and Pictorialism in *Once a Week*,” 46.

[20] Brian Maidment, *Reading Popular Prints, 1790-1870* (Manchester: Manchester University Press, 1996), 9.

[21] Lorraine Janzen Kooistra, *Poetry, Pictures, and Popular Publishing: The Illustrated Gift Book and Victorian Visual Culture, 1855-1875* (Athens: Ohio University Press, 2011), 24.

[22] Julia Thomas, *Pictorial Victorians: The Inscription of Values in Word and Image* (Athens: Ohio University Press, 2004), 6.

[23] Johanna Drucker, ”Humanities Approaches to Graphical Display,” *Digital Humanities Quarterly*

suggest alternative ways of defining the study of illustrations.

What does an image mean? As Maidment persuasively argues, historical illustrations cannot simply serve as evidence of things seen. Even though many illustrated newspapers insisted on the accuracy and fidelity of their images, these illustrations were thoroughly mediated from sketches to reworking on the block. Their content was often fabricated, sometimes based upon other prints or illustrations, sometimes envisioned from verbal testimonies or textual reports.[24] What we "see" in a newspaper illustration is less a historical record than a record of historical ways of seeing. As Maidment claims, the Victorian illustration "must be understood both as a naturalistic representational medium ... *and* as a shorthand non-naturalistic visual code built out of long traditions."[25] In *Graphesis*, Johanna Drucker emphasizes how the very concept of the visual has its own histories. Visual epistemologies change with cultural circumstance, or the "conventions, habits of reading and thought, and graphical expressions whose properties translate into semantic value."[26] It is not simply that we see with different eyes, but that we differently understand the very possibility of visual knowledge.

Add to these complications the technological layer of the image's remediation into microfilm and then digital forms, rendered through a variety of compression and display processes, and approached in a new context of its computational tractability as data. What are we seeing at all? What exactly is our data evidence of? While the extent of these biases seem enormous, they can also effect a powerful shift in researchers' points of view, a "productive unease" that may reveal other interpretive angles.[27] We are indeed seeing in a new way, with all the potential advantages of that viewpoint. In his own work visualizing collections of media, Lev Manovich describes the process as "a powerful mechanism of defa-

5, no. 1 (Winter 2011).

[24] For example, consider the first issue of the *Illustrated London News* (1842) including its triumphant announcement about its images' fidelity and its fabricated illustrations, including a fire in Hamburg for which the "illustrator had simply copied the original topographical print of the city [from the British Museum] and added flames and onlookers. Images of the royal fancy dress ball in the inaugural issue were based entirely on written press reports." See Gerry Beegan, *The Mass Image: A Social History of Photomechanical Reproduction in Victorian London* (Basingstoke; New York: Palgrave Macmillan, 2008), 54.

[25] Maidment, *Reading Popular Prints, 1790-1870*, 15. Joshua Brown describes this in similar terms: "rather than being a rigid representational form, illustrated journalism was constituted by a complex interaction between the creation and viewing of images that changed" over time; see *Beyond the Lines: Pictorial Reporting, Everyday Life, and the Crisis of Gilded Age America* (Berkeley: University of California Press, 2002), 4.

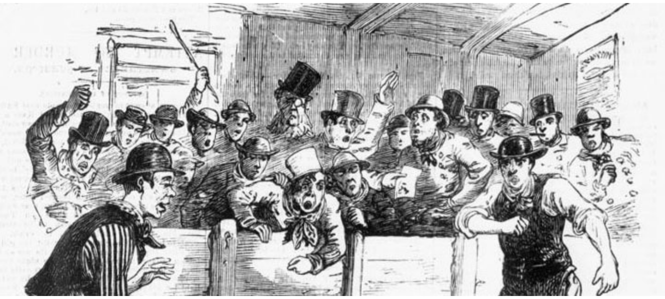
[26] Johanna Drucker, *Graphesis: Visual Forms of Knowledge Production*, MetaLABprojects (Cambridge: Harvard University Press, 2014), 51.

[27] Julia Flanders, "The Productive Unease of 21st-Century Digital Scholarship," *Digital Humanities Quarterly* 3, no. 3 (Summer 2009).

miliarisation ('ostranenie')-a device for seeing what we could have not noticed previously."[28] For Manovich as well as our own project, that process starts with transforming images into data, then transforming that data into new images or visualizations.

Content Analysis

We began our experiments by trying image processing and recognition methods we were already familiar with, attempting to test what techniques might succeed (or not) on historical illustrations. We started with face recognition, image segmentation, and object recognition, using "out-of-the-box" and open software, such as the OpenCV face detection using Haar Cascades as well as Adam Geitgey's face recognition based on dlib (https://github.com/ageitgey/face_recognition). Initial results were poor: out of a sample of 229 images with faces we selected from the *Graphic*, the face recognition software returned rates of true positives at 0.3144, false positives at 0.0131, and false negatives at 0.6856. The *IPN* scored even worse: out of a sample 328 images of faces, we got true positives at 0.2927, false positives at 0.0061, and false negatives at 0.7073. These techniques all use libraries of sample images (notably photographs) for training and comparison. Inputing images of historical illustration without adjusting that training data was naïve, though still useful to understand how much trouble such algorithms immediately have with these images, especially pictures drawn in lines. Consider the formidable challenge of the following illustration for face detection:



[28] Lev Manovich, "How to Compare One Million Images?," in *Understanding Digital Humanities*, ed. David M Berry (Basingstoke: Palgrave, 2012), 276.

Figure 2. detail from "Capture of a gang of dog fighters in Gravel Lane." IPN January 19, 1867, 1.

In Figure 2, the sketch artist and/or engraver renders faces in several different ways, from a quick sketch of simplified faces in the background to outright caricatures. Faces also appear at different sizes, not just within an illustration, but relative to the engraved line. In his work on the printed image, William Ivins introduces the concept of "tolerances" to explain the representational leeway in such reprographic methods.[29] These tolerances are especially noticeable in shading, for which illustrators had to use cross-hatching and stipple techniques to create tonal range. These textures are more pronounced in smaller-sized images, as in Figure 3:

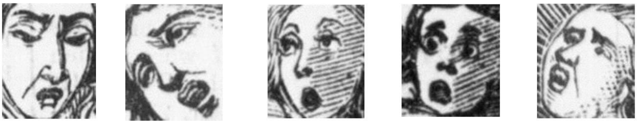


Figure 3. Extracted faces from IPN: January 5, January 12, and January 19, 1867 issues.

These close-ups suggest the how the tolerances of wood-engraved images, especially at smaller scales, introduce ambiguities in tonal range, backgrounds, features, and varied details, all deriving from combinations of lines, all engraved with an eye toward managing the clarity of a printed impression.

We also hoped to apply face recognition to the data set, tracking the appearance and recurrence of famous figures in the news, or perhaps the patterns of facial individuation in nineteenth-century illustration. There is some precedent for doing so. In their study of two illustrated nineteenth-century U.S. periodicals, Kevin Barnhurst and John Nerone claim that only political celebrities were individuated, and everyone else was rendered as a "personage": "a relatively fixed set of traits that spring from social class, race, position of power, physiognomy, style of dress, and personality."[30] But for this claim and others about large-scale changes, Barnhurst and Nerone do not explain their research methods or their standards for what, or how much, qualifies as evidence. With careful attention to individual engravings, Joshua Brown argues for large scale shifts in illustrating political celebrities, from the early simulation of "emulatory photographs" including "poses depicting solemn features and gazes averted in timeless reflec-

[29] William M Ivins, *Prints and Visual Communication* (Cambridge: MIT Press, 2001), 48.

[30] Kevin G. Barnhurst and John Nerone, "Civic Picturing vs. Realist Photojournalism the Regime of Illustrated News, 1856-1901," *Design Issues* 16, no. 1 (2000): 72.

tion" (partly due to the long-exposure time of commercial photography) to more idiosyncratic, and sometimes melodramatic, "character sketches."[31] These scholars propose several hypotheses which might be computationally testable, or at least provide launching points for further experiments. For instance, should facial recognition succeed, even its errors would be instructive, used to consider representational patterns in era in which phrenology and facial physiognomies were both explicitly and tacitly accepted as signs.

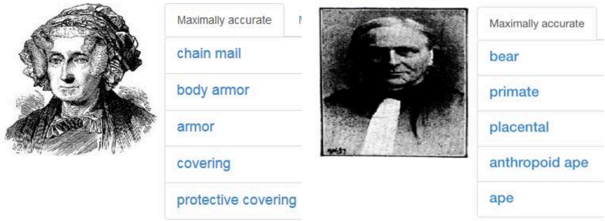
To make learning-based face detection algorithms work on our dataset, or even to measure the performance of such algorithms, we would need a training set of illustrations with each face labeled by a bounding box. A similarly labeled data set would be needed for object detection in these images, which we also hoped to try. But labeled datasets with these levels of granularity do not currently exist, and we despaired at the magnitude of labor required to create, test, and refine them, given the limits of funding and our interests in exploring multiple approaches. Seeking a workaround, we approached researchers involved with the Database of Mid-Victorian Illustration and the Illustration Archive, each of which "tags" images by hand according to a typology. Because of "the lack of a fixed bibliographic convention for describing illustration," as Julia Thomas explains, these projects created their own vocabularies and, with the Illustration Archive, relied on crowdsourced tagging to describe visual content.[32] Those tags may not be consistently applied, nor do they label specific locations within the images, which limits their utility for identifying image features. These projects were designed for human users to apply categories and sort by them, rather than for machine analysis.

We also considered trying some unsupervised approaches such as image segmentation and machine learning. For the lineated images especially, region-based segmentation is non-trivial because adjacent regions have similar texture and lack clear boundaries. Also, the engraved lines all over the images prevent the use of edge clues for segmentation. Seeking a different approach to classifying images based on content, we tried using an image classification network implemented by Caffe, a deep learning framework. Our initial test used Caffe's web-based demo (http://demo.caffe.berkeleyvision.org/), again without providing additional training data. Out of the box, this network can classify clear and simple pictures, such as a cat or a water pitcher. It struggles with even slightly more complex illustrations, thrown off by the hatchwork of engraved lines which it misreads, as in Figure 3, as "chainmail"; or confused by the hazier reproductions of nineteenth-century reprographic media, as in Figure 4: a halftone portrait it

[31] Brown, *Beyond the Lines*, 78, 167.

[32] Thomas, *Nineteenth Century Illustration and the Digital*, 40.

bafflingly understands as "placental" and "primate."



Maximally accurate

chain mail

body armor

armor

covering

protective covering

Maximally accurate

bear

primate

placental

anthropoid ape

ape

Figure 4. Screenshots from Caffe's web image classification demo.

Of course, deep learning algorithms must be trained on a tagged data set, "taught" to identify positive and negative examples. Again, we decided to pursue other approaches. Yet there are some promising developments in the field which suggest pathways for content-based image analysis using historical materials. For example, Giles Bergel and colleagues at the University of Oxford and the Bodleain Libraries developed an image recognition tool and semantic tagging for early modern broadsides and ballads.[33] Thomas Smits and Melvin Wevers working at the Dutch National Library have used a convolutional neural network to classify images from Dutch newspapers 1860-1940.[34] Neural Neighbors, a project of the Yale Digital Humanities Lab, also uses this technique to find pictorial tropes in the Beinecke Library's digital collection of historical photographs.[35] Andrew Piper, Chad Wellmon, and Mohamed Cheriet are steering the Visibility of Knowledge project to computationally identify and study nineteenth-century scientific illustrations in digital collections of books and periodicals.[36] Still, we wonder whether machine learning frameworks can compensate for the incredible variety of styles and visual information in illustrated periodicals, or how machine-learned classifications would align (or not) with the visual epistemologies of Victorian viewers. It is possible that these high-level visual features, if detectable, would provide further insights to our data, but, since we could not test them, we turned instead to what we could test with more promising results.

[33] Giles Bergel et al., "Content-Based Image Recognition on Printed Broadside Ballads: The Bodleian Libraries' ImageMatch Tool" (IFLA World Library and Information Congress, Singapore: IFLA Library, 2013).

[34] Thomas Smits, "Illustrations to Photographs: Using Computer Vision to Analyse News Pictures in Dutch Newspapers, 1860-1940" (DH 2017, Montréal, 2017).

[35] Yale Digital Humanities Lab, "Neural Neighbors: Pictorial Tropes in the Meserve-Kunhardt Collection," Yale Digital Humanities Lab, 2017.

[36] Andrew Piper, "The Visibility of Knowledge," *.TxtLAB @ McGill* (blog), September 15, 2016.

From Content to Image Features

In spring 2016, we presented our initial work to our funder in a room full of computer scientists and data analysts. One person came up afterwards, fascinated and flummoxed by how these seemingly self-evident illustrations defied the power of image processing algorithms to reckon with them. As he asked, "Why can humans recognize this stuff and computers cannot?" The question gets to the very heart of computer vision research. But it also formulates the problem in a way which makes it impossible to answer. The question assumes that our visual processes can transpose onto computational processes, which, were that even possible, we do not yet know how to do. Given what we can currently measure, we had been asking questions of the data in the wrong way. Another person pointed out a crucial distinction: between semantic-level analytics, in which you ask the computer to recognize something based on an understanding of what it is, and pixel-level analytics, in which you let the computer do what it is currently good at: measure stuff. Digital image processing, as Manovich explains, can aid the researcher by measuring visual features which either escape the human eye or, in a large-enough collection of images, exceed our ability to apprehend. For assessing image content, the human eye remains far better at visual interpretation, at negotiating multiple and complex modalities of representation.[37] Might we achieve more interesting results with a synthesis of what humans and computers each do well?

The turning point for our research came with treating images in a very different way: not as representations of seeable things, but as tractable data at the level of pixels. Instead of asking what do images show, we started asking what rudimentary "features" they comprise. One such measurement is pixel ratio, or simply how many light relative to dark pixels appear in an image. Perhaps this is not such a distortion. It could be argued that, until lithographic techniques were adapted for large-scale commercial printing in the early 1900s, relief printing is fundamentally binary: a raised surface prints, a cut-away portion does not. At the level of inscription, printing is black or white, inked or not, on or off. Digital images translate this information into pixels, whose configurations can be measured and compared seperately of what they represent. Returning to our data set, we converted the grayscale images into binary images (black and white) using a threshold of grayscale intensity 150 for both the *Graphic* and *IPN*. Then

[37] Manovich, "How to Compare One Million Images?," 251, 259.

the pixel ratio *R* for each image was computed by $R = \frac{Nthr}{N}$, where N_{thr} is the number of pixels with intensity below the threshold and is the total number of pixels in the image. While many programming languages might suffice for this work, Qian Ge used MATLAB to process all of the extracted images and generate sorted lists of images as well as distribution charts. We ran a similar process to measure another low-level feature: entropy. The entropy E of images is defined as $E = -\sum_i P_i \, log_2 P_i$, where P_i is probability of pixels with intensity i. A concept in information science, entropy is a measure of information as a range of possible signals. An image that only uses a limited range of pixels has low entropy; these usually look very light or very dark. An image saturated with a broad range of pixel values has high entropy. These images usually yield the more complex representations in the data set as well as capture many halftone images.

Though a simple process, sorting images by low-level features operationalizes a much different interpretive approach to nineteenth-century illustration than has been active in the field. Instead of studying selected images in their relation to texts, we batch processed images without initial regard to what they represented, their textual surroundings, or even their searchable metadata. This breaks the necessary relationship of image and text within illustration studies while suggesting a new approach to many of its abiding questions, including stylistics, interpictoriality, and graphical languages. Futhermore, this approach does not privilege its numerical measures but uses those measurements to create visualizations, inviting the human eye to reconsider these materials from a different remove. Natalie Houston underscores how visualization can help with exploratory data analysis, harnessing "the pattern-recognition powers of the human brain in order to enable comparison, differentiation, segmentation" within complex data sets.[38] Manovich has insisted on the value of visualizing not just data but large media collections, using images to describe images at a different scale.[39] Rather than privileging close or distant perspectives, Manovich recommends that researchers "compress massive media universes into smaller observable media 'landscapes,'" expressive of large-scale features while retaining enough detail to show "subtle patterns" in individual objects.[40]

Low-level parameterization of newspaper illustration enables just such an approach. While measures of image features do not by themselves tell you

[38] Natalie M Houston, "Visualizing the Cultural Field of Victorian Poetry," in *Virtual Victorians: Networks, Connections, Technologies*, ed. Veronica Alfano and Andrew M Stauffer (New York: Palgrave Macmillan, 2015), 125.

[39] Manovich, "How to Compare One Million Images?," 263.

[40] Manovich, "Media Visualization."

much about any single illustration, at slightly larger scales, they generate some intriguing clusters and trends, helping the researcher to compose mid-scale “landscapes” of visual patterns to interpret, including those graphs which Manovich calls “image plots.” For example, we sorted the *IPN* by pixel ratio and entropy. By imposing a set of sample images upon the chart, we noticed some distinct clusters (Figure 5).

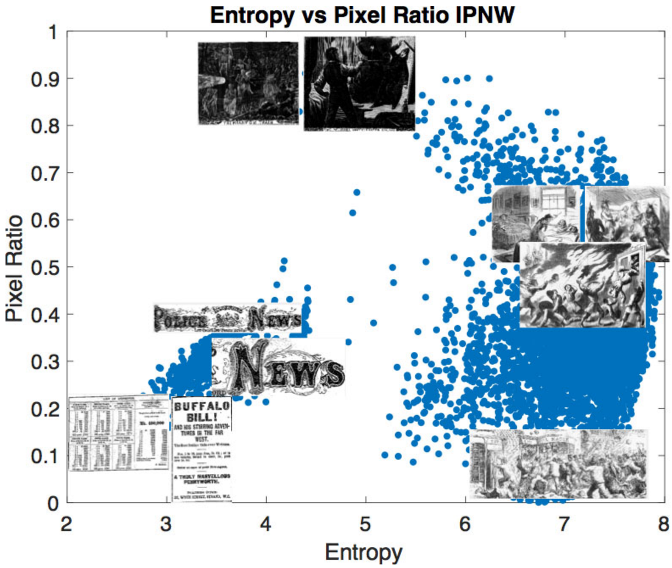


Figure 5. Plot with sample images of entropy vs. pixel ratio in the IPN.

Lacking a labeled data set, we do not measure the quantitative performance of the graph but use it instead as an exploratory visualization. We noticed that low pixel ratio, low entropy images tended to capture tabular information and display advertisements. It may be possible to collect large numbers of advertisements based on these parameters, rescuing for study a form of newspaper content which is frequently overlooked. This cluster also raises questions about the very thresholds of text and image, or when typography becomes a graphic of type, including wooden display typefaces and nineteenth-century “ASCII art.”[41] Our image extraction algorithm tried to harvest illustrations alone, but in also identifying these text-based display advertisements as images, it prompts us to consider modalities of graphical display in a broader sense.

In this case and others, we are using image processing and visualization to ask

[41] Jacob Harris, ”Solving a Century-Old Typographical Mystery,” *The Atlantic*, May 23, 2016.

questions of our data, exploring the research possibilities which unfold from using these techniques. Other clusters within the *IPN*'s distribution of features show an opportunity to focus on image content, as the human eye detects semantic patterns in an otherwise rudimentary graph of image features. For example, high-pixel ratio images tend to capture lots of depictions of night scenes (Figure 6).

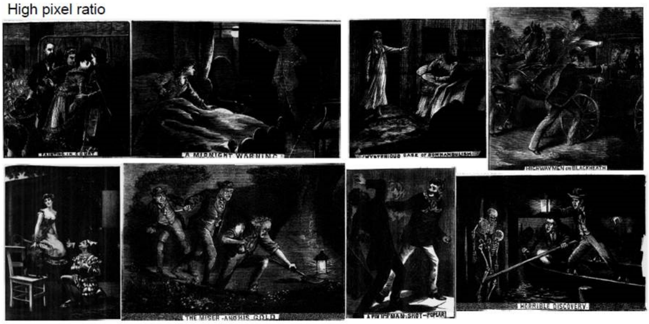


Figure 6. Sample images from high pixel ratio cluster in the IPN.

The computer sorts by pixels; the human interpreter considers the unfolding possibilities for historical interpretation, perhaps including reflections on vision and surveillance in illustrated periodicals, how these intersect with the genres of sensation in the nineteenth century, or even the cultural history of darkness. At this scale, we noticed subtler patterns in the images, too, such as the *IPN*'s own attempt to classify these images with generic captions. Consider the "Horrible Discovery" in the bottom right of Figure 6. Even as the sensational discovery of quayside skeletons in chains rattles us as a most singular event (or as "news"), the *IPN* presents it within a recurring generic category by which readers assimilate new information. That visual dynamic between generic framework and singular instance, on a small scale, expresses the dynamic of seriality and miscellaneity that James Mussell finds at the heart of periodical culture.[42] Seeing clusters of images exposes how that dynamic works within micro-genres of periodical illustration.

We approached the *Graphic* in the same way: computing its image features, then visualizing them in different ways.[43] After sorting the *Graphic* by pixel ratio, we noticed a set of maps in the low range (figure 7). As line diagrams, they have far

[42] Mussell, *The Nineteenth-Century Press in the Digital Age*, 24.

[43] This method of "digital image processing + visualisation" forms the core of Manovich's own definition of "cultural analytics." See "How to Compare One Million Images?," 262.

more lights than darks. But the *Graphic* is hardly known for geo-graphic representations which are absent from scholarship about the periodical.

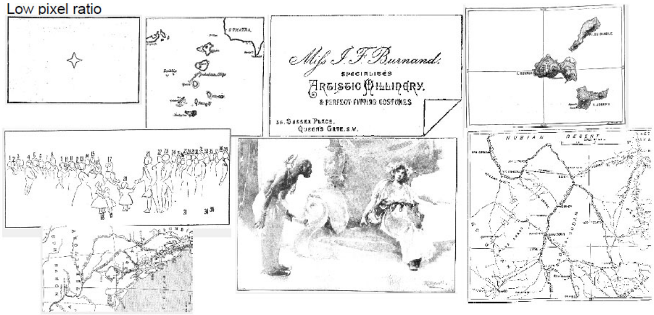


Figure 7. Sample images from low pixel ratio cluster in the IPN.

Why was it printing maps, and for what places and occasions? What geographical understanding could the *Graphic* presume in its readers? What political imaginary might these maps help to construct? And how might their production link to a broader discourse about graphic knowledge in the nineteenth century? The *Graphic*'s maps, captured within a collection of low pixel-ratio images, unfold into a world of questions about the history of visualization, geographical knowledge, and international networks of seeing.

In the high range for entropy in the *Graphic*'s images, we found the significant presence of halftone illustrations (figure 8).

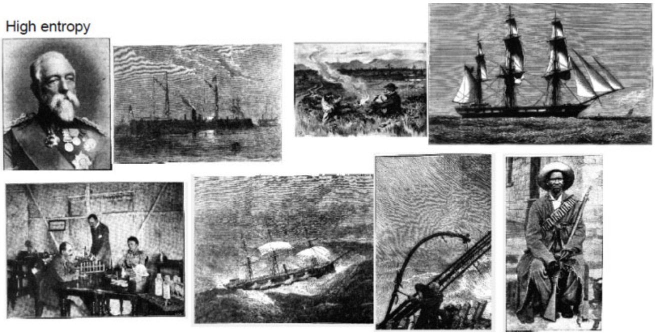


Figure 8. Sample images from the high entropy cluster of the Graphic.

The Graphic debuted the halftone technique for British periodicals in 1884, and it became widely adopted in the years following. The halftone turned photographs into relief printing plates by exposing them through a screen. This results in a mesh of relatively larger or smaller type-high dots which, when inked and printed, imitate photography's tonal range.[44] When magnified for the eye, or analyzed at the pixel-level, a halftone can be readily distinguished from a line engraving. Liu, Guo, and Lee have created an algorithm to identify halftone images by measuring selected patches from an image background, then comparing these against a threshold for entropy level. We adapted their technique to identify halftone images in the *Graphic*. Because of the *Graphic*'s reputation, some of these images are already known to scholars, such as its first halftone "The Midnight Sun" (September 6, 1884) which depicts the lowest point of the winter sun in Norway and wonderfully thematizes its own breakthrough as, quite literally, a "sun picture" (as in the earliest descriptions of photography). [45] It was also a one-off. According to the algorithm, halftones next appear a year later in the article "An Amateur Photographer at the Zoo" (September 5, 1885). Its multiple halftones are made from "instantaneous photographs" by C.J. Hinxman. As David Reed argues, this four-page story "has a substantial claim to be considered as the first photo-picture story to be published in any periodical in the world."[46] Photographic histories have a tendency to celebrate firsts in this manner, but halftone detection can also be used to identify larger batches and longer trends.

For instance, we measured the relative numbers of illustrations the *Graphic* published using different reprographic techniques. Standard narratives in the history of print illustration suggest that wood engraving effectively dies out by the end of the nineteenth century. Our results show an increasing presence of halftone images in the *Graphic* starting in the 1880s and achieving equal levels with wood engravings by 1900 (figure 9).

[44] For a thorough discussion of the process, see Dusan C. Stulik, "The Atlas of Analytical Signatures of Photographic Processes: Halftone" (Los Angeles: Getty Conservation Institute, 2013).

[45] David Reed, *The Popular Magazine in Britain and the United States, 1880-1960* (Toronto: University of Toronto Press, 1997), 34.

[46] Reed, *The Popular Magazine,* 35.

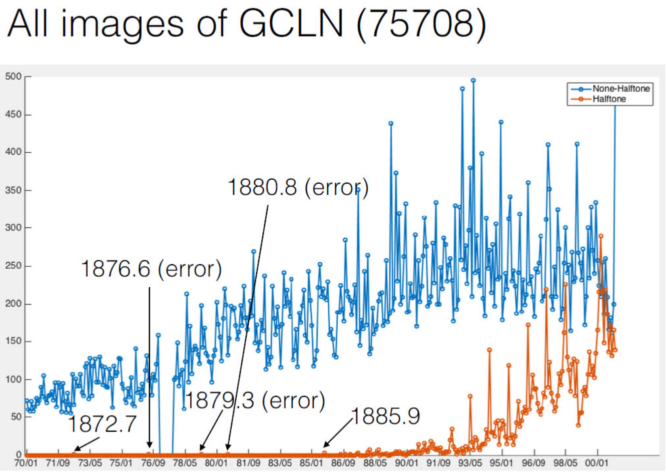


Figure 9. Illustrations classified as halftone or not in the Graphic, 1870-1900.

At least by 1900—when our data set ends—the *Graphic* continues to produce wood engravings and halftones in equal measure, though the trajectories are clear. The graph also lets us focus on the outliers like "firsts" and the peaks which turn out to be special issues, as well as confirm or add nuance to larger historical trends.

Halftone identification also returns results which complicate the story of late nineteenth-century illustration. Figure 10 was correctly identified as a halftone; when magnified, its telltale dot patterns can easily be seen. But the image is definitely not a photograph, considering the non-blurred action and how all the depth of field remains in focus.

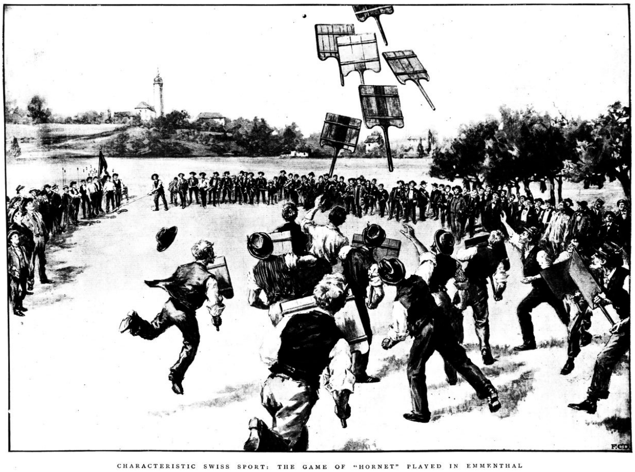

F.C.D.
CHARACTERISTIC SWISS SPORT: THE GAME OF "HORNET" PLAYED IN EMMENTHAL

Figure 10. Halftone/lithograph by "F.C.D." from The Graphic, August 4, 1900, 163.

The illustration was printed with a halftone process, but the source was a lithograph, another emerging technology of image making in the late nineteenth century. Basically, a hand-drawn print was photographed and then turned into a halftone relief plate. Scholars like Gary Beegan have pointed out how, during moments of technological shift, image-making techniques will adapt methods and styles from each other, complicating our understanding of the mass image as a hybrid, something not determined by a breakthrough technology.[47] In returning results for halftones which do not otherwise distinguish their image-making technique, the algorithm may help expose these interrelationships. Furthermore, as these medial complexities currently exceed computer vision's grasp, they suggest a development frontier for image analysis.

[47] Beegan, *The Mass Image*.

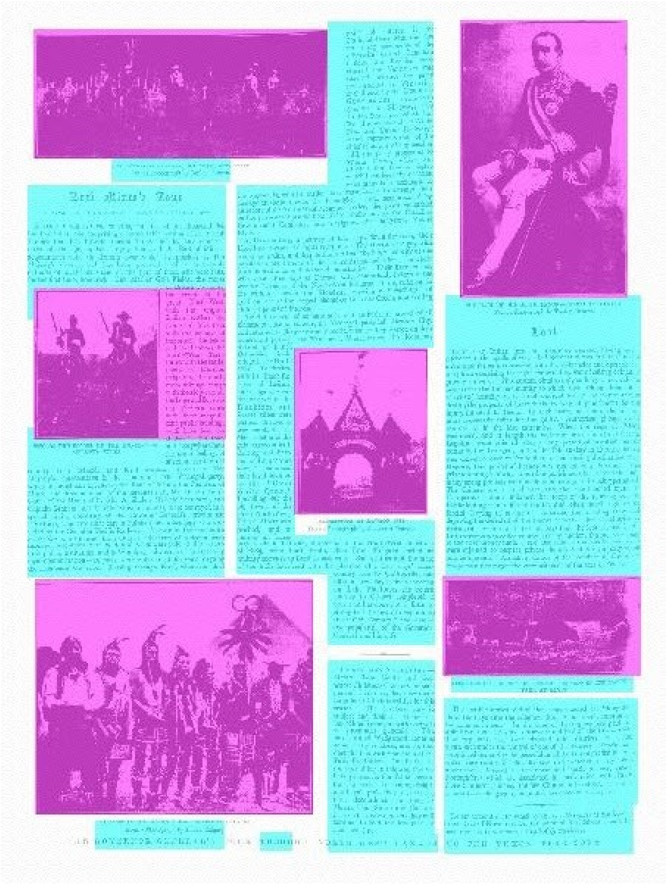


Figure 11. Sample page decomposition by text vs. image area.

In a related approach, we used image processing to explore regularities and changes in how whole pages of illustrated newspapers looked. Rather than examining only the extracted illustrations, entire page facsimile images can be treated as units of visual experience, as in related work on periodical page formats and bibliographic codes by Elizabeth Lorang, Natalie Houston, and Dallas Liddle.[48] In taking measurements of whole pages, we first used the initial

[48] For a variety of reasons, digitization methods have privileged the single page as a discrete research object. It still remains to treat not the single page, but the opening or two-page spread, as a visual unit. However, reconstructing two-page spreads is very difficult to automate: page-length can vary per issue; digital collections do not reliably include every page; the *Graphic* printed illustrations which spanned two facing pages which the digital collection collapses into a single object. Thus, our efforts

figure extraction algorithm to identify zones for images, and a related algorithm to identify textual materials (figure 11). We then computed several ratios based on those pixel dimensions, including the ratios of text and image to entire page size, and the ratio of text and image to each other. This method does admit some uncertainties, including the accuracy of automatically zoned text and image areas, and the inconsistency of page sizes as originally digitized and cropped in the digital collection. Because of this, we found the image/text ratio more useful as a heuristic, offering a glimpse at the changing modalities within a given periodical. Figure 12 shows the ratio for the *Graphic* during 1870-1900.

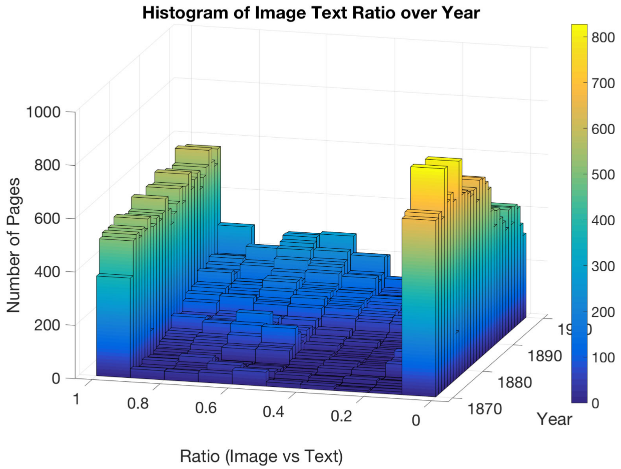


Figure 12. Image vs. text ratio in the Graphic, 1870-1900.

The chart shows the relative distribution of image or textual material on any given page. In 1870 (nearest the front of the graph), most pages were either fully devoted either to images or to text content. However, in the decades following (moving toward the back of the graph), that distribution becomes more mixed as the *Graphic* diversified its page layouts. Text and image become more integrated. This change could be related to several factors. For instance, before the application of stereotyping to pages with wood engravings, large volume periodicals often had to use two separate printing processes: a flat press for heavily illustrated pages and a cylindrical press for the rest. Once everything could be printed on the same cylindrical plate, page composition was greatly freed. Furthermore, the integration of text and image also reflects the gradual increase of visual display advertising in the *Graphic*. The changes also invite broader questions such as

to stitch together page spreads at scale were not yet reliable enough to measure.

about the emergence of graphic design which, as Drucker underscores, did not develop as an independent discipline until the early 1900s, previously consisting of tacit decisions made during the production process.[49] The shift towards integrated modalities on the page may track an increasing dependency of text and image, as Maidment has suggested, contrary to claims about how mass production loosens the image from its contexts.[50] It may also show display trends in late-nineteenth century's "new journalism" which, among its effects, shrunk news items to smaller "Tit Bits" (as in the eponymous periodical) and used bolder typographic and visual display to make the newspaper more accessible.

Describing images with images can also reveal what remains invisible to the computer and yet hides in plain sight, like the almost endless gallery of whiteness which *The Graphic* continuously enshrines in portrait vignettes (figure 13). These portraits were collected by computing GIST features or overall vectors for our image collections and sorting by similarity.

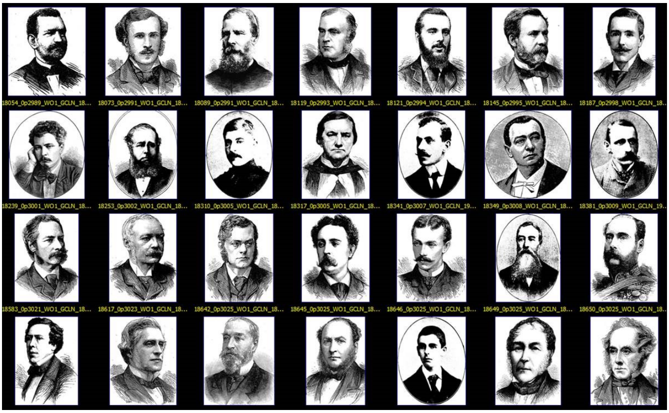

18054_0p2989_WO1_GCLN_18...
18073_0p2991_WO1_GCLN_18...
18089_0p2991_WO1_GCLN_18...
18119_0p2993_WO1_GCLN_18...
18121_0p2994_WO1_GCLN_18...
18145_0p2995_WO1_GCLN_18...
18187_0p2998_WO1_GCLN_18...
18239_0p3001_WO1_GCLN_18...
18253_0p3002_WO1_GCLN_18...
18310_0p3005_WO1_GCLN_18...
18317_0p3005_WO1_GCLN_18...
18341_0p3007_WO1_GCLN_19...
18349_0p3008_WO1_GCLN_18...
18381_0p3009_WO1_GCLN_19...
18583_0p3021_WO1_GCLN_18...
18617_0p3023_WO1_GCLN_18...
18642_0p3025_WO1_GCLN_18...
18645_0p3025_WO1_GCLN_18...
18646_0p3025_WO1_GCLN_18...
18649_0p3025_WO1_GCLN_18...
18650_0p3025_WO1_GCLN_18...

Figure 13. sample portrait images from The Graphic.

The computer sees the shape, but when visualized this way, we see much more. By sorting and showing illustrations by computational means, we can also reveal the ordinary, what goes unseen for being always seen, sometimes arriving with the shock of its homogeneity. We can also notice what is missing, many women's and black and native faces for instance, because they are absent from the data or from the collections we study. The absence of evidence significantly affects the research horizon for computational cultural history. As Ben Fagan argues

[49] Drucker, *Graphesis*, 26-27.

[50] Maidment, *Reading Popular Prints, 1790-1870*; Kate Flint, *The Victorians and the Visual Imagination* (Cambridge; New York: Cambridge University Press, 2000).

in “Chronicling White America,” analytics projects on digital newspaper collections may perpetuate the biases of the collection’s own history.[51] Furthermore, historical imagery conveys its own deep prejudices; as Joshua Brown and Sarah Blackwood claim, we must look beyond images entirely for better evidence of nineteenth-century African American visual culture.[52] Even should non-white faces appear in digital collections, they may very well be invisible to the computer. Researchers have exposed race and gender bias in technologies of visual representation reaching from machine learning algorithms to the normalization of color film.[53] Our illustrations originate from techniques called “white- and black-line engraving.” The biases of computer vision, in this context, look back to nineteenth-century image-making technologies which inscribed race as a highly visual phenomenon, yielding contrasting images which aspired to mark ontological difference.[54] The analysis of data whether at scales distant or close always presents a view from somewhere, as Lauren Klein has explained, and researchers must reckon with the problematics not simply of their tools but of their own politically-embedded scholarship.[55]

Throughout this article, we have raised a number of interpretive possibilities and problems to question the usefulness of computer vision techniques for studying historical illustrations. And also to underscore how relatively simple measurements might unfold into much more complex historical consideration, including of the visual epistemologies of the nineteenth century as well as the twenty-first. Our research means to emphasize a dynamic between basic image features and interpretive speculation that has been, at least for our project, the source of many productive conversations about what these measurements alternately distort and reveal. Compared to other methods used in digital humanities scholarship, computer vision techniques remain relatively difficult for scholars outside of image analytics research. Beginning with low-level image features can offer a more

[51] Benjamin Fagan, “Chronicling White America,” *American Periodicals: A Journal of History & Criticism* 26, no. 1 (March 30, 2016): 10-13.

[52] Brown, *Beyond the Lines*, 113; Sarah Blackwood, “ ‘Making Good Use of Our Eyes’: Nineteenth-Century African Americans Write Visual Culture,” *MELUS: The Journal of the Society for the Study of the Multi-Ethnic Literature of the United States* 39, no. 2 (2014): 42-65.

[53] Ian Tucker, ”‘A White Mask Worked Better’: Why Algorithms Are Not Colour Blind,” the Guardian, May 28, 2017; Joy Buolamwini and Timnit Gebru, ”Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” *Proceedings of Machine Learning Research* 81, no. 1 (2018): 1-15; Syreeta McFadden, ”Teaching The Camera To See My Skin,” *BuzzFeed* (blog), April 2, 2014. In this context, Tim Sherratt’s project *Invisible Australians* offers a powerful corrective, using computer vision techniques to repair racial bias in the historiography of Australia.

[54] See also Mandy Reid, ”Racial Profiling: Visualizing Racial Science on the Covers of *Uncle Tom’s Cabin*, 1852-1928,” *Nineteenth-Century Contexts* 30, no. 4 (December 1, 2008): 369-87.

[55] Lauren F Klein, ”Distant Reading after Moretti” (MLA Convention, New York, January 5, 2018).

accessible and effective step than charging into machine learning and content analysis. There remain fascinating opportunities for applying those techniques to historical illustrations. And there remain significant challenges to the access, training, and sharing of large visual collections. Developments in all of these areas may reveal as many problems as promising avenues of interpretation, hopefully enlivening the discourse about the history of visual epistemology and the millions of images in our cultural records.

Unless otherwise specified, all work in this journal is licensed under a Creative Commons Attribution 4.0 International License.