Examples of challenges and opportunities in visual analysis in the digital humanities

Holly Rushmeier,[a] Ruggero Pintus,[b] Ying Yang,[a] Christiana Wong,[a] and David Li[a]

[a]Yale University, Department of Computer Science, New Haven, CT 06511, USA
[b]CRS4, Sardegna Ricerche Edificio 1, C.P. 25, 09010 Pula (CA), ITALY

ABSTRACT

The massive digitization of books and manuscripts has converted millions of works that were once only physical into electronic documents. This conversion has made it possible for scholars to study large bodies of work, rather than just individual texts. This has offered new opportunities for scholarship in the humanities. Much previous work on digital collections has relied on optical character recognition and focused on the textual content of books. New work is emerging that is analyzing the visual layout and content of books and manuscripts. We present two different digital humanities projects in progress that present new opportunities for extracting data about the past, with new challenges for designing systems for scholars to interact with this data. The first project we consider is the layout and spectral content of thousands of pages from medieval manuscripts. We present the techniques used to study content variations in sets of similar manuscripts, and to study material variations that may indicate the location of manuscript production. The second project is the analysis of representations in the complete archive of Vogue magazine over 120 years. We present samples of applying computer vision techniques to understanding the changes in representation of women over time.

Keywords: digital humanities, layout analysis, spectral imaging

1. INTRODUCTION

Over the past decade, digitization efforts have made large collections of physical works available as electronic documents. Availability in electronic form makes it possible now to study large collections, rather than focussing on individual texts. This has expanded scholarship in the humanities, allowing scholars to study context that supplements traditional lines of inquiry. Many studies have used optical character recognition and focused on the textual content of books. Machine learning techniques such as topic modeling have been applied to learn about topical trends over time, and to attribute authorship to texts.[1] New work is emerging that is analyzing the visual layout and content of books and manuscripts. The work in the humanities uses techniques that are also being developed in other disciplines. For example, in the medical literature images in publications are being analyzed for similarity to assist researchers in finding related materials in medical literature.[2] In management, the appearance of print advertisements is being examined for their effect on investors.[3]

We present two different digital humanities projects in progress at Yale that present new opportunities for extracting data about the past, with new challenges for designing systems for scholars to interact with this data. The first project we consider is the layout and spectral content of medieval manuscripts. The second is the analysis of representations in the complete archive of Vogue magazine over 120 years.

2. DIGITALLY ENABLED SCHOLARSHIP OF MEDIEVAL MANUSCRIPTS

Digitally Enabled Scholarship of Medieval Manuscripts (DESMM) is a project funded by the Mellon foundation for collaborative work between Stanford and Yale universities in the development of open repositories for medieval manuscripts.[4] The project continues work that Stanford has been engaged in to produce interoperable repositories for medieval manuscripts. The work is a response to initial efforts to digitize medieval manuscripts and make them available in a manner that followed traditional physical library access. That is, the emphasis was on replicating the experience of viewing the manuscript in a physical library. Images of the manuscripts were made

Send correspondence to HER : E-mail: holly@acm.org

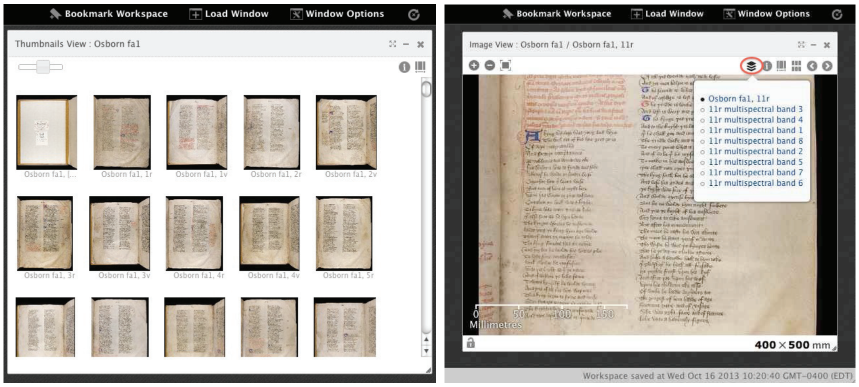

Bookmark Workspace
Load Window
Window Options
Thumbnails View : Osborn fa1
Osborn fa1, [...
Osborn fa1, 1r
Osborn fa1, 1v
Osborn fa1, 2r
Osborn fa1, 2v
Osborn fa1, 3r
Osborn fa1, 3v
Osborn fa1, 4r
Osborn fa1, 4v
Osborn fa1, 5r
Bookmark Workspace
Load Window
Window Options
Image View : Osborn fa1 / Osborn fa1, 11r
● Osborn fa1, 11r
○ 11r multispectral band 3
○ 11r multispectral band 4
○ 11r multispectral band 1
○ 11r multispectral band 8
○ 11r multispectral band 2
○ 11r multispectral band 5
○ 11r multispectral band 7
○ 11r multispectral band 6
0
50
100
150
Millimetres
400 × 500 mm
Workspace saved at Wed Oct 16 2013 10:20:40 GMT-0400 (EDT)

Figure 1. Screenshots show the Mirador canvas view – on the left, thumbnails of multiple pages, on the right, a page with multiple scanned images associated with it.

available in proprietary software viewers that differed for each collection. Because the process of digitization can be costly, viewing the manuscripts was often behind a paywall in an effort to raise funds.

It became evident however, that simply looking at manuscript pages does not exploit the value of having works in digital form. It also became evident that requiring pay for access is not a successful economic model. The goal of the interoperability effort is to create and distribute an open platform for storing and using medieval manuscript. By using an interoperable platform, scholars can study manuscripts from different collections in the same software environment. The basic structure behind the system developed by Stanford is a "canvas" for each manuscript page.[5] A canvas for a manuscript page can have multiple images of the page and additional information attached to it. In addition to making page images from different manuscripts easily available, the development of the system is continuing to facilitate advanced annotation and searches across collections. The viewing system known as Mirador[6,7] is built using the International Image Interoperability Framework.[8] Figure 1 shows screenshots of Mirador for selecting pages, and for selecting a version of a page to view from multiple scans associated with the same page canvas. While annotation and search are well studied problems, there are many ways that they can be implemented in a system. The current focus in DESMM is the development of techniques and formats for a stable system that will support multiple types of scholarly inquiry in the domain of medieval manuscripts.

The specific goal of the DESMM project at Yale is to press the system forward by discovering the needs of specific researchers studying different problems using manuscripts in the interoperability frame work. The Yale projects include creating major new editions of the first and second recensions of Gratian's Decretum (Anders Winroth, English), a study of the creation of English literature, ca. 1385 - ca. 1425 (Barbara Shailor, Classics, Alistair Minnis and Ardis Butterfield, English) and an exploration of the literary history of the Book of Hours (Jessica Brantley, English). These projects were chosen because although they all center around the use of medieval manuscripts, the questions to be answered, and the style of work to be performed in pursuing the answers are quite different. The goal is to find a common structure for data storage, annotation and viewing that serves all of the projects well, rather than tailoring the system to one particular study. The study of Gratian's Decretum, the basis for modern western law, requires search and linking through large numbers of manuscripts. First, the definitive version of the Decretum itself is needed through searching through pieces and copies scattered through various libraries. Then, the basis for Gratian's laws need to be traced through a wide variety of sources of secular law and church canon. This work is extraordinarily demanding on being able to find and record linkages. The creating English Literature project is quite different as its goal is to consider form and material attributes to ultimately understand who physically produced the manuscripts of early English literature such as Chaucer. Traditionally the identification of the scribes that copied works is performed through visual inspection by art historians and paleographers and by dialect studies. The DESMM project particularly tests the requirements for additional data describing physical attributes and and details of image acquisition beyond the image and subsequent annotations made to the image. Finally the Book of Hours project is quite different again in considering the evolution of the concept of a book by examining differences in large numbers of nearly identical copies of the widely used book of prayers. This project tests the use and recording the results of machine learning algorithms on the manuscripts.

Of these projects, the creating English literature and the Book of Hours projects include visual analysis, while the work on Gratian focuses on annotations and tracing the basis for Gratian's work. The projects have been opening questions such as what kind of annotations are useful – considering both manually entered annotations and machine generated annotation. How can and/or should access to annotate and view annotations be controlled? How should the generation of machine annotations be documented? Once annotations are available, how can they efficiently be searched? Additionally, in creating machine annotations, what type of data needs to be available about the initial digitization?

2.1 Creating English Literature, ca. 1385-ca. 1425

The first initiative in visual analysis in DESMM is multispectral scanning of manuscripts to examine material variations. Multispectral imaging has clear benefits for improving the ability to read manuscripts, and to reveal erasures. Multi-spectral imaging can also provide clues about how the manuscripts were made. Changes in pigments and form can provide clues about where the manuscripts were produced. Multispectral images, including data from the ultraviolet and infrared bands, can provide insight into where changes in materials occur. Changes in pigment can be quite subtle, and determining what variations are significant depends on scholarly interpretation. The challenges in this thread are narrowing down many thousands of possible changes to those that may be of interest to scholars, and allowing scholars to view and make notes about these changes.

A pilot study focussed on a work by John Gower from the Beinecke Library at Yale.[9] The problem was to measure the difference between the materials used to produce flourishing and comments in the document, and to compare the spectrum of each to a set of potential pigments. A multispectral imager with sharply defined bands in the UV, visible (just three in the pilot), and infrared was used. The measurements showed clear differences between the pigments used in the two areas of the manuscript. They also suggested potential pigments that were more likely to have been used to produce the work. Physical measurements would be needed to make a final complete determination.

In the larger scale DESMM project, over one thousand pages were imaged with a eight wavelength band (one UV , six visible and one IR) device. An existing image viewing program, Hyper3D,[10] was modified to assist scholars in looking at the changes of spectra at manuscript locations where it has been hypothesized that there is a change in the scribal hand (see Fig. 2). While differences have been found in key areas, a problem that arises is how to determine when two spectra are significantly different. There are natural variations in the observed spectra as the letters and illustrations are completed, as the pen is dipped at different points, and as the materials available to produce the pigment vary. Since most pigments use organic materials, the spectra may change just to variations in the growing conditions of the plants used to extract pigments. To address the problem of pigment variations, we are developing techniques to automatically segment pages by spectra to determine the number of colors used per page, the spectra of each of the apparent colors on each page, and the deviation of the spectral values from the mean for each color. This analysis will allow us to help determine what is a significant variation in the spectrum.

2.2 A Literary History of the English Book of Hours

The second DESMM effort involves analyzing the visual layout in versions of the Book of Hours from late medieval England. Multiple copies of the Book of Hours are by design nearly identical in content. Of interest to scholars though is how the individual copies vary, with the inclusion of secular poetry or music. The Book of Hours was a "best seller" of its time.[11] For many people, it was their only interaction with a book, and so formed their impression of what a book was and what is purpose was. The emergence of variations in this very common book is important in understanding the history of the concept of a book. Understanding innovations in what might be in a book requires finding the variations in the hundreds of copies of the Book of Hours that have survived.

Finding variations without automated analysis is a long and tedious visual inspection process. We have been developing algorithms to segment pages into various types of text blocks and illustrations. We then can present the scholar with samples of particular visual features that are found in a large collection of manuscripts.

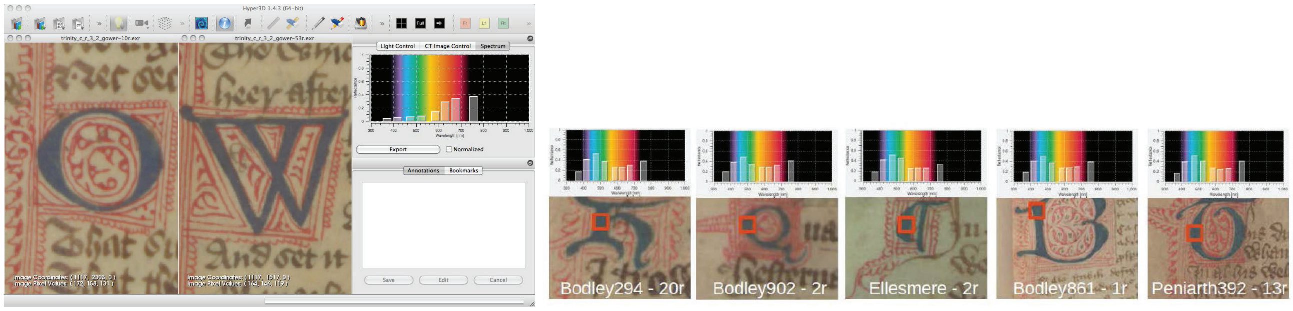


Figure 2. Screenshots show the Hyper3D interface for comparing spectral values on manuscript pages (left), and examples of spectra from several manuscripts (right).

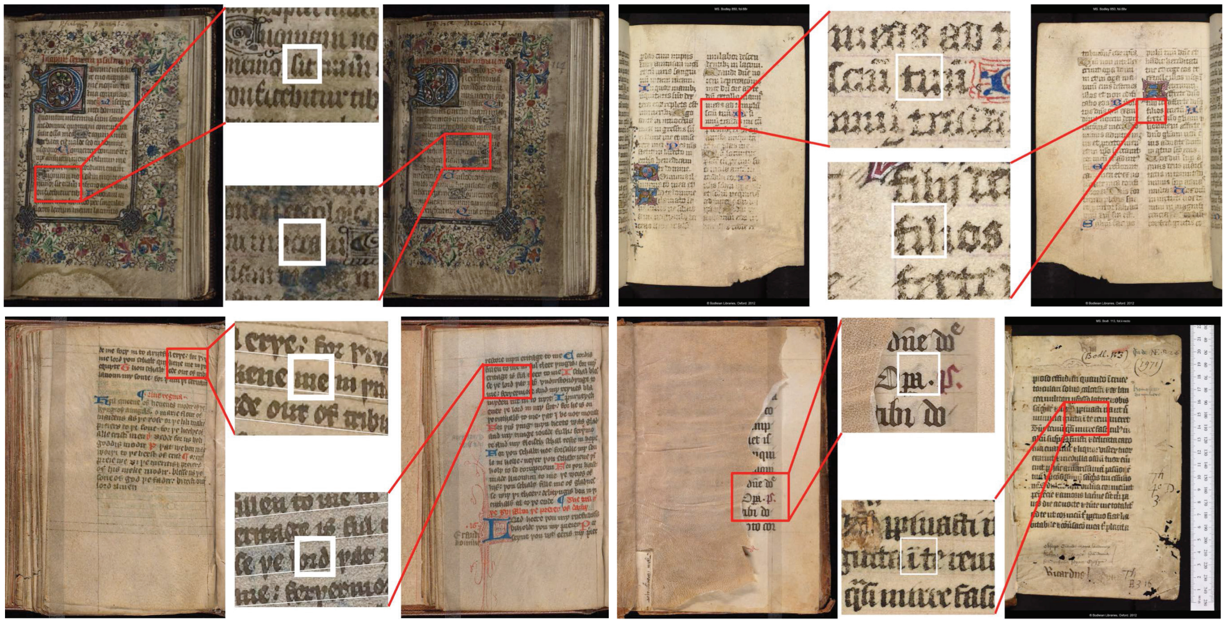


Figure 3. Examples show our approach successfully estimating text height in manuscripts with various types of damage.

Challenges in this research thread are developing methods for humanist scholars to indicate the visual features they are interested in finding, and presenting the results in a meaningful manner. We draw on previous work for segmenting handwritten manuscripts. Within the area of handwritten manuscripts (which include things such 19th and 20th century correspondence), we focus on the problems faced where there is substantial intermingling of text and illustration, and varying states of damage to the the original (see Fig. 3).

In analyzing the page layouts, we use a divide and conquer strategy.[12,13] We developed a technique to estimate the height of text automatically on a per page basis. Because we want to analyze many books with thousands of pages, we sought a completely automatic method for this, rather than a machine learning technique that requires user input for training. We use a multi-scale analysis to identify text and compute a text height per page.[12] Testing the accuracy of this analysis is itself a challenge. Manual analysis of 100 randomly chosen manuscripts found a variation of 15 % in manually determined text line heights per page – setting the standard for how accurate we can expect an automatic analysis to be. Comparison of our automatic analysis with manually determined text height gave a difference of only 14 %. On a further test of 15552 pages, blocks of the automatically found text height were superimposed on each page. These results were then examined manually, and an accuracy of 98% was found over the whole dataset.

We further segment the pages using the text height to carve out text blocks, and areas of capital letters and

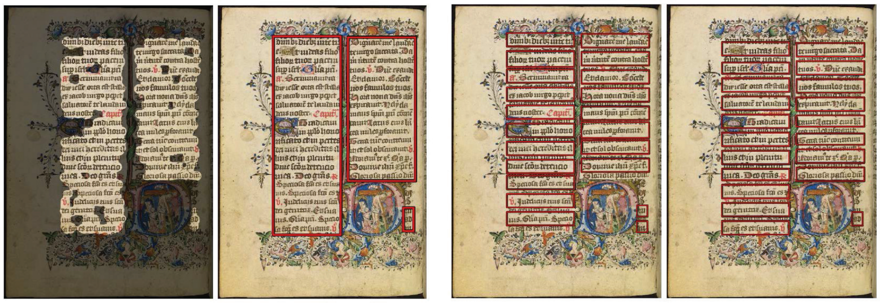


Figure 4. Examples of segmenting pages –far left: a dense segmentation of text, near left: text blocks, right: text lines segmented for transcript (odd lines on left, even on right).

illustrations within the text blocks, Fig. 4. The approach we use is to automatically identify pages with high and low amounts of text using our text height indicator and hue diversity. On these exemplar pages, we identify text and images automatically, using the text high indicator. We then find SIFT features on these labeled pages, and use the labeled pages to run SVM to learn the features associated with text and illustrations in the book. The model learned can then be used to classify all of the locations on all of the pages as text or illustration based on their SIFT features. The technique was tested by manually annotateding 56 randomly selected pages from a collection of 2724 pages. Precision and recall for text blocks was over 90%. Precision and recall for labelling text lines were respectively 69 % and 98%.

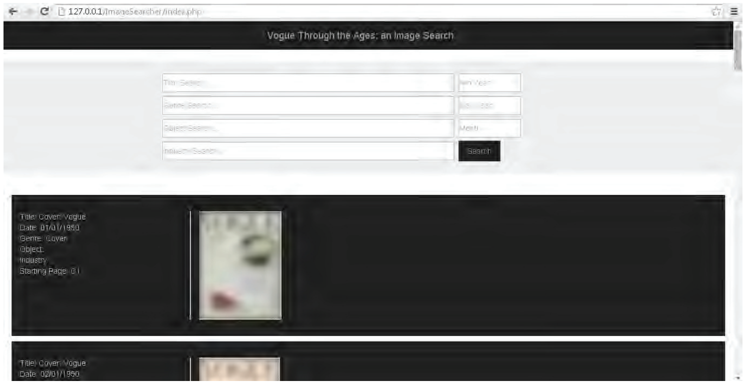
Identifying text blocks reduces the work needed for our current work in word spotting. Our work in progress on word spotting focuses on identifying words with characters, such as the thorn character, that are peculiar to a particular language (e.g. differentiating English from French and Latin), indicating areas where some possibly non-standard text may occur. Within the areas label illustrations, we seek to separate out capital letters and line fillers. Line fillers in particular, may indicate areas of verse, and have a particular aspect ratio that is easy to identify.

In the context of the interoperability project, this work is raising the issue of how and/or whether to include automated results of layout analysis in the manuscript repository. The automated analysis can produce large quantities of data, and is only meaningful if the analysis performed is properly documented. The work also raises the issue of how to make use of manual annotations to build better automated results. The form and vocabulary for manual annotations by scholars can effect the usefulness of the manually entered information in performing new analysis.

3. ROBOTS READING VOGUE

The second project is the analysis of images in the Vogue magazine data archive. The "Robots Reading Vogue"[14] project was initiated by Peter Leonard (Yale Digital Humanities librarian) and Lindsay King (Yale Arts librarian). This archive was provided by Conde Nast to Yale with pages that were segmented and annotated. The archive provides a unique view of attitudes about and representations of women. Textual analysis has already shown and validated clear trends in editorial policy through time and the according to the viewpoints of each editor-in-chief.[15] For example, an n-gram[16] analysis shows a marked decrease in the ratio of the instances of the word "women" to "girls" from the 1920's to the 1980's. N-grams also show the decline of the usage of the word "frock" from a high around 1920 to almost nil in 1950, and the increase of the word "pants" from nearly zero to steady usage from 1970 onwards. Topic modeling showed a change in the themes of articles by year, with terms associated with women's health (e.g. "breast cancer") rather than just fashion becoming prominent in the 1980's during the tenure of Editor-in-Chief Grace Mirabella.

Visual analysis can help form some hypotheses of both the perceptions of women and perhaps how women perceive themselves. In early work we have developed a system to browse the segmented page images and the metadata associated with them.[17] The interface is shown in Figure 5. Entering simple search terms such as "smoking" or "red carpet" is a way to browse imagery from across the decades to form new questions. The



Title: Cover: Vogue
Date: 01/01/1950
Genre: Cover
Object:
Industry:
Starting Page: 01

Title: Cover: Vogue
Date: 02/01/1950

Figure 5. Specialized search program for browsing images from Vogue (copyrighted images blurred for purposes of illustration).

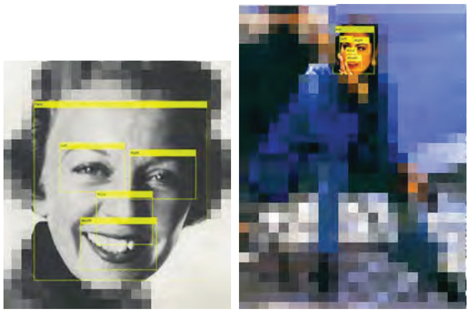


Figure 6. Two images marked with the face finder. The image at the left has high face-ism, the image at the right low face-ism. Copyrighted images pixellated for use in illustration.

search program is also useful for pulling out large quantities of images with annotation data to do specialized analytical studies.

The segmented pages were used in an initial study making use of face finding algorithms to collect statistics about images of women.[18] Three effects were studied – face-ism, facial proportions and photographic composition. As initial studies, no definitive conclusions can be drawn. However these studies do show the potential of visual analysis in studying the archive.

Faces were detected using the Viola and Jones algorithm,[19] as implemented in the MATLAB[20] function vision.CascadeObjectDetector. The algorithm is effective in detecting full frontal faces. Experiments were run with various options for determining the presence of a face. On the Vogue archive, accepting a face if at least 3 facial components (eyes, nose mouth) were detected was found to be effective. Testing on 300 random samples found a 79% success rate in detecting faces, with the errors having few false positives. This success rate was adequate for performing initial proof-of-concept tests.

A total of more than 849,000 images were pulled from the archive for processing. Images were excluded if they were less than 1/8 of the page in size. Using the parameters defined for the face finder, more than 287,000 images were marked as containing faces. For these images the size of the largest face in the image, measures of distances between facial features, and the image with the faced marked (see Fig. 6) was stored, along with the image metadata. .

Face-ism is a concept developed in psychology by Archer[21] designed to measure facial prominence. Face-ism of is defined as the distance from the top of head to chin divided by the distance from the top of head to the lowest portion of the body in the image. Previous work showed higher face-ism for men than for women, and higher regard for periodicals with higher face-ism. To simplify analysis, in our study the ratio of the facial area to the area of the full image was used as a proxy for measuring face-ism. In analyzing all genres of image (e.g.

covers, advertisements, etc.) the ratio increases over time, as shown in Fig. 7. The full data is very noisy, but plotting the median face-ism per issue shows the trend. In analyzing covers, face-ism appears to increase in the 1960's and 70's, and then decrease again in the 1980's Fig 8. This may reflect attitudes in society and/or the influence of the particular editors-in-chief – with Vreeland starting in 1963, Mirabella in 1971, and then Wintour taking over in 1988.

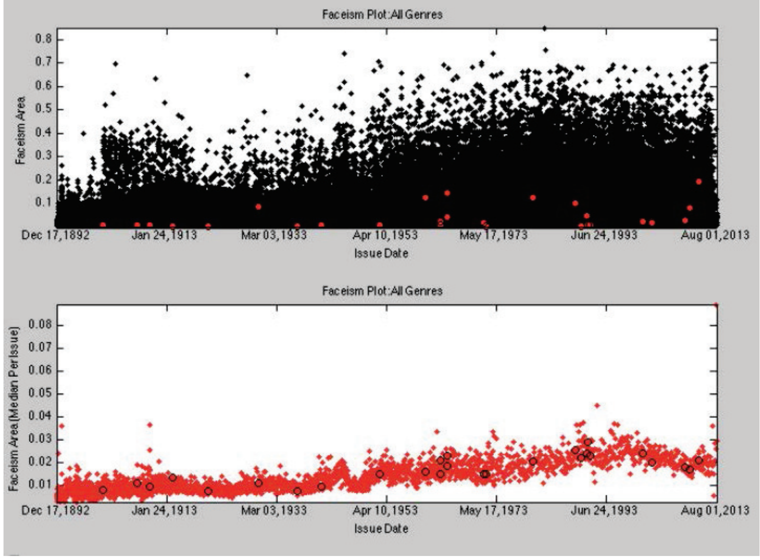


Figure 7. Plots of the trend in face-ism over time for all genres. All data is plotted at the top, only median face-ism is plotted below. Hollow dots are plotted for dates for important changes affecting the magazine, such as change in Editor-in-Chief.

Second we used the ratios between various facial measurements, such as distance between eyes and size of eyes, to look for variations of the "perfect face" as indicated by faces on the magazine cover. The data for facial features was very noisy. Plots and statistical analysis of the various facial measurements relative to the overall face showed no statistically meaningful trend. At least at the scale that the algorithm used to find facial features, no changes over time or by Editor-in-Chief were discovered.

Finally, the position of faces in different classes of images (i.e. covers, fashion shoots, advertisements) were studied to look for trends in photographic composition. In general strong trends for changes over time in composition were not found. Faces were generally found in all genres in the center above the middle. The relative location of faces in all images are plotted in Fig. 9 as semi-transparent dots, so that the density of face positions can be seen.

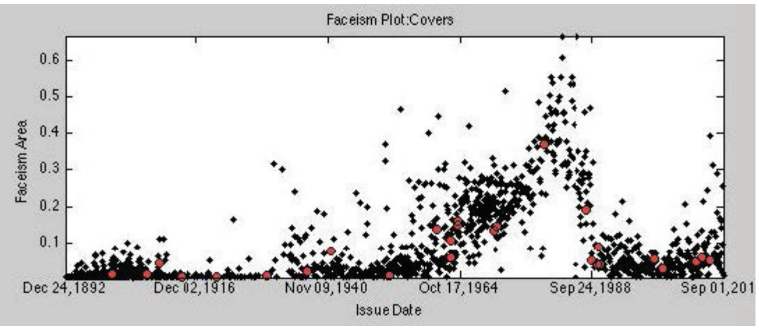


Figure 8. Plots of the trend in face-ism over time only for covers.

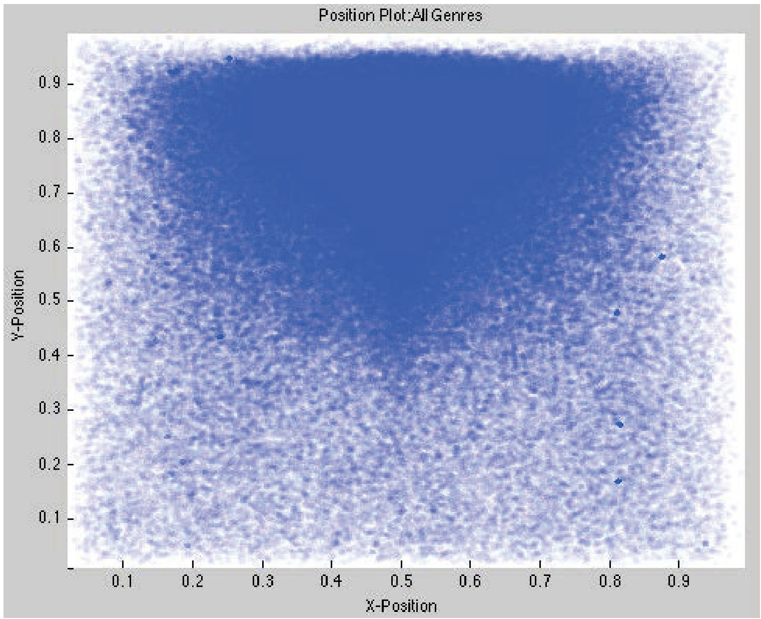


Figure 9. Each semi-opaque dot represents the relative location of a face in an image– faces tend to be in the middle and upper half of the image..

4. CONNECTIONS BETWEEN PROJECTS

The digital representation of the extensive Vogue archive, and the positive and negative aspects of dealing with it, can inform new approaches for representing the data in medieval manuscripts. On the positive side, the format used for the Vogue archive has some advantages for storing various types of data. On the negative side, the data is often incomplete, and setting up a system with the current documents to add and to verify additional information would take a great deal of effort. In that respect, the annotation system being developed in DESMM could inform the development of a system to annotate the Vogue archive.

Semi-automatic methods can be developed to do tasks that are difficult for the computer but are designed to reduce the burden on humans. The success of face finding in Vogue suggests face finding in the illustrations in medieval manuscripts. Additional image object identifiers could be applied to the manuscripts to find various types of illustrations – birds, buildings, flowers, etc. The study of spectral values and color variations in medieval manuscripts suggests a study of the variation of colors used through time in Vogue.

The Digital Humanities has grown rapidly over the last five years at Yale. DESMM and Vogue are just two of many projects. The project participants are joining a growing community that meet to share experiences and new tools.

ACKNOWLEDGMENTS

We acknowledge the assistance of Meg Bellinger, Michael Appleby, Carly Bogen, Barbara Shailor Jessica Brantley, Alistair Minnis and Ardis Butterfield and Lindsay King and Peter Leonard. The DESMM project is funded by the Mellon Foundation.

REFERENCES

[1] Mimno, D., "Computational historiography: Data mining in a century of classics journals," *J. Comput. Cult. Herit.* **5**, 3:1–3:19 (Apr. 2012).

[2] Xu, S., McCusker, J., and Krauthammer, M., "Yale image finder (YIF): a new search engine for retrieving biomedical images," *Bioinformatics* **24**(17), 1968–1970 (2008).

[3] Madsen, J. and Niessner, M., "Is investor attention for sale? the role of advertising in financial markets," *The Role of Advertising in Financial Markets (November 18, 2014)* (2014). http://dx.doi.org/10.2139/ssrn.2506872.

[4] Yale Digital Collections Center, *Digitally Enabled Scholarship with Medieval Manuscripts* (2012 (accessed January 3, 2015)). http://ydc2.yale.edu/research-support/digitally-enabled-scholarship-medieval-manuscripts.

[5] Sanderson, R., Albritton, B., Schwemmer, R., and Van de Sompel, H., "Sharedcanvas: a collaborative model for digital facsimiles," *International Journal on Digital Libraries* **13**(1), 3–16 (2012).

[6] Yale Digital Collections Center, *Canvas Viewer/Mirador* (2013 (accessed January 3, 2015)). http://ydc2.yale.edu/canvas-viewermirador.

[7] Sanderson, R., Albritton, B., Emery, D., Noel, W., and Porter, D., "Distributed repositories of medieval calendars and crowd-sourcing of transcription," *Open Repositories 2014* (2014).

[8] IIIF Community, *International Image Interoperability Framework* (2014 (accessed January 3, 2015)). http://iiif.io/.

[9] Kim, M. H. and Rushmeier, H., "Radiometric characterization of spectral imaging for textual pigment identification," in [*The 12th International Conference on Virtual Reality, Archaeology and Cultural Heritage (VAST11)*], *VAST'11*, 57–64, Eurographics Association, Eurographics Association (2011).

[10] Kim, M. H., Rushmeier, H., ffrench, J., Passeri, I., and Tidmarsch, D., "Hyper3D: 3D graphics software for examining cultural artifacts," *ACM Journal of Computing and Cultural Heritage* **7**, Article. No: 1 (02/2014 2014).

[11] Getty Center, *The Medieval Bestseller: Illuminated Books of Hours* (2002 (accessed January 3, 2015)). http://www.getty.edu/art/exhibitions/books/.

[12] Pintus, R., Yang, Y., and Rushmeier, H., "Athena: Automatic text height extraction for the analysis of text lines in old handwritten manuscripts," (2014).

[13] Pintus, R., Yang, Y., Gobbetti, E., and Rushmeier, H., "A talisman: Automatic text and line segmentation of historical manuscripts," EUROGRAPHICS Workshops on Graphics and Cultural Heritage, EUROGRAPHICS Workshops on Graphics and Cultural Heritage (10/2014 2014).

[14] Leonard, P. and King, L., *Robots Reading Vogue* (2014 (accessed January 3, 2015)). http://dh.library.yale.edu/projects/vogue/.

[15] Leonard, P., "Mining large datasets for the humanities," in [*IFLA WLIC, 16-22 August 2014, Lyon, France*], (2014).

[16] Banerjee, S. and Pedersen, T., "The design, implementation, and use of the ngram statistics package," in [*Computational Linguistics and Intelligent Text Processing*], 370–381, Springer (2003).

[17] Li, D., "Vogue through the ages: A foundation for image analysis," (2014). Yale Computer Science Senior Project Report.

[18] Wong, C., "Vogue: An analysis of vogue fashion photography's implications about the female face," (2014). Yale Program in Applied Mathematics Senior Thesis.

[19] Viola, P. and Jones, M. J., "Robust real-time face detection," *International journal of computer vision* **57**(2), 137–154 (2004).

[20] Mathworks, "MATLAB," (2014). http://www.mathworks.com/products/matlab/.

[21] Archer, D., Iritani, B., Kimes, D. D., and Barrios, M., "Face-ism: Five studies of sex differences in facial prominence.," *Journal of Personality and social Psychology* **45**(4), 725 (1983).