Abstract

In this article we establish a methodological and theoretical framework for the study of large collections of visual materials. Our framework, *distant viewing,* is distinguished from other approaches by making explicit the interpretive nature of extracting semantic metadata from images. In other words, one must 'view' visual materials before studying them. We illustrate the need for the interpretive process of viewing by simultaneously drawing on theories of visual semiotics, photography, and computer vision. Two illustrative applications of the *distant viewing* framework to our own research are draw upon to explicate the potential and breadth of the approach. A study of television series shows how facial detection is used to compare the role of actors within the narrative arcs across two competing series. An analysis of the FSA-OWI corpus of documentary photography is used to establish how photographic style compared and differed amongst those photographers involved with the collection. We then aim to show how our framework engages with current methodological and theoretical conversations occurring within the digital humanities.

1 Introduction

Digital humanities' (DH) focus on text and related methodologies such as distant reading and macroanalysis has produced exciting interventions (Jockers, 2013; Underwood, 2017). Yet, what about that which we see and hear? Cultural forms predicated on visuality and sound have long shaped our daily experiences. Disciplines such as art history, film studies, media studies and music continue to show how visual and aural objects reflect and shape cultural values. These disciplines have been joined by sound studies and visual culture studies, which have also ardently argued for the centrality of audio (Attali, 1977; Schafer, 1993; Sterne, 2003) and visual (Jay, 1993; Martin, 1994; Mitchell, 1994; Mirzoeff, 1998) forms to our mediated lives. Building on decades of scholarship from the across these fields, the call to take seriously sound culture, visual culture and moving images as objects of study in digital humanities is amplifying (Clement, 2012; Posner, 2013; Acland and Hoyt, 2016; Manovich, 2016).

As a part of this chorus, we argue that DH should consider, what we call, *distant viewing* – a methodological and theoretical framework for studying large collections of visual material. *Distant viewing* is distinguished from other approaches by making explicit the interpretive nature of extracting semantic metadata from images. In other words, one must 'view' visual materials before studying them. Viewing, which we define as an interpretive action taken by either a person or a model, is necessitated by the way in which information is transferred in visual materials. Therefore, in order to view images computationally, a representation of elements contained within the visual material—a code system in semiotics or, similarly, a metadata schema in informatics—must be constructed. Algorithms capable of auto-

matically converting raw images into the established representation are then needed to apply the approach at scale. Therefore, the active process of analyzing large collections of visual materials computationally is the act of *distant viewing*.

In the sections that follow, we establish the motivation and goals of *distant viewing* from an interdisciplinary perspective. The method draws on scholarship from semiotics and visual cultural studies that theorize the cultural function of images and how they produce meaning differently than other cultural forms. Given the dominance of textual analysis in the humanities, we in particular focus on the contrasting elements of linguistics and visual knowledge production. We then illustrate how these differences are paralleled in the computational sciences by exploring the relationship between natural language processing and computer vision. We conclude by positioning *distant viewing* within other theoretical frameworks developed in DH.

2 Distant Viewing Framework

2.1 Meaning making in visual materials

Work in visual semiotics has established that the way meaning is encoded in images differs from text. Textual data is described by characters, words, and syntax. Read within a particular cultural setting, these elements are interpreted as having meaning. Linguistic elements serve as a code system where signs, such as words, correspond to objects primarily by convention (Saussure, 1916; Dubois *et al.,* 1970*;* Pierce, 2000). The word 'camera' in English is typically used to refer to a type of equipment capable of capturing photographs or moving images from observations of light. The link between the six-letter word and the definition is induced by millions of English speakers having previously agreed upon the defining relationship.[i] The same relationship exists in spoken language between the pronunciation of the

word (IPA: /kæmrə/) and the same underlying concept of cheese as a certain class of milk products. In French, the word 'fromage' serves as a code for the same concept. Grammatical constructs such as verb conjugation, plurality, and object-verb relationships operate similarly within a particular language to produce higher level meanings between individual words.

Images function differently. Visual culture studies, informed by semiotics and rhetorical studies, explores how images signify and communicate (Barthes 1977b; Hill and Helmers, 2004; Kress and van Leeuwen, 2006). Visual forms such as paintings and photographs illustrate and circulate concepts through characteristics such as lines, color, shape, and size.[ii] An image serves as a link to the object being represented by sharing similar qualities. One may recognize a painting of a particular person by noticing that the painted object and person in question shares properties such as hair style, eye color, nose shape, and typical clothing. The representational strategies of images, therefore, differ from language. While often rendered meaningful in different ways through language, visual material is 'pre-linguistic, a "truth" of vision before it has achieved formulation' (Scott, 1999, p. 20).

A photograph, for example, in its most basic form is a measurement of light through the use of either chemically sensitive materials or a digital sensor.[iii] One does not need to know a particular language in order to distinguish the objects represented within a photograph (Scott, 1999). As Roland Barthes argues, there is a 'special status of the photographic image: it is a message without a code' (1977a). It is not necessary with photography to construct an explicit mapping between the visual representation and what is being represented.

The relationship between the photograph and the object being represented by the photograph is signified by shared features between the object and its photo. This not to suggest that photographic images are somehow culturally agnostic.

The culturally coded elements of photographic images coexist with the raw measurements of light. The cultural elements are exposed through the productive act of photography—what Barthes refers to as the image's 'connotation' (Barthes, 1980). Why was a particular shot taken by the photographer? Why was this image developed in a particular style? What effects were applied to the digitized image? How was the image cropped? Where and how was the image placed within a newspaper page or displayed in a frame? The answers to all of these questions critically depend on the cultural influences and motivations present in every stage of the creative process. The cultural elements, then, serve as a second layer of meaning constructed by the raw elements of a photograph.[iv] The presence of both coded and uncoded messages in photographic images, which Barthes considers the 'photographic paradox', points to the added complexity of working with visual materials compared to textual data (Barthes, 1977a). While a person looking at a photograph can encode and decode the objects and therefore meaning of a photography, the process of making explicit these decisions is exactly what makes photographs so difficult to study computationally. There is an additional layer of interpretation demanded by visual material that has to be made explicit. Such conditions are not only characteristic of the way these different cultural forms signify but extend to computational approaches.

A quick comparison to text analysis helps elucidate the difference in code systems between images and text and how computers process them differently. The explicit code system of written language provides a powerful tool for the computational analysis of textual

corpora. Methods such as topic modeling, TF-IDF, and sentiment analysis function directly by counting words, the smallest linguistic unit that can be meaningfully understood in isolation (Saussure, 1916). The interpretive act of understanding these units may be delayed until the models are applied.[v] It is, for example, only after applying latent Dirichlet allocation (LDA) and finding a topic defined by the words 'court', 'verdict', and 'judge' that we are forced to decode the meanings of the words. Images afford no such coded units on which to aggregate. Raw pixel intensities hold no meaningful information out of context. Even if we split the pixels of an image into objects, the pixels that represent a particular object in one photograph will be completely different than those pixels representing the exact same object from a different perspective or moment in time. To the computer, there is only one "Eiffel Tower," but there are millions of unique photographs of the Eiffel Tower described by different sets of pixels. The difference between text and images is exacerbated when considering non-proper nouns. The word 'dog' represents a concept describing a particular species of animals. A photograph of a dog directly represents only one particular instance of a particular animal. Whether the photograph should be considered as a representation of a particular dog, a certain dog breed, the class of all dogs, the class of all pets, or all mammals depends on the larger context of the image. When analyzing images at scale, one needs to explicitly decide what is 'actually' being represented by each object prior to aggregating and analyzing the collection.

In order to conduct a computational analysis of a large digitized corpus of photographic materials, it is necessary to develop a code system for describing each image. A description of the elements constituting the image must be added as a new layer of meaning to the raw pixels. For example, the image on the left-hand side of Fig. 1 might be described as: 'an older man sits on a horse overlooking a field'. The process of coding images in this way is

both destructive and interpretive. Many elements of the image are lost in this description, and no amount of words could ever fully capture the entirety of the original photograph. At the same time, the description of this space as a field and assumption that this is an older man are both interpretations based on a particular cultural reading of the image that may not match that of the photographer or the intended audience. The difference between elements contained in the raw image and the extracted structured information used to digitally represent the image within a database is known as a *semantic gap*. This gap is a known challenge within the field of information retrieval because, unlike linguistic data, the "meaning of an image is rarely self-evident" (Smeulders *et al.*, 2000, p. 1375). The challenges of interpretation of visual material is further expounded when studying large corpora by applying the process of converting images into a coded system with an automated logic.[vi]

2.2 Computing with visual materials

The contrasting way in which images and text are digitally stored and interpreted mirror the semiological differences that exist between mediums. Text is written as a one-dimensional stream of characters. These characters are written in an encoding scheme that defines an agreed upon mapping from 0s and 1s into symbols. Compression software can be used to store the encoded text using a smaller amount of disk space, but this compression must be reversible. That is, it is possible to go from the compressed format back into the raw text without losing any information. Images are stored in a different format. While displayed on the screen as an array of pixels, images can be stored in a number of compressed formats. Many common compression formats, such as JPEG, use a lossy compression algorithm. These algorithms only approximately reconstruct the original image. Also, it is possible in any storage format to rescale an image to a lower resolution.[vii] This process saves hard drive space

but results in a grainier version of the original. The fact that digital images can be scaled to a smaller size highlights the lack of a formal code system within photographic images. If an image consisted of a code system, lossy compression would require losing some of the coded elements. However, for moderate amount of compression, all of the semantic information within an image can be retained following compression.

The framework of *distant viewing* calls for the automatic extraction of semantic elements of visual materials followed by the aggregation and visualization of these elements via techniques from exploratory data analysis. The extraction and representation of semantic elements constitutes the construction of an explicit code system. There are many types of elements that can be represented by such a code. Simple examples include the dominant color pallets, lighting, and moving image shot breaks. More complex examples include object detection, facial recognition, and the detection of camera movement. The process of extracting metadata from visual materials involves assigning a semantic meaning to the array of pixels, which serves as an explicit code system.

Algorithms allow for the creation of an automatable and trainable code system to view visual materials. Such computational techniques for understanding the objects present within an image dominate the current research agenda of computer vision. The first step in understanding objects in images is the task of object detection, identifying what objects are present in an image. Algorithms for image detection are typically built by showing a computer example images from a set of predefined classes (typically, containing several thousand object types). The computer detects patterns within the images and uses these to find similar objects in new images. In addition to knowing what objects are present, researchers also want to identify where in the image an object is located. This step, object localization, is

now often done simultaneously with the object detection (Redmon *et al.*, 2016). Finally, once an object is detected and localized, computer algorithms can be taught to conduct object qualification. Algorithms for face detection, for example, may include identifying the identity of the person, their facial expression, and the direction in which they appear to be looking (King 2009; Baltrušaitis, 2016).

Computer vision research has been deeply concerned with creating explicit code systems for decades, and there have been significant gains in just the past two to three years. Major progress has been made in approaching human-like accuracy on several annotation tasks (Szegedy *et al.*, 2017). At the core of these improvements is the successful use of deep learning models—a class of general-purpose algorithms that find latent structures within large datasets—for object detection and localization (He, 2016). The application of deep learning has been greatly assisted by improvements in hardware specifically designed for the required calculations.[viii] While not necessary for the application of models to new datasets, hardware accelerated algorithms have been behind nearly all recent advances in the field of computer vision. Finally, software architectures such as TensorFlow and Caffe have made it possible for researchers to quickly prototype new models and share them with the community. Of course, there is still significant work to be done in realizing much larger goals of full artificial intelligence. Algorithms are, currently, only able to achieve human-like results on relatively constrained tasks (Goodfellow and Bengio, 2016). At the moment they struggle to quickly generalize results to entirely new tasks. However, the current state of research in computer vision is sufficiently developed to be useful for extracting well-defined semantic elements from photographic materials.

In Fig. 1, we see several examples of the types of codes that can be extracted as metadata about an image. The first frame shows rectangles, known as bounding boxes in computer vision, describing the location and object type of a person, a dog, and a horse. The second example illustrates the detailed semantic information that can be extracted from facial features by identifying the location of eyes, noses, and mouths of three women. Additional algorithms can build off of these facial features to summarizing emotions and indicating the identity of people found in one image. Similarly, either image could be described by an auto-generated linguistic summary. These various features describe particular elements of each image by a specific system of codes, which can take on the form of either structured data (coordinates and labels, as in Fig. 1) or linguistic data. These extracted elements do not attempt to capture all of the elements of an image; as mentioned, the interpretive act of coding images is necessarily destructive. The metadata here, also, does not directly attempt to measure higher-order meanings such as the themes, mood, or power dynamics captured by an image. However, much like the relationship between words and cultural elements in text, these elements can often be discerned by studying patterns in the extracted features.

The automated process of making explicit the culturally coded elements of images is what we call distant viewing. Unlike text, we must assign a semantic meaning to a set of pixels. Elements may be as simple as a color to as complex as an emotion. We now turn to two examples to illustrate how this general framework can be applied to explicit humanities research questions.

3 Examples

Like other methods in DH, *distant viewing* calls for the exploratory analysis of extracted and aggregated metadata in order to view larger patterns within a corpus that may be difficult to

discern by closely studying only a small set of objects. The term distant is used to signal that such an analysis is designed to be conducted at a large scale. One risk of using the word distant is that such a term may suggest claims to objectivity. The method actually argues for the opposite. Rather, the code system that is required for distant viewing is culturally and socially constructed. This approach also is not at the expense of close viewing. In combination with subject matter expertise and subsequent close analyses, the approach of studying high-level trends allows for scholars to ask and address new and existing questions. To illustrate the *distant viewing* method, we turn to two examples.

3.1 Detecting narrative arcs in American television

Working with TV studies scholars Annie Berke and Claudia Calhoun, we have applied our *distant viewing* framework to a comparative study of narrative style of American situational comedies. Many media scholars have studied the formal properties of moving images, such as framing and blocking, as they pertain to feature-length films. Television, particularly during the Network Era (1954-1975), has in contrast often been portrayed as formulaic, middlebrow, and lacking in stylistic form. Scholars who have studied network-era television have most often been interested in the way television has produced or challenged race and gender hierarchies (Lipsitz, 1990; Spiegel, 1994; Douglas, 1995; Acham, 2005; Desjardins, 2015). Often missing from cultural studies approaches are accounts of form and style. We wanted to augment these studies with a computational analysis of the television shows themselves. Consisting of hundreds of hours of material, this task was a perfect candidate for the application of *distant viewing*.

Our initial approach was to study the situational comedies *Bewitched* (1964-1972) and *I Dream of Jeannie* (1965-1970). The underlying research question is to identify how

gender is being performed and represented through formal elements present in the show. These series were chosen as an initial study because they ran during the same time period, with very similar story lines, on competing networks. Both series also had very high ratings and continue to hold significant cultural capital.[ix] To analyze the stylistic features of these sitcoms, we identified the location of main characters in each frame and found the time codes of shot and scene breaks. These features serve as the code system in our analysis. Feature were identified by modifying the face detection and identification from the deep-learning model OpenFace (Baltrušaitis, 2016) and developing our own shot detection algorithm (Arnold and Tilton, 2017). Applying these over just one season of episodes from one of the shows yielded over one million detected faces and nearly five thousand shots. In order to work with such a large collection of data, we have written optimized software for extracting features from moving images. This software, the Distant Viewing Toolkit (DVT), has been made available under an open source license and is currently in active development thanks to funding from the United States National Endowment for the Humanities.

By looking at the placement and patterns of the main characters over the 1966-1967 season, we detected two noteworthy patterns. As shown in Fig. 2, Samantha, the main female character on *Bewitched*, is rarely absent from the show for more than a single scene. Although still constrained to a domestic existence centered on a 1960s nuclear household, the plot of each episode has moved into the domestic household as well. Samantha appears to carry and move the entire narrative arc. In contrast, the character Jeannie is often absent from significant portions of an episode. In some episodes she appears to function as a plot device rather than the main character. Jeannie appears at the start of the episode to cause trouble, disappears during the middle acts, and then reappears for the concluding scene. The

differences between the way in which two leading female actors function within the narrative arcs, and the role of the domestic space within each narrative, contrasts with research that often lumps *Bewitched* and *I Dream of Jeannie* together as serving similar cultural purposes (Jagtiani, 1995; Meyer 1998). But, as our *distant viewing* of just one season of television illustrates, the formal qualities of the two shows are quite divergent. The differences are important indicators of how the series were produced and interpreted, and these differences should be taken into consideration when looking at the impact of particular television series on U.S. culture and society.

3.2 Stylistic features in documentary photography

The *distant viewing* framework does not necessarily require the explicit identification of objects within images. It is possible to 'view' an image, that is, to represent the image in a code system, in other ways. A convolutional neural network (CNN) is a particular type of deep learning model particularly well-suited for image analysis. It operates by applying a sequence of compression algorithms to the original image. Unlike other compression algorithms, however, the goal of those in a neural network is not to represent the original input with a smaller approximation. Instead, the compression tasks attempt to extract increasingly complex semantic features from the image. For example, the first few algorithms may detect gradients followed by edges, textures, shapes, objects, objects in context, and entities. When applied to object classification tasks, the final compression algorithm produces a set of probabilities over the available object types. The algorithms in a neural network, known as layers, are 'trained' by a mathematical optimization task attempting to figure out which algorithms produce the most accurate probabilities as a final output. While the final layer of the model

is applicable only to object classification over a predefined set, the output of the other compression algorithms are known to be generalizable to other image processing tasks. This has a number of applications. For example, new classification tasks can make use of transfer learning—where only the last layer of a neural network is re-trained on a new dataset—to build classification algorithms using much smaller datasets and limited computing resources.

Because the raw pixels do not encode meaningful information, the exact color intensities typically differ substantially when comparing the raw pixel values in two very similar images. One option for comparing two images is to run the images through the compression algorithms in a neural network and then compare how similar the compressed versions of the image are. As the neural network is attempting to represent higher-level features in the image, two similar photos will have very similar representations. It is possible to use this behavior to perform image clustering using a neural network. A large set of images is first compressed using the inner layers of a neural network. Then, images are clustered together if their representations are similar. These clusters can be used for tasks such as a recommendation system for photographic collection (Arnold *et al.*, 2016) or to identify thematic or stylistic patterns.

We have applied image clustering to the one hundred and seventy thousand images from the FSA-OWI photographic collection. The FSA-OWI collection consists of approximately 170,000 photographs taken between 1935 and 1943 by the U.S. federal government. Photographs in the collection depict daily life across the United States during the Great Depression and the Second World War. Images include iconic photographs such as Dorothea Lange's *Migrant Mother* (1936) and Arthur Rothstein's *Young Girl at Gee's Bend*. Due to the

large size of the corpus and prominence of many of the photographs, much of the research on the visual and artistic elements of the FSA-OWI collection as a work of art has been focused on specific photographers. Our use of a distant viewing approach provides a way of analyzing and exploring stylistic features across dimensions such as individual photographers, place, and space.

Using the compression algorithms from the internal layers of the InceptionV3 model, image clustering reveals several important links across the FSA-OWI corpus (Szegedy *et al.*, 2015). Fig. 3 shows five example images from the collection and their seven closest neighbors. Each set of similar images centers around detection of the same dominant object: horses, wooden houses, pianos, train cars, and cooking pots. The first two photographs of pots show the exact same scene taken from slightly different angles. Most connections, however, show different instances of the same class of object, possibly from different perspectives. Often these photos are taken many years apart, by different photographers, on opposite sides of the United States. The state-of-the-art InceptionV3 algorithm is still a long way from replicating the entire human visual system. However, the deep learning model is able to detect the essence of what defines an object such as a horse, piano, or train car. The only 'mistake' is the second to last photograph in the line of photos depicting pianos. While resembling someone sitting at a standing piano even to the human eye, this image in fact depicts a worker pulling a large sheet of paper from a machine in a paper mill.

The distances induced by the deep learning model help to understand similarities between the photos taken by the staff photographers involved in the FSA-OWI project. Fig. 4 depicts a network of the twenty staff-photographers with the largest set of credited photo-

graphs. Each photographer, represented by a node in the graph, is connected to the photographer whose photos most resembled their own using the distant metric induced by the penultimate layers of the InceptionV3 neural network model. The network illustrates several meaningful relationships. Ben Shahn is connected to his close mentor Walker Evans. Four photographers associated with the OWI portion of the corpus—Siegel, Palmer, Hollem, and Feininger—are connected only to one another for they all focus on homefront war mobilization. This suggests that scholarship on the visual style of the corpus may want to distinguish the work of these four photographers from the rest of the collection. The remaining cluster consisting of Liberman, Parks, Collins, Rosner, and Bubley selects those photographers most closely associated with photographing dense urban spaces—namely, New York City, Washington D.C., and Chicago. The network also echoes recent scholarship on the FSA-OWI collection by suggesting the central role that women photographers had in shaping the collection (Appel, 2015; Brennan, 2015): the two most central photographers in the network are Marjory Collins and Marion Post Wolcott.

4 What's in a Name?

Our use of the term *distant viewing* draws off of the concept of distant reading for literary history.[x] Though the term 'distant reading' was originally coined by Franco Moretti (2000), within literary theory it refers to a larger methodological movement 'originally framed by scholars…who worked on the boundary between literary history and social science' (Underwood, 2017).[xi] How do we read ten thousand books? How might computational methods make analysis at scale possible? What might we learn? These questions and the provocations they assert offer important additions to traditional methods of close reading. Yet, the very term reaffirms the privileging of text in the digital humanities and in the humanities writ-

large. Other forms of culture such as visual, aural, and embodied are secondary. The use of reading in distant reading is not an expanded notion such as framing culture as a 'text' that is 'read' as argued by anthropologist Clifford Geertz and now pervasive throughout fields like cultural studies (Geertz, 1973). Distant reading is about text as word culture, which the very use of the word 'reading' discursively signals.

The emphasis on distant reading as words, for example, led Tanya Clement to configure the computational analysis of sound as 'distant listening'. If hearing is the passive perception of sound, listening is the active decision to perceive sound. Such a computational approach required developing a new tool set. Clement and her team developed High Performance Sound Technologies for Access and Scholarship, known as HiPSTAS, to analyze spoken word audio. *Distant viewing* follows Clement's implicit critique of word culture by signaling through the very name of the method that the object of study is visual culture forms.[xii]

One might then ask why not use terms like culturomics, macroanalysis, or cultural analytics. The first two bring us back to a focus on text despite their more expansive names. In 2011, two Harvard scholars announced in *Science* what they called culturomics, a transdisciplinary area of study across the humanities and social sciences that quantitatively analyzed culture (Michel *et al.*, 2011). The paper and the emerging field the authors announced alongside their findings garnered praise (and quickly criticism) for their problematic use of the Google n-gram viewer to draw conclusions about culture. Despite the term's potentially broader configuration, scholars of culturomics formulated cultural analysis at scale as limited to text.

Also advocating for the use of computational methods to analyze humanities data at large scale, Jocker's labeled his method macroanalysis. As with distant reading, he is concerned with understanding the structures of literature by using quantitative computational analysis. He suggests, however, that macroanalysis emphasizes 'the systematic examination of data … [using a] quantifiable methodology' whereas distant reading suggests that the computer is engaged in the 'interpretive act of "reading"' (Jockers, 2013, p. 25). His term more accurately reflects the work of this method for 'this is no longer reading that we are talking about,' he provocatively argues (Jockers, 2013, p. 25). Like distant reading and culturomics, this method is about text.

Cultural analytics on the other hand is more expansive than *distant viewing*. Formulated in 2005 by Lev Manovich, the term was used to describe 'the analysis of massive cultural datasets and flows using computational and visualization techniques' (Manovich, 2016). Manovich anticipated the need to and persuasively argued for the study of visual culture at large scale. A particular focus of Manovich and his lab has been social media visual culture, but, as he writes, 'cultural analytics is interested in everything created by everybody' (Manovich, 2016). Such a broad framework is shared by the recently launched *Journal of Culture Analytics* that defined the term as the 'computational study of culture'. The journal's editor Andrew Piper writes, 'what unites it [cultural analytics] is a belief that computation can show us things about culture that previous media and their metonymic impasses could not' (Piper 2016). We share this belief that cultural forms and see Piper and Manovich as key interlocutors. Because of this, we understand *distant viewing* as a method within cultural analytics.

While what we name methods can be discursively powerful, it may not alone be the reason a new term should be developed. Rather, we argue, *distant viewing* is not simply a name change—distant reading or macroanalysis of images. Rather, the very way that different cultural forms make meaning and the logic of what the computer processes when it reads text and views images are of different kinds. As a result, these different conditions shape the series of assumptions and processes that we use to study culture at scale. Because it is important to signal and reveal the methodological and theoretical assumptions of knowledge production, we argue that when we are applying the kind of computational study of visual culture we have outlined in this paper, we are *distant viewing*.

5 Conclusions

In this article we have focused on the need for coding visual materials prior to exploratory data analysis of visual corpora. *Distant viewing* does not, however, specify the particular characteristics that this interpretation should take. Rather, this framework, which calls for establishing code systems for specific computational interpretations through metadata extraction algorithms, should be a major focus of study within DH.

Scholars from many disciplines have established formal elements for use in the close analysis of non-textual data. Visual culture studies uses formal analysis such as composition, focus, and repetition alongside cultural studies to analyze how cultural forms like photographs make and circulate meaning. Visual rhetoric establishes elements including composition, tropic features, and nature, regarding the production and publication of visual materials. In moving images, film theorists draw on concepts such as blocking, framing, and mise–en–scène in order to discuss and catalogue the thematic and stylistic features. Performance studies, while not always pegged to visual representations, has established ontologies for

describing the interaction between people and object within a defined space. In developing interpretive algorithms, we should find ways to extend such features identified for close analysis to large-scale analyses. Doing this requires both formalizing the close analysis features, where necessary, and building algorithms that can automatically detect and describe them within a fixed metadata schema while recognizing that we are actively and explicitly creating a code system.

Interpretive strategies within the *distant viewing* framework are only useful if they can be applied efficiently to large corpora. To ensure that we are accomplishing this task while making the most of advances in computer vision requires ongoing partnerships, such as ours, between humanists and computational scholars. As we have shown in this article, there are many connections that go far beyond simple tool-building. Uniting theories such as visual semiotics and deep learning, we have the potential to pursue work in both fields and open up new areas of scholarship. By producing an explicit framework that encourages such partnerships, we believe *distant viewing* will serve as a catalyst for achieving these goals.

References

Acham, C. (2005). The Richard Pryor Show. In Alvarado, M., Buonanno, M., Gray,H., and Miller, T. (eds) *The SAGE Handbook of Television Studies*. London: SAGE Publishing.

Acland, C. and Hoyt, E., (2016). *The Arclight Guidebook to Media History and the Digital Humanities*. Sussex: REFRAME Books.

Appel, M. J., (2015). The Duplicate File: New Insights into the FSA. *Archives of American Art Journal*. 54(1): 4-27.

Attali, J., (1977). *Bruits: Essai sur l'Economie Politique de la Musique*. Paris: Presses Universitaires de France.

Arnold, T., Leonard, P. and Tilton, L., (2016). Knowledge Creation Through Recommender Systems. *Digital Scholarship in the Humanities*, 32(3): ii151-ii157.

Arnold, T. and Tilton, L. (2017). Distant Viewing TV: An Introduction. https://statsmaths.github.io/blog/dtv-introduction/ (Accessed 15 November 2017).

Baltrušaitis, T., Robinson, P. and Morency, L.P., (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision, 2016 IEEE Winter Conference*, 2016(1): 1-10.

Barthes, R. (1977a). The Photographic Message, In: Wells, L. (eds), *Photography: A Critical Introduction.* London: Routledge.

Barthes, R., (1977b). Rhetoric of the Image. In Heath, S. (ed., trans.) Image - Music - Text. New York: Hill and Wang.

Barthes, R. (1980). *Camera Lucida: Reflections on Photography*. New York: Hill and Wang.

Bender, K., (2015). Distant Viewing in Art History. A Case Study in Artistic Productivity. *Digital Art History,* 1(1): 101-110.

Brannan, B., (2015). Perpetual Pioneers: The Library of Congress Meets Women Photojournalists of World War II. In Kadar, M. (ed) *Working Memory: Women and Work in World War II*. Waterloo, ON: Wilfrid Laurier University Press.

Ciula, A. and Øyvind, E. (2017). Modelling in Digital Humanities: Signs in context. *Digital Scholarship in the Humanities*. 32(1): i33-i46.

Clement, T. (2012). Multiliteracies in the Undergraduate Digital Humanities Curriculum. In Hirsch, B. (eds) *Digital Humanities Pedagogy: Practices, Principles and Politics*. London: Open Book Publishers.

Desjardins, M. (2015). *Father Knows Best.* Detroit, MI: Wayne State University Press.

Douglas, K. (1995). *Media Culture: Cultural Studies, Identity and Politics Between the Modern and Postmodern.* London: Routledge.

Dubois, J., Edeline, F., Klinkenberg, J.-M., Minguet, P., Pire, F., Trinon, H., (1970). Rhétorique Générale. Paris: Larousse.

Geertz, C., (1973). *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Goodfellow, I. and Bengio, Y. (2016). *Deep Learning*. Cambridge, MA: The MIT Press.

He, K., Zhang, X., Ren, S. and Sun, J., (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017(1): 770-778.

King, D.E., (2009). dlib-ML: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul), 1755-1758.

Kress, G. and van Leeuwen, T., (2006). *Reading Images: The Grammar of Visual Design*. London: Routledge.

Hill, C. and Helmers, M., (2004). Defining Visual Rhetorics. London: Routledge.

Jagtiani, A., (1995). Television Portrayal of Women in the 1960s Case Study Situation Comedies of the Sixties The Dick Van Dyke Show, Bewitched, I Dream of Jeannie. Ph.D. thesis, The Ohio State University.

Jannidis, F. and Flanders, J. (2013). A Concept of Data Modeling for the Humanities. Lincoln, NB: Digital Humanities, pp. 237–39.

Jay, M., (1993). *The Denigration of Vision in Twentieth-Century French Thought.* Los Angeles: University of California Press.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literacy*. Urbana-Champaign, IL: University of Illinois Press.

John, M., Kurzhals, K., Koch, S., and Weiskopf, D. (2017). A Visual Analytics Approach for Semantic Multi-Video Annotation. In *Proceedings of the 2nd Workshop on Visualization for the Digital Humanities*. Online.

Lipsitz, G. (1990). *Time Passages: Collective Memory and American Popular Culture*. Minneapolis, MN: University of Minnesota Press.

Manovich, L. (2016). The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. *Cultural Analytics.* http://culturalanalytics.org/2016/05/the-science-of-culture-social-computing-digital-humanities-and-cultural-analytics/. (Accessed 10 December 2017).

Meyer, M., (1998). I Dream of Jeannie: Transsexual Striptease as Scientific Display. *TDR*, 35(1): 25-42.

Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J. and Pinker, S., (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014): 176-182.

Mitchell, W. J. T. ,(1994). *Picture Theory*. Chicago: University of Chicago Press.

Mirzoeff, N. (1998). *The Visual Culture Reader*. London: Routledge.

Moretti, F., (2000). Conjectures on World Literature. *New Left Review*. Jan-Feb: 54-68.

Peirce, C. (2000). Writings of Charles S. Peirce: A Chronological Edition. Indianapolis, IN: Indiana University Press.

Piper, A. (2016). There Will Be Numbers. *Cultural Analytics.* http://culturalanalytics.org/2016/05/there-will-be-numbers/. (Accessed 10 December 2017).

Posner, M. (2013). Digital Humanities and Film and Media Studies: Staging an Encounter. *Workshop della Society for Cinema and Media Studies Annual Conference*, Chicago**.**

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., (2016). You Only Look Once: Unified, Real-time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017(1): 779-788.

Panofsky, E., (2013). *Studies in Iconology: Humanistic Themes in the Study of Art History.* London: Oxford University Press.

Ross, S. (2014). In Praise of Overstating the Case: A review of Franco Moretti, Distant Reading (London: Verso, 2013). *Digital Humanities Quarterly*, 8(1). http://www.digitalhumanities.org/dhq/vol/8/1/000171/000171.html (Accessed 10 December 2017).

Saussure, F., (1916). *Cours de Linguistique Générale*. Paris: Payot.

Schafer, R. M., (1993). *The Soundscape: Our Sonic Environment and the Tuning of the World*. Rochester, VT: Destiny Books.

Scott, C. (1999). *The Spoken Image: Photography and Language*. London: Reaktion Books.

Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R., (2000). Content-Based Image Retrievalat the End of the Early Years. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12): 1349-1380.

Spigel, L. (1994). *Make Room for TV: Television and the Family Ideal in Postwar America*. Chicago, IL: University of Chicago Press.

Sterne, J., (2003). *The Audible Past: Cultural Origins of Sound Reproduction.* New York: Duke University Press.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017(1): 1-9.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A., (2017). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of Association for the Advancement of Artificial Intelligence*, 2017(1): 4278-4284.

Underwood, T., (2017). A Genealogy of Distant Reading. *Digital Humanities Quarterly*, 11(2). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (Accessed 17 December 2018).

Figure Captions

Fig. 1 Detected features in two color photographs from the Farm Security Administration–Office of War Information (FSA–OWI) archive, a collection of documentary photography taken by the United States Government from 1935 to 1943. The right image shows detected images from the YOLOv4 algorithm (Redmon *et al.*, 2016): a horse, a person, and a dog. The library can detect 9000 classes of images. On the left are detected facial features—such as the location of eyes, nose, mouth, and jawline—from the three women as detected by the *dlib* library (King 2009).

Fig. 2 A plot showing frames where the main female character had *not* been seen for at least two minutes. Each row depicts a particular episode, roughly 25 minutes in length, from the 1966-1967 seasons of *Bewitched* and *I Dream of Jeannie*. Faces were detected using the distant viewing toolkit (Arnold and Tilton, 2017).

Fig. 3 The left column of this grid of photographs shows images selected from the Farm Security Administration - Office of War Information (FSA-OWI) archive, a collection of documentary photography taken by the United States Government between 1935 and 1943. To the right of each image are the seven closest other images in the collection using the distant metric induced by the penultimate layers of the InceptionV3 neural network model (Szegedy *et al.*, 2015). Notice that each row detects images with a similar dominant object: horses, wooden houses, pianos, train cars, and cooking pots.

Fig. 4 A network of the twenty staff-photographers with the largest set of credited photographs from the Farm Security Administration - Office of War Information (FSA-OWI) archive. Each photographer was connected to the photographer whose photos most resembled their own using the distant metric induced by the penultimate layers of the InceptionV3 neural network model (Szegedy *et al.*, 2015).

Notes

[i] That is, in Pierce's trichotomy of signs, most words function as *symbols* with no logical connection between the word and the concept represented by the word (2000). Proper nouns, in contrast, function as an *index* where the word has a direct relationship to the concept being represented. To illustrate how these signs function differently, note that proper names typically do not change in translation. Finally, a very limited number of words that exhibit onomatopoeia, such as 'sizzle' and 'splat', serve (in theory) as *icons*.

[ii] It is possible to have an image *which itself depicts another code system*, such as the scanned image of a textual document. The analysis in this article generally focuses on image collections that do not consist of such self-contained code systems.

[iii] Throughout, we use the term 'photograph' to broadly include manual and digital still photography, film, video and any other methods for recording a measurement of light in an attempt to replicate the human visual system.

[iv] Erwin Panofsky's three levels of understanding in art history—primary, iconography, and iconology—offer a similar description of the levels of interpretation within the secondary meaning of visual materials (Panofsky, 1939).

[v] Some pre-processing must first be applied as the text typically needs to be split apart into words by the process known as *tokenization*. For written text, however, this process can usually be accomplished unambiguously through a simple deterministic algorithm (language such as Japanese and Cherokee that make use of a syllabary require additional work).

[vi] Work by both Jannidis and Flanders (2013) and Ciula and Øyvind (2017) has argued that models should themselves be considered a semiotic system within DH. This line of work complements the claim here in understanding the role modeling within *distant viewing* but is not directly related. We are arguing that the *output* of the model yields the code system of interest here rather than the models themselves.

[vii] Here we are referring to images as stored by raster graphics, which is only format capable of storing photographic data. Vector graphics, which are inherently digital in nature, store information such as lines, bounding boxes or geographic outlines. Formats for vector graphics function differently and cannot be arbitrarily rescaled. These types of graphics are generally outside the discussion of image corpora presented in this article.

[viii] These systems make use of graphical programming units (GPUs), original designed for high-performance computer gaming. GPUs do not have the ability to perform all of the general-purpose programming that can be run on a CPU. They are, however, able to compute their limited set of operations extremely fast. These operations correspond to those required for training and running neural networks.

[ix] *Bewitched* was reimagined as a major theatrical film in 2005. Both shows continue to be actively syndicated in the United States.

[x] We are aware of a few other uses of the term *distant viewing* within the computational humanities, including Bender (2015) and John *et al.* (2017), in which it is used broadly to refer to the computational analysis of large corpora of visual materials.

[xi] For further discussion of the modern term 'distant reading'—particularly in the wake of claims made against Moretti as part of the #MeToo movement—see Lauren Klien's "Distant Reading after Moretti" (2018).

[xii] Though, the undergirding theory of *distant viewing*—that one must develop a code system and algorithmically convert objects into this code system before analysis—can be sensibly applied to audio and audio-visual materials. For example, our Distant Viewing Toolkit (Arnold and Tilton, 2017) extracts both visual and audio features from moving images.