No cover image available

The Oxford Handbook of the Sociology of Machine Learning

(In Progress)

Christian Borch (ed.), Juan Pablo Pardo-Guerra (ed.)

https://doi.org/10.1093/oxfordhb/9780197653609.001.0001

Published: 2023 **Online ISBN:** 9780197653630 **Print ISBN:** 9780197653609

CHAPTER

Analyzing Image Data with Machine Learning

Han Zhang

https://doi.org/10.1093/oxfordhb/9780197653609.013.17

Published: 18 December 2023

Abstract

Visual information such as images and videos is abundant in the digital age. However, visual data are currently underutilized in social science research. Previous sociological studies that rely on visual information as data either use qualitative interpretation or small-scale quantitative analysis, which is not suitable for large-scale datasets. This chapter first provides a historical overview of automated image analysis using computer vision. It then discusses the state-of-art deep learning algorithms and summarizes the two steps of using deep learning algorithms to automatically analyze large image datasets. It uses various publications using large-scale image datasets ranging from social media images, satellite and street view images, and historical archives to illustrate how automatic image analysis was carried out in recent research. Finally, the chapter discusses ways in which social scientists can better utilize visual information in their future research.

Keywords: image data, computer vision, machine learning, deep learning, social media, satellite image, street view image, historical archive

Subject: Social Research and Statistics, Social Theory, Sociology

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

Visual Information

"A picture is worth a thousand words." This statement is substantiated by both psychological research and historical evidence. The dual-coding theory in psychology posits that humans process visual information more rapidly than text-based information (Paivio, 1990). Studies have demonstrated that visual information is more effective at capturing attention, increasing credibility, and enhancing audience engagement (Pieters & Wedel, 2004). Examining human history, we discover that the earliest known paintings date back 40,000 years, while written text emerged only around 3,000 B.C.E.—approximately 5,000 years ago. Evidently, the power of visual communication has been acknowledged and deployed far longer than written language has been.

In the pre-digital era, the storage and production of visual data were significantly more cumbersome than they were for textual data. Furthermore, during the early stages of the internet, text-based information prevailed because of bandwidth and storage constraints for images and videos. However, the digital revolution has entirely altered the landscape of data storage and distribution. Notably, a vast corpus of analog data, such as historical maps and images, is now being digitized. An even larger volume of visual data is being inherently generated in digital formats, encompassing a variety of videos, photographs, and digital maps found on popular social media platforms like YouTube, TikTok, Instagram, Pinterest, and Google Maps. Pew Research's 2022 survey reveals that three out of the four most widely used social media platforms in the United States primarily rely on visual information: YouTube (first), Instagram (third), and Pinterest (fourth), while Facebook, which incorporates both text and images, occupies the second position. Among younger generations, the preference for creating and consuming visual content is even more pronounced, as exemplified by the popularity of TikTok. Consequently, the dominance of visual data over text-based data is set to increase in the foreseeable future.

Visual information holds a distinct advantage over textual information, as it is more easily transferable across cultures. Comprehending written text in another language often necessitates the assistance of interpreters or translation software, which can introduce errors and distortions. However, visual content can cross language barriers more easily. For instance, social media images posted on Chinese platforms are more likely to convey their intended meaning to an American social media user without the need for interpretation or translation. Consequently, the inherent transferability of visual information across cultures makes it a powerful tool for communication and understanding, particularly in a globalized world where diverse audiences are increasingly connected through digital media. This has significant implications for developing methods: Statistical models trained on images are more easily "transferable" than texts across cultures are.

Despite the growing prevalence of visual data in contemporary society, social science research on the analysis of image data has not kept pace with the developments in text-as-data analysis techniques that have gained popularity since the 2010s. While text-as-data analysis has facilitated the development of methods, data collection procedures, and empirical research, there has been a marked disparity in the level of attention given to image data analysis. Although a nascent community of scholars has emerged around the analysis of visual data, there remains a dearth of research in this area compared to the significant body of work dedicated to the analysis of textual data. Given that visual information is becoming increasingly abundant in many spheres of social life, the relative paucity of research on image data analysis is cause for concern. There is an urgent need for social science researchers to develop robust and rigorous methods for analyzing visual data and to apply these methods to empirical research in order to capture the rich and complex information contained in these data.

This chapter presents a comprehensive review of the history of visual data analysis in both social sciences and computer sciences, with a particular focus on the period prior to 2010. Additionally, this chapter will

discuss the recent advances in computer vision research that have revolutionized the study of visual information and enabled automated analysis on a large scale. Subsequently, the chapter will describe the efforts of social scientists to introduce and develop computer vision methods for analyzing visual data. These efforts will be categorized by the types of data used, the methods deployed (supervised and unsupervised), and the substantive topics that have been examined empirically. This chapter places particular emphasis on images. Image data serve as the building block for more complex types of visual information such as videos, which can be viewed as a panel of images aligned in time.

History of Using Visual Information in Sociology

The occasional use of images in sociological studies has a long history; much of ethnographic studies uses photos taken by authors to illustrate the field environment. However, systematic studies that use image as data are relatively less common, and most existing studies using image data tend to focus on how visual information has been essential in spreading social movements, political events, and political communication. Notably, in 2012, *American Behavioral Scientist* organized a special issue titled "The Power of Pictures: Images of Politics and Protest." Within this special issue, Corrigall-Brown and Wilkes (2012) compared how textual contexts differ from photographs in media reports of the Oka Crisis protests in 1990 between Indigenous peoples and Quebecois and Canadian authorities. They found that image framing of the events was more nuanced than textual content was. They produced 1,439 subject-level items of information portrayed in the photographs, such as perceived relative power and emotional facial expressions. Fahmy and Neumann (2012) examined 647 images in major news media outlets and found that their selection of photos portraying the Gaza War shaped viewers' perception of the war, which could ultimately affect public opinions more broadly. Vliegenthart (2012) examined 225 posters from 23 Dutch political parties, focusing on the elements that changed from 1946 to 2006. Keith (2012) compared the visual coverage of the 60th and 65th anniversaries of the liberation of Paris. Rohlinger and Klein (2012) collected 2,093 news images related to abortion in various U.S. news media reports (print and electronic). The review of these articles in the special issue of *American Behavioral Scientist* provides insight into the relative focus of prior studies using image data.

Another branch of scholars has expressed interest in using visual data to study microinteractions. Although Collins (2009) did not fully use the video data to develop his famous micro-situational theory of violence emergence, followers of Collins have attempted to analyze the emergence of violence using CCTV video footage (Levine et al., 2011). Collins (2015) later envisioned that analyzing image and video data could foster a new generation of research on studying microinteractions, a core tradition in sociological analysis that had historically rejected quantitative analysis. Collins noted that in addition to ethnographic observations, videos also capture everyday life. The type of information captured range from passively recorded sources like CCTV videos to actively recorded sources like social media sites, all of which could be used to observe the behavior of bodies, emotions, and everyday interactions.

Methodologically, some scholars use images in a qualitative, interpretative approach. Others try to quantitatively analyze the data to map images into predefined categories to generate measurement for theoretical constructs (Corrigall-Brown & Wilkes, 2012; Fahmy & Neumann, 2012; Vliegenthart, 2012). This approach belongs to the broader literature on content analysis. For instance, the classic textbook on content analysis by Krippendorff (2004) defines the goal of content analysis as analyzing "texts, images, and expressions that are created to be seen, read, interpreted, and acted on for their meaning." Although almost all chapters in Krippendorff (2004) focus on text-based content analysis as examples, these methods be easily transferred to analyzing image content, from sampling to constructing analytical constructs and conducting intercoder reliability analysis.

For example, for their study of more than 2,000 images from U.S. news media reports about abortion, Rohlinger and Klein (2012) began by collecting images, which they then manually classified into predefined "visual landscapes" (p. 172), which they describe as "professionally constructed images that are designed to capture an issue for a broader public." (p. 175). Examples included TV anchors, pro-life demonstrations or protests, government officials, and government buildings. The main analysis is presented as estimating the proportion of each visual landscape and their variations across different abortion events (e.g., around *Roe vs. Wade*, 1973, or presidential election debates on abortions). This content-analysis style of quantitative approach is straightforward, but it has two significant limitations. First, it transforms each image into a single number, effectively depriving the images of their rich context. Second, this manual classification is not scalable for analyzing millions of images, as is commonly needed in the age of big data. In the next section, I introduce automated image analysis that can use machines to process millions of images.

Automated Image Analysis

In automated image analysis, researchers utilize images as data for two distinct objectives, which are categorized as *supervised* and *unsupervised* tasks in the field of computer science. On one hand, in supervised tasks, researchers aim to generate a measure of a clearly defined theoretical construct from images. For instance, they may be interested in quantifying the levels of anger, disgust, happiness, and sadness in a collection of politician images (Boussalis & Coan, 2021). On the other hand, unsupervised tasks involve researchers working with a large set of images without a specific goal of what to measure (Zhang & Peng, 2022). The primary objective of unsupervised tasks is generally to explore the dataset and identify prevalent patterns or clusters within it. The distinction between supervised and unsupervised tasks is not always well-defined, and it is not unusual for scholars to perform unsupervised clustering before determining the precise measurements that they wish to obtain and seeking the support of supervised machine learning algorithms.

Image Representation

Before machines apply supervised or unsupervised methods, they need to transform natural data such as text and images into numeric values. Compared with textual content, images have a natural representation in modern computers, as they are represented as matrices with numeric values or pixels. For example, a typical 8-bit 100x100 grayscale image is represented as a 100x100 matrix, with each cell (i.e., pixel) taking values ranging from 0, which indicates black, to 255, which indicates white. One can also flatten the matrix into a long vector by stacking each row (or each column) together. In a typical red, green, and blue (RGB) color image, each image becomes a three-dimensional matrix with each matrix representing the extent of red, green, and blue color.

The pixel representation of images presents an intrinsic challenge: high dimensionality. For instance, a typical RGB image consisting of 800 × 600 pixels has a vector representation of length $800 \cdot 600 \cdot 3 = 1{,}440{,}000$. Using this vector as variables in a regression model would result in a model with 1.44 million variables. Such high dimensionality not only poses significant challenges in terms of storage requirements but also makes algorithms more complex and time consuming to run on such large datasets.

As a result, transforming images into shorter yet meaningful vector representations is a critical step in performing image analysis. In academic terms, we transform pixel representations of images into a low-dimensional vector representation. This transformation allows for more manageable data storage and easier algorithm operations. Moreover, such compression may effectively remove some noise information that is irrelevant to the problems at hand. The importance of low-dimensional transformation may be

underestimated for social scientists, particularly among individuals accustomed to analyzing survey and administrative data using regression analysis.[1]

During the initial stages of automated image analysis, specific methods were designed to convert images into low-dimensional vectors for particular tasks. These methods included algorithms capable of extracting contours of objects, detecting angles, or identifying color changes (Lowe, 2004).[2] More advanced techniques combined features extracted from different methods, such as the bag-of-visual-words model (Lowe, 2004; Koniusz et al., 2017).

Since the 2010s, deep learning has dominated every aspect of automated image analysis. It offers a unique way to map images into low-dimensional vectors, also referred to "representation learning" in the computer science literature. In fact, the founding fathers of deep learning, in their review article, LeCun et al. (2015), emphasize that the fundamental change of deep learning algorithms is their ability to "learn from data using a general-purpose learning procedure" that is designed by human researchers to "automatically discover the representations needed for detection or classification (p. 436)." In simpler terms, we no longer need to invent ten different algorithms for ten different tasks, as previous scholars did. Instead, deep learning can train one model on a massive dataset with millions of images and apply and adjust the model (referred to as fine-tuning) to specific tasks rather than start from scratch. This is because deep learning has multiple layers (hence "deep"), with the initial layers being capable of extracting basic features such as edges, while the later layers can learn to combine these features. The final layers form a low-dimensional vector whose length typically ranges from hundreds to thousands, which is several magnitudes smaller than the length of the vector that has been transformed from raw pixel representation. Although this low-dimensional vector lacks interpretability (which I will address later), empirical research has found that using the vector as a variable in subsequent regression or classification tasks achieves superior performance compared to non-deep learning algorithms.

In terms of practical considerations, here I provide some recommendations for applied researchers who may be new to analyzing image data. *Architecture* refers to the structure of a deep learning model (e.g., the number of layers and how internal layers are constructed), *dataset* means the input datasets and output categories, and *model* refers to the combination of an architecture, input dataset, and the types of outcomes produced. Social scientists who are familiar with regression models spend more time choosing the regression specification, which is equivalent to the architecture in a model, and spend less time on the dataset, because the input and outcome are clearly defined in the research design process. However, when using deep learning algorithms for image analysis, especially with pre-trained models, it is essential to pay more attention to the dataset aspect—specifically, to (a) the input dataset used to train the models and (b) the categories in the original, pre-trained models. The architecture is less critical unless one decides to train his or her own model, which can be resource intensive and requires advanced programming skills. This is a substantial mental shift for social scientists.

Scholars first need to compare the input dataset and the outcomes in the pre-trained model, as well as their own dataset and the types of outcomes they want to obtain. If both are similar across the pre-trained model and the problem at hand, the vectors extracted from the pre-trained models will likely be more useful. For example, the well-known ImageNet dataset, which ignited the deep learning revolution, does not contain many human faces, but it contains other types of image categories (e.g., animals). If one's research goal is to examine the presence of pets, a pre-trained model based on ImageNet is ideal, and it is in fact the default choice for many software programs. However, if one's research goal is facial recognition or an analysis of facial expressions, it is not suitable to use models trained on the ImageNet dataset. Instead, one should find a model with the same architecture that has faces in the training data and is designed for facial recognition, such as VGGFace (Parkhi et al., 2015).

Since social scientists' datasets are typically much smaller than those used to train large deep learning models, achieving similar performance to pre-trained models is unlikely. Therefore, it is advantageous for researchers to use a pre-trained model from the outset. One only needs to input the images into a pre-trained model, extract the last layer, and use that as the input. Knowledge of Python programming is beneficial, as many mature packages from TensorFlow to Keras and PyTorch offer readily available pre-trained models.

If a suitable model whose training data are similar to researcher's own cannot be found, an alternative is to fine-tune the model. Fine-tuning involves retaining the initial layers of a pre-trained model—as they extract basic features—and enabling the modification of the last few layers so that the algorithm training process can learn how to combine these features.

If fine-tuning still does not meet one's needs, she can train her own model. However, this comes at the cost of needing to collect a large-scale training dataset (typically larger than hundreds of thousands) and managing the training process to avoid overfitting issues. As such, this approach is not recommended for beginners.

Classification, Clustering, and Validation

As long as low-dimensional vectors are obtained, the next step is to use them in supervised or unsupervised tasks and to validate the results. This step is not unique to image analysis; scholars need to perform the same process when analyzing image, visual, or other types of quantitative data. Therefore, I will briefly introduce the workflows. Scholars can refer to standard textbooks like Hastie et al. (2009) and those written for social scientists, such as Grimmer et al. (2021).

In the first step, scholars map their low-dimensional vectors into predefined categories (supervised tasks) or allow the algorithm to automatically discover categories from the low-dimensional vectors. Some researchers prefer an end-to-end approach, meaning that they do not extract the low-dimensional vectors but keep them in the original deep learning model, which takes the low-dimensional vector as input and uses a fully connected neural network for classification (e.g., VGG) (Simonyan & Zisserman, 2015) or clustering (Kipf & Welling, 2016). Others extract the vectors and input them into supervised (e.g., linear or logistic regression, decision trees, and random forests) or unsupervised (e.g., PCA, K-means, hierarchical clustering) methods that suit the problem at hand. Both approaches have been widely used, with the first approach theoretically yielding better performance but being slightly more challenging to implement.

For supervised tasks, scholars have a predefined list of categories that they want machines to assign images to. Hence, the next step is to validate image classification results. If scholars apply existing models, they may not always have training data, a portion of which can be used as validation data. In this case, scholars must create their own validation data. Essentially, they need to sample a list of images in their dataset, keep them out of the training process (if there is a fine-tuning process), and use the trained or pre-trained model to classify the held-out validation dataset. Standard metrics for evaluating article classification performance include F1, precision-recall curve, and ROC curves (receiver operating characteristic curve) (Grimmer et al., 2021).

If scholars are conducting unsupervised tasks, validation becomes more challenging (but arguably more important). The difficulty lies in the absence of a "gold standard" for good clustering results. Consequently, one must try different values of K (the number of groups for the dataset), review results, and assess if they make theoretical sense. Zhang and Peng (2022) discussed several data-driven and theory-driven approaches to validate clustering results in the context of image analysis.

Examples of Automated Image Analysis in Sociology

Article algorithms have enabled scholars to analyze data at an unprecedented scale. I next discuss more recent literature that analyzes, in some cases, millions of images using automated image analysis. I categorize them according to the types of big data sources used, as these sources determine the types of questions that scholars have been able to address thus far.

Social Media Images

Social media platforms provide easily accessible data sources containing visual information. Because of the historical legacy of social movements and broader political communication, scholars studying these topics make up the majority of the field that focuses on these data. However, as there are relatively fewer publications by sociologists, I also include work by political scientists and communication scholars, given that many of the methods can be applied to traditionally sociological topics as well.

Researchers studying social movements have used images from protests either to identify protest events and their features or to examine the mobilization power of such protests. Zhang and Pan (2019) utilized images and text information from Chinese social media posts to construct one of the first large-scale protest event datasets for China. Traditional sources of protest event data, such as newspapers, have been under tight control, making it difficult for protest events to appear in printed forms (Cheng & Lu, in this volume). However, social media, despite censorship, still offer plenty of space for citizens to discuss protests they have participated in or witnessed. Zhang and Pan (2019) found that incorporating image data as a source is crucial, whereas using only text content, as previous generations of protest scholars have done, does not achieve satisfactory machine classification performance. Sobolev et al. (2020) used Twitter protest images to estimate protest size with crowd size estimation algorithms, finding their estimations to be as accurate as journalist reports are whereas text-based estimations provided subpar performance. Casas and Williams (2019) and Steinert-Threlkeld et al. (2022) provided evidence that images are critically important during mobilization because of their ability to arouse emotions such as anger.

Another group of studies focuses on political communication, featuring work from scholars in both political science and communication fields. Among communication scholars, many studies have explored the role of image features in generating increased social media attention. For example, Peng (2018) examined media biases in portraying politicians in over 13,000 images during the 2016 U.S. presidential election. Boussalis and Coan (2021) collected footage from a two-hour debate in the 2016 U.S. Republican primary, splitting the video into over 200,000 static images with politicians' faces. They detected emotions from these faces using Microsoft's Face API and used those as independent variables in studies to examine a focus group's support level when viewing each frame and the associated emotions. They found that anger emotions triggered more support during debates. Peng (2021) first performed unsupervised clustering to identify the patterns over 59,000 images posted by U.S. politicians on Instagram. He identified two large categories: those that focus on self-personalization, featuring politicians in private, non-political settings, versus images that still adopt the traditional style that features politicians in political environments. He then found that self-personalization is more effective in leading to more user engagement on Instagram. Zhang and Peng (2022) later systematically discussed unsupervised methods for clustering images and using social media images as examples.

Satellite and Street View Images

Another major source of image data comes from night light images, daytime satellite images, and street view images from map providers such as Google. The use of these three data sources has emerged over time in different fields of study. Initially, the values of satellite images were noted in geography and urban studies; this was followed by economists who were interested in measuring economic growth and development; finally, computer scientists began working on big data methodology. Now, sociologists have also recognized the value of these data sources and have begun to explore the possibilities (Lund, in this volume).

Night light images are the most intuitive and easiest to start with. Variations in night light density capture the variations in economic development levels. Analyzing night light density is also relatively straightforward, without much need for deep learning algorithms. In the simplest sense, one just needs to add the pixel values of each night light image, which offers a quick measure of the development levels of a region. However, more concerns have been raised in the past decade about the added value of such data: Areas with substantial night light density also tend to be developed regions where official statistics and survey data on socioeconomic status are usually available. However, developing areas lack both variations in night light density and official statistics and survey data (Gibson & Boe-Gibson, 2021; Gibson et al., 2021). As a result, the added value of night light data beyond existing data is not substantial. Worse still, the variations in night lights are usually not salient enough to explore fine-grained regional variations. This dilemma has prompted many scholars to consider alternative data sources, such as daytime satellite images, which I will refer to as satellite images.

Daytime satellite images share the advantage of being a global-scale data source, and they can provide a look back in time, offering the potential to construct panel datasets (Donaldson & Storeygard, 2016; Jean et al., 2016; Yeh et al., 2020). Using daytime satellite images requires more methodological effort because it is demanding for humans to judge what types of images reflect more developed regions. Typically, scholars collect a list of images, tag them (label the outcome variables that they want to measure, such as GDP or poverty level), and then feed them into supervised article algorithms, following the standard article exercises. In other cases, scholars will first segment the images to identify the relevant objects that they believe to be important for their analysis.

In recent years, street view images have gained recognition and have been utilized as data sources (Hwang & Sampson, 2014; Hwang & Ding, 2020). In contrast to night light or daytime satellite images, street view images are captured by map providers' vehicles as they navigate through cities, providing a closer view of urban environments rather than a perspective from thousands of miles above. The difference in viewing angle and distance potentially offers more detailed information, but it comes at the cost of reduced comprehensiveness and timeliness compared to satellite images. This is because the decision of where and when to collect street view images largely depends on the choices made by companies like Google or other local map providers, such as Baidu in China, and the frequency of data collection can vary considerably across cities.

Street view images from Google, Baidu, and other map providers also offer valuable close-up information about urban landscapes. The analysis steps are similar to those involved in using daytime images. Researchers can directly train supervised article algorithms to map the input images to predefined outcome variables. Alternatively, some take a two-step process: First, they perform semantic segmentation or object detection to identify relevant objects that they believe are related to outcomes. Then they train a predictive model that links the objects to the outcomes. For instance, Gebru et al. (2017) deployed a two-step process. They first performed object detection to extract cars from over 50 million street view images in 200 U.S. cities. Then they built a regression model that linked car types to various demographic outcomes at the

community level. The rationale behind this approach is that the cars that residents drive in a community can indicate the socioeconomic status of those residents at an aggregate level.

Historical Archives

Historical archives have also been digitized and deployed as another source of visual information. Unlike social media, satellite, and street view photos, which originate from a few large, established sources such as big companies or government agencies, historical archives can come in various formats, different sources, and may require distinct cleaning processes. Therefore, it is hard to tell scholars which sources in which they should find image data that fit their needs. I discuss two exemplary articles below to serve as inspirations. Consider recent research by Cantú (2019), which investigates factors that contribute to altered vote tallies in the 1988 presidential election in Mexico. The author applied a convolutional neural network, a standard image classification algorithm, to images of ballots to produce a binary prediction of whether vote tallies were altered. The machine-predicted binary variable was then regressed on a set of independent variables to understand which factors are associated with vote tally alteration.

Muller-Crepon et al. (2021) utilizes historical maps. Their research examines whether regions with low state capacity are more likely to experience civil conflicts. They argue that traditional measures of state capacity, such as tax revenue, are not precise enough. Instead, they propose that the road network (e.g., the driving distance between two cities) to capitals measures the control of central government over local authorities and should serve as a more accurate measure. Scholars could use map providers and their navigation tools to measure the road network at the present time. However, this approach is not feasible for Muller-Crepon et al. (2021), as they aim to analyze road networks in the 20th century. In order to create such a measure, Muller-Crepon et al. (2021) digitized historical maps in Africa from 1966 to 1990. They then used convolutional neural networks to detect roads from the digitized network and produced a measure of the time it took to travel from national capitals to any city.

Future Directions of Visual Information and Its Opportunity for Social Scientists

In this section, I explore some potential future directions that may emerge in the foreseeable future. The most straightforward extension involves using video as data. Most current studies that utilize video data treat it as a panel of static images (Boussalis & Coan, 2021) and do not further consider the temporal order of these static images, which can lead to a loss of information related to intrinsic differences between video and a collection of images. Nassauer and Legewie (2021) argue that Levine et al. (2011) provide a good example of an approach that seeks to reconstruct the sequences of violence emergence during microinteractions, which requires teasing out the time order of different types of static images rather than treating them as exchangeable over time. Although Levine et al. (2011) manually constructed their sequence, the idea of constructing the temporal patterns based on static images from videos is an area that deserves more scholarly attention.

The second direction involves combining visual and non-visual information together in analysis. Notably, text-as-data has led the methodological development of the social sciences in the past decade (Grimmer & Stewart, 2013). However, the growing trend toward using image-based and visual media suggests that visual information will become only more important in the future. I argue that focusing on either text (as most text-as-data scholars do) or images (as do most of the studies discussed in this chapter) is not ideal because the former discards image information and the latter discards text information. Zhang and Pan (2019) offer a unique example in which evidence is presented that text and images are both essential in

detecting protest events from social media. This serves as a good example of the need to develop methods for integrating text and image analysis in future studies.

Third, large models have created more opportunities for future image analysis. On one hand, these large models provide the potential for significantly faster speeds for users to perform supervised or unsupervised tasks. For instance, OpenAI's Contrastive Language-Image Pre-training, or CLIP, model presents practitioners with excellent opportunities for generating text descriptions from images. Meta's Segment Anything Model can ease the second step of semantic segmentation and allow researchers to conduct supervised tasks more easily. On the other hand, with much AI-generated visual content, it is interesting to examine the consequence of such content on human behaviors and society, which is naturally an empirical research problem.

References

Boussalis, C., & Coan, T. G. (2021). Facing the electorate: Computational approaches to the study of nonverbal communication and voter impression formation. *Political Communication*, *38*(1–2), 75–97.
Google Scholar WorldCat

Cantú, F. (2019). The fingerprints of fraud: Evidence from Mexico's 1988 presidential election. *American Political Science Review*, *113*(3), 710–726.
Google Scholar WorldCat

Casas, A., & Williams, N. W. (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, *72*(2), 360–375.
Google Scholar WorldCat

Collins, R. (2009). *Violence: A micro-sociological theory*. Princeton University Press.
Google Scholar Google Preview WorldCat COPAC

Collins, R. (2015). Visual micro-sociology and the sociology of flesh and blood: Comment on Wacquant. *Qualitative Sociology*, *38*(1), 13–17.
Google Scholar WorldCat

Corrigall-Brown, C., & Wilkes, R. (2012). Picturing protest: The visual framing of collective action by First Nations in Canada. *American Behavioral Scientist*, *56*(2), 223–243.
Google Scholar WorldCat

Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, *30*(4), 171–198.
Google Scholar WorldCat

Fahmy, S., & Neumann, R. (2012). Shooting war or peace photographs? An examination of newswires' coverage of the conflict in Gaza (2008–2009). *American Behavioral Scientist*, *56*(2), NP1–NP26.
Google Scholar WorldCat

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, *114*(50), 13108–13113.
Google Scholar WorldCat

Gibson, J., & Boe-Gibson, G. (2021). Nighttime lights and county-level economic activity in the United States: 2001 to 2019. *Remote Sensing*, *13*(14), Article, 2741.
Google Scholar WorldCat

Gibson, J., Olivia, S., Boe-Gibson, S., & Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, *149*, Article 102602.
Google Scholar WorldCat

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*(1), 395—419.
Google Scholar WorldCat

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.
Google Scholar WorldCat

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2). Springer.
Google Scholar Google Preview WorldCat COPAC

Hwang, J., & Ding, L. (2020). Unequal displacement: Gentrification, racial stratification, and residential destinations in Philadelphia. *American Journal of Sociology*, *126*(2), 354–406.
Google Scholar WorldCat

Hwang, J., & Sampson, R. J. (2014). Divergent pathways of gentrification: Racial inequality and the social order of renewal in Chicago neighborhoods. *American Sociological Review*, *79*(4), 726–751.
Google Scholar WorldCat

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794.
Google Scholar WorldCat

Keith, S. (2012). Forgetting the last big war: Collective memory and liberation images in an off-year anniversary. *American Behavioral Scientist*, *56*(2), 204–222.
Google Scholar WorldCat

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv.
https://doi.org/10.48550/arXiv.1609.02907

Koniusz, P., Yan, F., Gosselin, P.-H., & Mikolajczyk, K. (2017). Higher-order occurrence pooling for bags-of-words: Visual concept detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(2), 313–326.
Google Scholar WorldCat

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.
Google Scholar Google Preview WorldCat COPAC

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
Google Scholar WorldCat

Levine, M., Taylor, P. J., & Best, R. (2011). Third parties, violence, and conflict resolution: The role of group size and collective action in the microregulation of violence. *Psychological Science*, *22*(3), 406–412.
Google Scholar WorldCat

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.
Google Scholar WorldCat

Muller-Crepon, C., Hunziker, P., & Cederman, L.-E. (2021). Roads to rule, roads to rebel: Relational state capacity and conflict in Africa. *Journal of Conflict Resolution*, *65*(2–3), 563–590.
Google Scholar WorldCat

Nassauer, A., & Legewie, N. M. (2021). Video data analysis: A methodological frame for a novel research trend. *Sociological Methods & Research*, *50*(1), 135–174.
Google Scholar WorldCat

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.
Google Scholar Google Preview WorldCat COPAC

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. *Procedings of the British Machine Vision Conference*, *41*, 1–12.
Google Scholar WorldCat

Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, *68*(5), 920–941.
Google Scholar WorldCat

Peng, Y. (2021). What makes politicians' Instagram posts popular? Analyzing social media strategies of candidates and office holders with computer vision. *International Journal of Press/Politics*, *26*(1), 143–166.
Google Scholar WorldCat

Pieters, R., & Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of marketing*, *68*(2), 36–50.
Google Scholar WorldCat

Rohlinger, D. A., & Klein, J. (2012). Visual landscapes and the abortion issue. *American Behavioral Scientist*, *56*(2), 172–188.
Google Scholar WorldCat

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the Third International Conference on Learning Representations*. Computational and Biological Learning Society, 1–14.
Google Scholar Google Preview WorldCat COPAC

Sobolev, A., Chen, M. K., Joo, J., & Steinert-Threlkeld, Z. C. (2020). News and geolocated social media accurately measure protest size variation. *American Political Science Review*, *114*(4), 1343–1351.
Google Scholar WorldCat

Steinert-Threlkeld, Z., Chan, A., & Joo, J. (2022). How state and protester violence affect protest dynamics. *Journal of Politics*, *84*(*2), 798–813*.
Google Scholar WorldCat

Vliegenthart, R. (2012). The professionalization of political communication? A longitudinal analysis of Dutch election campaign posters. *American Behavioral Scientist*, *56*(2), 135–150.
Google Scholar WorldCat

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic wellbeing in Africa. *Nature Communications*, *11*(1), Article 2583.
Google Scholar WorldCat

Zhang, H., & Pan, J. (2019). CASM: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, *49*(1), 1–57.
Google Scholar WorldCat

Zhang, H., & Peng, Y. (2022). Image clustering: An unsupervised approach to categorize visual data in social science research. *Sociological Methods & Research*.

Notes

1 In most typical survey and administrative datasets, the dataset has a significantly higher number of rows (e.g., thousands) than columns (e.g., dozens) representing variables. Consequently, there is little demand to transform each row vector into a low-dimensional vector because their length will only be dozens of columns. Instead, greater attention is typically devoted to selecting appropriate statistical models and determining which variables should be included in the analysis.

2 Each of these algorithms transformed pixel representations into low-dimensional vectors. Once the algorithm extracted the contours, for instance, what was inside the contour became irrelevant and the values were assigned as 0, which rendered them effectively irrelevant for statistical analysis.