Automated Visual Analysis for the Study of Social Media Effects: Opportunities, Approaches, and Challenges

Yilang Peng[a], Irina Lock[b], and Albert Ali Salah[c]

[a]Department of Financial Planning and Housing and Consumer Economics, University of Georgia, Athens, Georgia; [b]Institute of Communication Science, Friedrich Schiller University Jena, Jena, Germany; [c]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands

ABSTRACT

To advance our understanding of social media effects, it is crucial to incorporate the increasingly prevalent visual media into our investigation. In this article, we discuss the theoretical opportunities of automated visual analysis for the study of social media effects and present an overview of existing computational methods that can facilitate this. Specifically, we highlight the gap between the outputs of existing computer vision tools and the theoretical concepts relevant to media effects research. We propose multiple approaches to bridging this gap in automated visual analysis, such as justifying the theoretical significance of specific visual features in existing tools, developing supervised learning models to measure a visual attribute of interest, and applying unsupervised learning to discover meaningful visual themes and categories. We conclude with a discussion about future directions for automated visual analysis in computational communication research, such as the development of benchmark datasets designed to reflect more theoretically meaningful concepts and the incorporation of large language models and multimodal channels to extract insights.

Communication scholars have increasingly recognized the importance of analyzing visual content when investigating media effects, or "within-person changes in cognitions (including beliefs), emotions, attitudes, and behavior that result from media use" (Valkenburg & Peter, 2013), on social media platforms. Visual content abounds in social media environments. In 2021, approximately 81%, 40%, 31%, and 21% of U.S. adults used YouTube, Instagram, Pinterest, and TikTok, respectively, which are all platforms heavily oriented toward the sharing of visual content (Auxier & Anderson, 2021). Prior research also reveals the differential kinds of effects that visual information has on people's attitudes and behaviors: in comparison to textual communications, people generally tend to pay attention, remember content, or feel a wide range of emotions when exposed to visuals (Casas & Webb Williams, 2019; Houts et al., 2006). In addition, social media users frequently use visual media to express identities, develop relationships, exchange information, and participate in public affairs (Casas & Webb Williams, 2019; Liu et al., 2017; Nesi et al., 2018; Pouwels et al., 2021). The self-effects resulting from visual media production, which deals with how media affect message creators or senders' perceptions, attitudes, and behaviors (Valkenburg, 2017), has gained increasing scholarly attention. For scholars who wish to understand how media content affects social media users' perceptions, attitudes, and behaviors, it is essential to incorporate visual media in the investigation.

Yet, the sheer amount of visual content poses a serious analytical challenge, and the rise of video content further complicates this task. Automated visual analysis, which involves the use of computer algorithms and tools to replicate human abilities to extract meaningful information and insights from

CONTACT Yilang Peng yilang.peng@uga.edu Department of Financial Planning and Housing and Consumer Economics, University of Georgia, Athens, Georgia

This is the uncorrected author proof, please check the DOI for the published version. Copyright with Taylor & Francis Group, LLC, 2023. Please cite as "Peng, Y., I. Lock, A.A. Salah, "Automated Visual Analysis for the Study of Social Media Effects: Opportunities, Approaches, and Challenges," Communication Methods and Measures, accepted for publication. DOI: http://dx.doi.org/10.1080/19312458.2023.2277956.

visual media, becomes critically needed (Joo & Steinert-Threlkeld, 2022; Webb Williams et al., 2020). Computer vision, a field of study dedicated to enabling computers to understand digital images and videos, provides communication scholars with many tools to analyze social visuals on a large scale. The images-as-data and video-as-data approaches call for more attention to the large number of online visuals in social scientific investigation (Joo & Steinert-Threlkeld, 2022; Webb Williams et al., 2020). There has been a burgeoning body of scholarship that uses automated visual analysis to characterize visual content, such as estimating characteristics of social media movements (Steinert-Threlkeld et al., 2022; Zhang & Pan, 2019) and investigating politicians' visual communication strategies (Bossetta & Schmøkel, 2023; Haim & Jungblut, 2021; Peng, 2018, 2021).

Still, the adoption of automated visual analysis in studying social media visuals is in its early phase, and existing studies have so far focused on a limited set of computer vision tasks (Chen et al., 2021). While the importance of studying visual content in advancing media effects theories is widely acknowledged, how existing computer vision tasks can contribute remains uncertain. We argue that one challenge hindering the adoption of automated visual analysis is a disconnection between the outputs from computer vision models and the theoretical concepts that are meaningful in studying media effects. Many computer vision tools, such as facial recognition and object recognition, are not designed to measure theoretical concepts typically found in communication research, and their utility in advancing communication theories needs to be further interrogated.

This article aims to address these gaps. Specifically, we provide guidance to scholars who are new to automated visual analysis and wish to understand the potential of these techniques, as well as their problems and limitations. We first review the theoretical opportunities for studying social media visuals and how automated visual analysis can advance media effects research in multiple directions. We then present several groups of computer vision techniques that could help communication scholars investigate social media visuals. We aim to demonstrate how these analytical tools can potentially be applied to fulfill various research needs in concept measurement and theoretical development, with the hope of stimulating further research in this domain. We end with a discussion of several pitfalls of computational tools our community should further engage with and future directions for this area of research.

Theoretical opportunities for studying social media visuals to understand media effects

The visual turn brought about by the digital communication environment has resulted in a constant flow of new visual content on social media (Araujo et al., 2020). The long-known fact that visuals exert differential effects on individuals' emotions, attitudes, and behavioral intentions than text (Kress & Van Leeuwen, 1998) has also triggered heightened interest in the effects of visual social media. Automated visual analysis can process large amounts of visual data; it thus has the potential to contribute new insights to the puzzle of small media effects (Valkenburg et al., 2021). We argue that automated visual analysis can contribute to our understanding of social media effects in three directions. First, automated visual analysis informs social media effects research by answering what kinds of visual contents people are exposed to on social media, how visuals are composed, and what patterns are observable across contexts. Second, it helps analyze how social media visuals affect the emotions, attitudes, and behaviors of different audiences. Lastly, the methods shed light on how users' characteristics shape visual posts and investigate the self-effects of sharing visuals. By presenting potential research questions regarding a variety of topics (e.g., visual politics, misinformation, body image) that span across multiple communication subfields, including political, health, and interpersonal communication (Table 1), we aim to inspire future communication research using automated visual analysis in and beyond these areas.

Description of visual social media content

Descriptive content analysis allows scholars to map content categories or themes in the information environment and locate important content attributes to draw inferences on communication theory.

Table 1. Theoretical opportunities for applying automated visual analysis and exemplar research questions in different topical domains.

Approach	Visual politics	Mis-/disinformation	Digital activism	Body image	Digital connections
Description of visual media content (What kinds of visuals are present in social media environments?)	How frequently do politicians display different emotional expressions on social media? Bossetta & Schmøkel, (2023); Farkas & Bene (2021)	How widespread is visual misinformation and what visual features characterize misinformation content? Chen et al., (2022); Yang et al., (2023)	What kinds of visual frames are used by climate change activists on social media? Molder et al. (2022)	To what extent are women subject to sexual objectification in music videos? Aubrey & Frisby, (2011)	What kinds of visual messages are used by young people to provide social support to their peers? Nesi et al. (2018) Pouwels et al. (2021)
Effects of visual social media contents on audiences (How do visuals affect message receivers' perceptions, attitudes, and behaviors?)	How do politicians' personalization strategies on Instagram affect audience engagement and impressions? Farkas & Bene, 2021; Peng, 2021	To what extent does cultivation theory apply to visual misinformation and thus have negative effects on democracy? Allen et al. (2020); Grinberg et al. (2019); Riddle & Martins, (2022)	How do visuals of different emotional appeals mobilize participants in online protests? Casas & Webb Williams, (2018)	How does exposure to sexually objectified models on Instagram affect viewers' body dissatisfaction and well-being Brown & Tiggemann, (2016); Kleemans et al. (2018); Vendemia & DeAndrea, (2018)	How do different types of smartphone activities affect individuals' sense of belonging, relationship satisfaction, and subject well-being? Muise et al. (2022)
Visual production, producer characteristics, and self-effects (How are visuals on social media influenced by producers' characteristics? How do making and sharing visuals affect message producers' perceptions, attitudes, and behaviors?)	How do politicians of different genders and political ideologies differ regarding their visual self-presentations on social media? Brands et al. (2021); Xi et al. (2020)	How do disinformation campaigns with different appeals use visuals to attract potential audiences? Bastos et al. (2023)	How does posting a selfie or meme in activism affect the sender's political identity and future participation? Gil de Zúñiga et al. (2014)	How do fitness influencers of different genders differ in the level and type of objectification on Instagram? Murashka et al. (2021)	How does sharing a visual message such as a group selfie affect the sender's experience of friendship closeness? Nesi et al. (2018); Pouwels et al. (2021)

Only when the content is known can its effects be studied (Krippendorff, 2018). Automated visual analysis helps scholars quantify the patterns and coverage of various content attributes, categories, or themes on a large scale just as in texts (Hornik et al., 2022; Riddle & Martins, 2022; Peng, 2017). Examples of visual attributes available in existing computer vision tools include: (1) human faces, including facial identities and facial attributes such as emotional expressions, orientation, skin status, and demographic characteristics (e.g., gender, age); (2) human bodies and their gestures and activities; (3) objects, scenes, and settings, as well as the locations of objects; (4) text contained in visuals; (5) aesthetic features such as brightness, color percentages, and visual complexity; (6) visual categories based on clustering, and (7) for videos, movement patterns among objects and spatial distances. For studying social media effects particularly, large-scale descriptions of visual contents allow researchers to (1) describe media use and exposure, (2) identify components and features of visuals that may affect

audiences, and (3) detect patterns and generalize across visual contents. The following discusses these applications in the contexts of political (misinformation) and health (body image) communication and digital activism (see Table 1).

The analysis of visual content in social media is complementary to the wide-spread text-based analyses. For example, research on misinformation illustrated that only a very small fraction of US-Americans daily media diet is composed of fake news (Allen et al., 2020) and even in election periods, misinformation is spread online by a very small audience (Grinberg et al., 2019). This helps to interpret the prevalence of the problem of misinformation and to better assess its potential detrimental effects on democracy (Lorenz-Spreen et al., 2023). Yang et al. (2023) demonstrated that misinformation is more widespread in visual posts compared to text posts on Facebook, showing the importance of incorporating visual analysis to comprehend the scope of misinformation on online platforms (Peng et al., 2023). By quantifying what content is consumed, for instance via tracking digital trace data or data donations of news readers (Ohme et al., 2023), scholars can more accurately predict the effects of these information diets for audiences (Wojcieszak et al., 2021) and better inform theory development on such current topics.

Second, automated visual analysis can identify what makes up visual content of different types. Computer vision algorithms were applied to thousands of science conspiracy videos, and uncovered the components of their visual frames (Chen et al., 2022). Conspiracy videos were darker and showed less variation in color than non-conspirative videos. Such findings can help explain what triggers the effects of conspiracy videos in terms of information processing.

Third, a big advantage of automated versus manual content analysis is that it can detect patterns in the data that are harder to find and generalize from a limited sample (van Atteveldt & Peng, 2021). For example, by analyzing sexual objectifications of females versus males across genres in music videos, Aubrey and Frisby (2011) found that female artists are more objectified than male musicians, which might have an impact on women's self-perception. Yet, the results were limited in terms of generalization to a specific time frame, genres, and charts. By means of automated visual analysis, the research could be expanded to more diverse music videos, allowing broader conclusions and inferences in terms of the effects of such content on women's self-objectification, also over time. There are many other examples of visual framing analysis for societal questions. Molder et al. (2022) used inductive visual framing analysis to check whether visual frames by climate change activists on social media are similar to one another. Typically, manual analysis is performed on several hundred images, but automated analysis can process much more in a shorter amount of time. Thus, approaching this research question with automated visual analysis would allow identifying patterns of climate change activist framing over time and across cultures, because it allows digestion of much more data.

Visual social media content and effects on audiences

Visual themes, attributes, and characteristics in visual social media content affect outcomes such as emotion, well-being, attitudes, and behaviors of users. Upscaling visual content analyses can inform experimental effects research in public health and communication on social media, because automated visual analysis allows assessing the prevalence of overarching themes and does not rely on a small sample of frequently watched videos or viewed image posts. In health communication, for instance, prior research has shown that exposure to idealized body images typically found on visual social media platforms affect particularly young women's self-image and may have negative effects on their well-being (Brown & Tiggemann, 2016; Kleemans et al., 2018; Vendemia & DeAndrea, 2018). These experimental studies relied on a few popular videos, whose prevalence is unknown. Influencers' visual posts can have far-reaching effects on their follower base. In the case of fitness influencers, objectified body images could provoke social media users' anxiety about body image and harm their mental well-being. These sexualized objectified portrayals could also propagate stereotypes about genders (Carrotte et al., 2017; Murashka et al., 2021). Yet, these experimental effects are limited in

generalizability which can be mitigated by analyzing the prevalence of such harmful visual patterns across several influencer accounts and posts.

Furthermore, automated visual analysis can inform the link between macro-level themes and micro-level outcomes. Traditionally, media effects scholars often rely on linkage analysis, which combines content analysis of media messages and measures of media use (using surveys, digital trace data) to investigate how certain media attributes affect outcomes such as beliefs, attitudes, and behaviors (de Vreese & Semetko, 2004; Otto et al., 2022). Many communication theories such as cultivation or selective exposure theory establish a link between broad macro-level themes and individual effects (Krcmar et al., 2016). From a longitudinal visual content analysis of US-American TV programs showing a slight increase in violence depictions coupled with a lack of anti-violence themes, Riddle and Martins (2022) infer, following cultivation theory, that frequent viewers will translate the increased violence to their daily lives. Large-scale analyses of visual content can inform such theoretical propositions by more realistically describing the reality of the information ecology (Wojcieszak et al., 2021) and thus lead to a better understanding of these connections.

A subfield of political communication that has been on the forefront of the adoption of computer vision tools is the analysis of visual politics. Studies have used emotion analysis tools to quantify facial expressions in politicians' photos and videos, and investigate how these facial displays affect audience engagement or how they differ by gender, political ideology, incumbency status, and across social media platforms (Bossetta & Schmøkel, 2023; Haim & Jungblut, 2021; Joo et al., 2019; Peng, 2018, 2021). Work in this area has revealed that visual portrayals of politicians indeed shape outcomes such as audience engagement on social media (e.g., in the forms of virality metrics; Farkas & Bene, 2021; Peng, 2021) and trait perceptions in lab settings (Olivola & Todorov, 2010). Still, much work is needed about how exposure to visual representations of politicians (e.g., in terms of personalization; Farkas & Bene, 2021) on social media shapes viewers' emotions, impressions, voting intention, or readiness to mobilize for protest (Casas & Webb Williams, 2019). Here, automated visual analyses aid because they can observe patterns across cultural context, for instance, to enable generalizations of media effects in national (e.g., Hungary: Farkas & Bene, 2021) vs supranational politics (e.g. European Union), or can quantify patterns in citizens' information diets that likely affect attitudes and voting intention.

To investigate media effects, automated visual analysis should not replace but rather complement other approaches in the content-based social media effects paradigm. For instance, in a linkage analysis, the same logic that applies to text analysis in prior work that links content exposure to media effects also works for visual analysis. This would require researchers to collect visuals, likely in combination with text content, from people's information diets and use survey methods to track their changes in beliefs, attitudes, and behaviors. Thus, patterns, themes, or characteristics of visual media can be studied as antecedents of specific media effects, and can be compared with textual information.

Visual social media production and self-effects

The production and sharing of visual media is a common practice in the domains of political and health communication. The characteristics of users affect how visuals are produced and in turn the produced visual content has self-effects on the producers. The wide accessibility of camera phones allows users to easily take and share photos. Selfies have become an important visual category shared on social media platforms and often signal certain social or political identities (Liu et al., 2017). Self-presentations of politicians on social media have been found to differ based on political orientation or gender (Brands et al., 2021; Xi et al., 2020). Automated visual analyses would allow cross-platform and cultural comparisons of these self-presentations to see whether general trends are observable in the personalization and impression management of politicians worldwide, or over time.

Next to selfies, actors on social media also make and share a large number of visual artifacts such as memes and captioned images. The way producer characteristics shape visual social media content is underexplored. Large-scale analysis of personal or organizational characteristics of accounts in combination with, for instance, climate change memes, can reveal who engages in the discussion to

better understand memes' effects on spreading content, people's risk awareness, or civic engagement on climate change (Kovacheva et al., 2022; Zhang & Pinto, 2021). Bastos et al. (2021) took such an approach when analyzing the visual framing strategy of a state-propaganda organization, revealing polarizing tendencies aimed to appeal to different target audiences.

In addition, while most studies on media effects examine the impact of media content on message recipients, self-effects refer to "the effects of messages on the cognitions (knowledge or beliefs), emotions, attitudes, and behavior of the message creators/senders themselves" (Valkenburg, 2017, p. 2). Micro-level effects of such self-production on political attitudes, body image, and subjective well-being are largely unknown. Automated visual analysis allows analyzing larger quantities of selfies, memes, microclips, and other self-produced visuals (for specific analytical tools, see below). By automating the analysis of these contents, self-effects can be studied across various cultural and topical contexts to understand in how far body images, for instance, are impacted by this cultural practice and might, at a societal level, result in changing body norms. For example, the development of data donation methods (Ohme et al., 2023) allows researchers to access visual materials produced by participants, which can be automatically analyzed and connected to various downstream media effects, such as body image and emotion (measured with surveys or experience sampling methods), or societal trends or norms. Thus, automated visual analysis also allows investigating the associations between micro and macro level effects.

In the domain of political communication, scholars have identified "expression effects:" political expression on social media (e.g., sharing thoughts about a political issue on social media) could serve as an antecedent and gateway to other types of political participation, such as making donations and attending offline protests (Gil de Zúñiga et al., 2014). Since such expressions are also shared visually, e.g., in memes (Beskow et al., 2020), these findings hold important theoretical implications for the political activation potential of citizens via social media and can aid in building or confirming theory.

Sharing visuals such as selfies and group photos on social media likely impacts message senders' interpersonal relationships and subjective well-being. As highlighted in the transformation framework (Nesi et al., 2018), social media has transformed how adolescents view and experience their friendships: for example, individuals may feel compelled to publicly display and validate friendships to others in social networks by sharing visual posts and engaging with others' posts. These "relationship displays" serve multiple social functions, including boosting self-image and providing public "proof" of connections with friends (Manago et al., 2008). The creation and sharing of relationship displays likely affect outcomes such as relationship satisfaction, social support, and well-being, which is an area for future investigation regarding self-effects.

One way to investigate self-effects in visual production is to combine automated visual analysis with data donation methods (Ohme et al., 2023) and linkage analysis. For example, researchers need to gather the visual materials participants have created and shared (e.g., friendship displays on social media). Automated visual analysis can quantify patterns and characteristics of these produced visuals, which are then linked to psychological outcomes such as relationship satisfaction and subjective well-being measured in surveys or experience sampling methods.

The gaps between computer vision tools and theoretical concepts

Despite the many opportunities presented by computer vision tools to examine social media effects, scholars still face some known challenges. Tools for specific object recognition are improving every year, and generalization to visual categories follows suit. But understanding the cultural and social meanings associated with visual categories is essential for achieving a thorough understanding of content.

In an influential review of multimedia retrieval, Smeulders et al. (2000) surveyed how image processing and computer vision tools addressed the problem of automatically categorizing image content and identified as one major challenge the "sensory gap" between the feature description and the semantic interpretation of the image. Much work at the time focused on finding objects in

images, and the appearance of an object under different illumination, pose, and occlusion conditions, as well as dealing with the intrinsic variability of objects. Later, objects were generalized to visual categories, and it became possible to detect the presence of general categories like "police," "flag," and "protest." In the two decades that followed the survey of Smeulders et al. (2000), the field advanced tremendously. Deep neural network models had orders of magnitude more tunable parameters than other popular approaches and were able to learn complex internal representations, provided that they were trained with vast amounts of data. Starting with the ImageNet model (Krizhevsky et al., 2017), which had 60 million parameters to detect a thousand objects, the field has quickly arrived at a point where a short prompt of text could be turned into a plausible image of convincing appearance (Ramesh et al., 2022). What caused this was not only theoretical progress in the design of the machine learning models enabling rich abstractions for high-level concepts, but also the availability of large-scale datasets that could be used in their training.

While computer vision techniques can accurately recognize visual categories, the social and cultural implications associated with these categories in visuals pose another layer of analytical challenge (Zellers et al., 2019). Some objects (such as flags and guns) can be important visual symbols in contexts such as political communication and some of these categories (such as violence) have been adopted in existing content analysis in social scientific research (Joo & Steinert-Threlkeld, 2022; Joo et al., 2019; Zhang & Pan, 2019), but some symbols are more subtle, and concern relationships and contextual effects. For instance, it is unlikely for an automatic system to notice the subtext embedded in the positioning of people in a room (e.g. politicians around a table), unless a specific rule is created for the interpretation of such content. D'Errico and Poggi (2019) developed a system to detect humility in a politician's stance by analysis of facial and vocal cues, but such systems require data collection and annotation for each target variable.

To scholars who wish to adopt computer vision tools to analyze media effects, many computer vision tools and algorithms are not designed to measure theoretical concepts such as visual framing, but instead to perform tasks for other purposes such as facial recognition, object recognition, and text detection. The gap between computer vision outputs and theoretical concepts relevant to social media effects presents a unique analytical challenge for communication science. Scholars who have predefined visual concepts to measure, such as the presence of collective actions in social media images (Steinert-Threlkeld et al., 2022; Zhang & Pan, 2019), can proceed to collect human annotations and train prediction models, ensuring a direct connection between theoretical concepts and computational measures. Yet, when using results from existing computer vision algorithms, researchers have to figure out how these measures reflect theoretical concepts and sometimes need to creatively adapt these measures. For instance, prior research has used motion detection to capture the extent of physical crossover amomg members of Congress across the aisle, which may indicate political polarization (Dietrich, 2021), underscoring the importance of contextualizing computer vision analytical techniques in social science inquiries.

Analytical approaches to advancing media effects research with automated visual analysis

We now introduce major analytical approaches that could be applied in a content-based social media effects paradigm, paying special attention to the potential of each computer vision technique in measuring visual attributes that are theoretically meaningful. We cover three approaches to automated visual analysis (Figure 1) directly adopting outputs from existing computer vision application programming interfaces (API) and libraries, (2) applying supervised learning to measure a visual attribute, and (3) applying unsupervised learning to discover visual categories or themes. We also highlight each approach's advantages and limitations in Table 2.

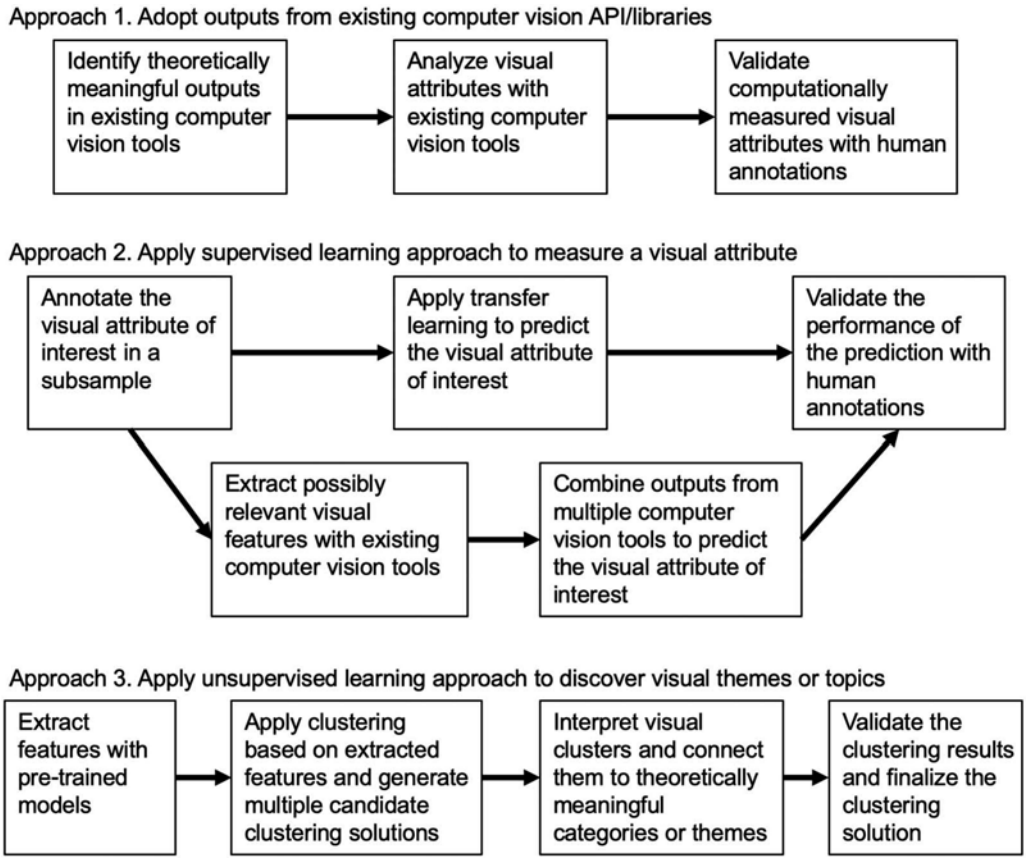

Figure 1. Workflow of the three proposed approaches.

Directly adopting outputs from existing computer vision APIs and libraries

First, to incorporate computer vision outputs in social scientific inquiries, we can directly justify the theoretical significance of certain computer vision outputs. This would require researchers both to know what visual features are available from existing computer vision tools and to identify the ones that can be theoretically significant to media effects research. Today, many publicly available computer vision APIs and libraries can perform a range of standard tasks. Figure 2 showcases various computer vision techniques that could be employed by communication scholars (also see Table S1). These tasks, however, are not intended to measure communication concepts or to advance theoretical inquiries. Still, some of the computer vision outputs may have theoretical significance, as they can indicate communicative patterns in specific contexts or be linked to media effects outcomes such as impression formation.

In political communication, the way politicians are visually presented can influence how viewers perceive them. Visual framing can highlight specific characteristics of politicians, such as ordinariness and compassion, and make them more noticeable to audiences (Grabe & Bucy, 2009). Multiple computer vision tools can analyze visual representations of politicians on social media. One popular technique is to use emotion recognition tools to quantify the facial expressions of politicians (Bossetta & Schmøkel, 2023; Boussalis & Coan, 2021; Haim & Jungblut, 2021; Jungblut & Haim, 2021; Peng, 2018). Emotional expressions displayed by politicians can communicate certain traits, such as friendliness or aggressiveness (Grabe & Bucy, 2009; Haim & Jungblut, 2021; Peng, 2018). For example, displaying happiness can make a politician appear friendly and approachable, and can signal a sense of

Table 2. A comparison of approaches of automated visual analysis regarding purposes, advantages, and limitations.

	Commercial APIs	Open-source libraries	Customized supervised learning	Customized unsupervised learning
Purpose	Researchers perform common tasks and link results to relevant theoretical concepts.	Researchers perform common tasks and link results to relevant theoretical concepts.	Researchers train models to predict visual concepts or attributes that interest them.	Researchers discover potential visual categories, topics, or themes in a dataset.
Flexibility	Restricted by the services provided.	A greater variety of computer vision tasks are available.	Can be flexibly customized to measure certain visual attributes.	Allow exploration of the dataset without predefined categories or attributes.
Technical expertise	Low	Moderate	Moderate to high	Moderate to high
Cost	Relatively inexpensive but may be expensive if applied to a very large dataset. Require human validation.	Typically freely available. Require human validation.	Require the collection of task-specific data and human annotations.	Require minimal human annotations upfront but necessitate human validation afterward.
Replication	Algorithms remain a black box. Difficult to replicate analysis and compare results over time.	Relatively easy to replicate analysis if training datasets, scripts, and models are shared.	Relatively easy to replicate analysis if training datasets, scripts, and models are shared.	Relatively easy to replicate analysis if training datasets, scripts, and models are shared.
Privacy	May violate the privacy of participants when researchers upload data to API providers.	Depends on how researchers collect and analyze participantdata.	Depends on how researchers collect and analyze participant data.	Depends on how researchers collect and analyze participant data.
Bias	May contain demographic or cultural biases.	May contain demographic or cultural biases. Biases may be mitigated by package developers.	May contain demographic or cultural biases. Biases may be mitigated by researchers.	May contain demographic or cultural biases. Biases may be mitigated by researchers.

Note. While our proposed workflow considers adopting computer vision APIs and open-source libraries as the same approach, this table presents them separately to highlight their distinctions in aspects such as privacy and replication.

popularity and ordinariness. The photo posted by Amy Klobuchar, a U.S. senator of the Democratic party, in which she is seen smiling while at a bar, is such an example (Figure 3a). The display of emotional expressions is also highly gendered, with women politicians expected to show more happiness and less anger than men (Boussalis & Coan, 2021; Jungblut & Haim, 2021; Peng, 2018). Therefore, researchers can quantify the emotional expressions of politicians on social media and study their effects on outcomes such as perceived traits of politicians, attitudes toward politicians of different genders, and voting intention.

Object recognition tools can provide content tags that describe common objects or settings in images Figure 3(c-f). In visual politics, the presence of certain objects may have significant meaning. For instance, researchers have observed a trend toward personalization of politics, where politicians focus on presenting their private lives on social media to build more intimate connections with their followers (Peng, 2021). This approach also involves the emphasis of certain personality traits, such as compassion and relatability. Politicians often use visual cues, such as appearing with babies or pets, to convey these traits (Grabe & Bucy, 2009). In Figure 3(c), Thom Tillis, a U.S. senator in the Republican party, is seen holding a baby, and in Figure 3(d), Nikki Haley, a former U.S. ambassador to the United Nations, is shown with her dog. Content tags like "baby," "toddler," and "dog" may indicate a personalized communication strategy. On the other hand, politicians may also use visual symbols to project a sense of professionalism and leadership. In Figure 3(e), the U.S. Vice President Kamala Harris is seen addressing a group of supporters in front of the U.S. flag. Content tags like "flag," "group," and "crowd" may convey a sense of professionalism in politics, as well as her broad appeal to voters.

Also, objects and settings presented in the visual portrayals of politicians may serve as visual cues that subtly convey ideology, which in turn shape how viewers judge politicians' ideology and voting

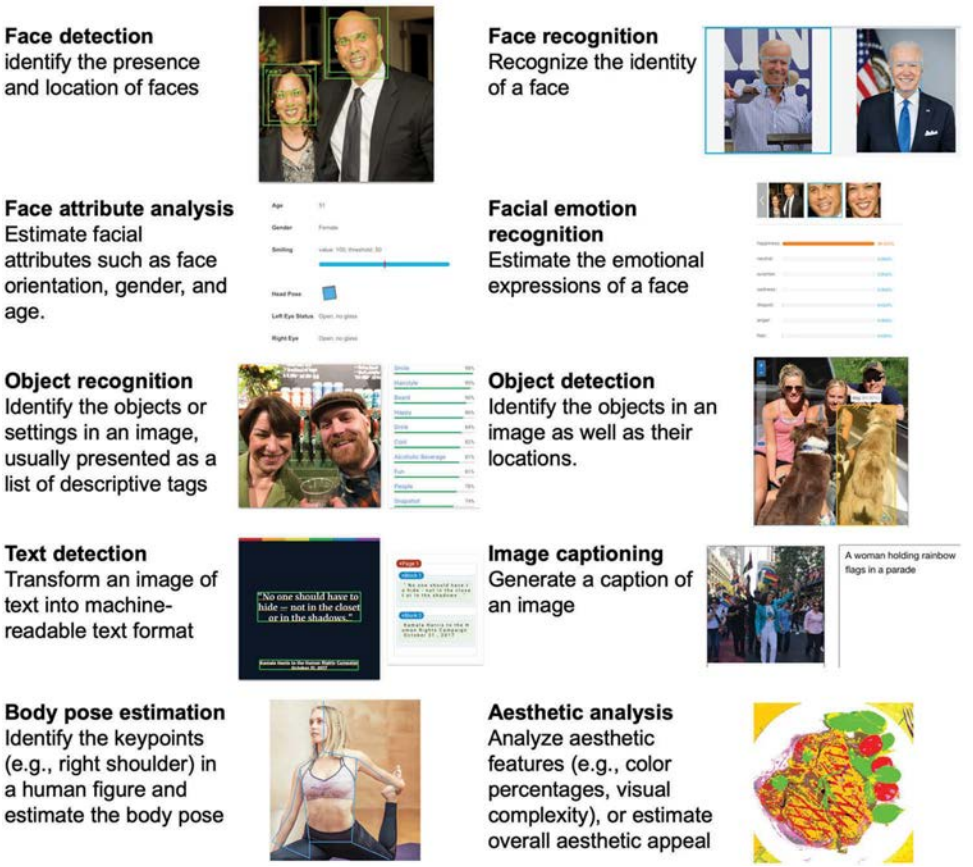

Figure 2. Common computer vision tasks available in existing API or libraries. High-definition version is available at: https://osf.io/8s4jv/. Links to computer vision APIs are provided in table S1.

intention (Dan & Arendt, 2021; Hiaeshutter-Rice et al., 2021). In U.S. politics, Republicans are associated with objects like guns, hunting, and churches, while Democrats are linked to things like veganism, piercings, and lattes (Hiaeshutter-Rice et al., 2021). In Figure 3(f), Lisa Murkowski, a U.S. senator in the Republican party, is pictured with a hunting gun. Content tags such as "gun" and "hunting" may signal her ideological commitments and issue positions related to gun control. Finally, politicians can directly convey textual messages in images to communicate their ideological commitments and issue positions. In prior research, a substantial proportion of politicians' Instagram posts include text overlay (Peng, 2021), which can be extracted using text detection tools and analyzed further with automated text analysis tools Figure 3(b).

The meaning of computer vision outputs can vary depending on the context and researchers must apply their domain expertise to determine the theoretical relevance of computer vision outputs across different fields. For example, in the study of digital connections, computer vision techniques may aid in elucidating the self-effects of friendship displays. While previous studies have established a link between social media usage and friendship closeness, the underlying mechanisms remain an area of ongoing research (Pouwels et al., 2021). To advance this line of inquiry, researchers could use data donation methods to access visuals shared on participants' social media feeds and survey psychological outcomes related to relationship closeness. Researchers can then employ facial detection and recognition tools to identify selfies and group photos and quantify the frequency of friendship displays involving specific individuals. Emotion recognition tools could measure the emotional display of people in friendship displays, providing further insight into the quality of friendships.

Figure 3. Existing computer vision tools applied to the analysis of U.S. politicians' visual posts on instagram: a. Facial attribute analysis, including emotion recognition (Google Vision); b. Text detection (Google Vision); c, d, e, f. Object recognition (Microsoft Azure). High-definition version is available at: https://osf.io/epx3v. Links to computer vision APIs are provided in table S1.

Validation is a critical step in adopting existing computer vision APIs or libraries. In such cases, researchers often choose a random subset of their data and involve human coders (e.g., crowdsourced workers) to annotate visual features. These annotations are then compared to the outputs generated by these off-the-shelf tools. By treating human annotations as the ground truth, researchers can assess tool performance using metrics like correlation, precision, recall, F1 score, and confusion matrix. For example, prior research has applied emotion recognition tools from exiting commercial APIs to study the emotional expressions of politicians. While these computer vision APIs generally performed well in terms of recognizing happiness and anger, they did not well identify other emotions (Peng, 2018). When specific tools are lacking or inaccurate for a particular task, researchers consider annotating some of their own data in order to adopt a supervised approach, such as fine-tuning a pre-trained model (see below).

Automated visual analysis may present some challenges to open science practices. Many computer vision APIs are developed by commercial entities, and their training datasets, algorithms, and procedures often remain unclear to communication researchers. Furthermore, these companies may update their algorithms without providing transparent information on the changes made, making it difficult for researchers to ensure the replicability of their analyses, as analyses based on the output of these commercial services may differ when replicated at a later point in time. At the very least, researchers need to specify the time frame and API version of their analysis and demonstrate how well these APIs perform on independent validation datasets, which can be compared across different time points. For replication, open-source solutions are preferable. For example, on GitHub one can find many packages for tasks like face detection and analysis. Researchers can replicate a given study by using the same open-source package (and version) originally used by the authors, especially when replication code is provided upon publication.

Applying supervised learning approach to measure theoretical concepts

Scholars sometimes investigate how a particular visual attribute or category is linked to media effects. For instance, they may examine how the level of violence in protest images on social media mobilizes future public participation (Steinert-Threlkeld et al., 2022) or how the degree of sexual objectification in fitspiration images shapes viewers' body image and exercise intention. To measure these characteristics, a supervised approach is appropriate. Such a task would involve collecting a subset of data with validated human annotations and training a supervised prediction model, such as a convolutional neural network (CNN) (Joo & Steinert-Threlkeld, 2022; Webb Williams et al., 2020). Typically training a neural network from scratch needs a very large amount of data. Since it is not possible for social scientists to annotate millions of images manually (which would defeat the purpose of automated analysis), there are multiple ways to address this challenge.

Transfer learning

First, researchers can adopt transfer learning techniques such as fine-tuning a pre-trained model, i.e., repurposing a model that is trained on a related task by keeping most of it unchanged and fine-tuning a portion of it on the new task (Webb Williams et al., 2020). This approach was made popular by Chatfield et al. (2014), which used a model pre-trained on 1.2 million images for a recognition task, and re-purposed it for new problems with only thousands of annotated training samples. Webb Williams et al. (2020) provides a general pipeline for utilizing transfer learning in supervised learning for a specialized task in social scientific research. After selecting a baseline neural network model trained on a related task, researchers would annotate a small number of samples (typically in the order of hundreds to thousands per category) with multiple annotators (to ensure reliability), and use the subsample to train the last few layers of the neural network. Since CNNs acquire a coarse-to-fine feature representation, the early layers contain generic representations, similar to the human visual system, which allows them to be used for other visual tasks.

There is no fixed rule to determine the number of images needed for human annotations in transfer learning. Generally, the accuracy of prediction models increases with the amount of training data. Also, if the dataset and visual labels researchers are working with are similar to the content categories in the pre-trained models, this may require fewer resources. In prior research, the number of annotations varies from a few hundred (Araujo et al., 2020; Webb Williams et al., 2020) to tens of thousands (Steinert-Threlkeld et al., 2022; Zhang & Pan, 2019). For example, Steinert-Threlkeld et al. (2022) trained a CNN to predict the presence of protests and different types of protest images (e.g., state violence) with about 40,000 annotated images. In Araujo et al. (2020), three image categories were selected and about 1,000 images were annotated manually.

Combine computer vision model outputs

A second approach to repurpose existing computer vision tools is to incorporate outputs from existing computer vision models and train a prediction model using supervised learning algorithms (Araujo et al., 2020; Beskow et al., 2020; Joo et al., 2014). For example, to predict the traits politicians aim to convey in images (e.g., trustworthy), Joo et al. (2014) used computer vision tools to measure visual characteristics including facial displays, body gestures, and scene contexts, and then built a prediction model combining these outputs. Beskow et al. (2020) proposed "Meme Hunter," which incorporated results from facial analysis and text detection in a neural network that predicts different types of internet memes. This approach also applies to video analysis. Joo et al. (2019) utilized computer vision libraries to extract facial and body features, which were then integrated into a neural network to predict various nonverbal behaviors of presidential candidates in video footage of debates.

To adopt such an approach would require researchers to understand (1) which visual characteristics are available in existing computer vision models and (2) to identify visual characteristics that should pertain to the concept they are aiming to measure, thus translating an abstract theoretical concept into concrete and computationally measurable visual features. Here, we use the measurement of fitspiration images as one example. Fitspiration is an increasingly popular trend on social media such as Instagram, which promotes physical health fitness by showing inspiring examples, but it might also lead to more social comparison and heightened body anxiety (Cataldo et al., 2021; Easton et al., 2018; Murashka et al., 2021). Importantly, fitspiration content may feature visual characteristics of objectification (Tiggemann & Zaccardo, 2018), i.e., "degrading someone to the status of merely a body that is mainly valued for its use to and pleasure of others" (Fredrickson & Roberts, 1997).

To computationally measure fitspiration content and its characteristics (e.g., level of objectification), researchers first manually annotate a subset of images and train a supervised learning model that incorporates outputs from multiple computer vision tools (Figure 4). First, facial detection and person detection models detect the presence of faces and human figures, as well as their relative size, which could be corresponding to objectification features (e.g., a photo featuring a person's body without clearly detected faces could be an objectified portrayal). In addition, body skeleton analysis can recognize the keypoints in a detected human figure, which can be adopted to either train a classifier about different physical activities (e.g., running) or poses (e.g., a sexy pose), or fed into a clustering algorithm to extract major body gestures in this set of images (Joo et al., 2014). Object recognition can recognize specific objects and scene categories, such as food and gyms, which are related to the fitness context. Alternatively, researchers can extract features with a pre-trained model trained on scene categorization and object classification. Text detection can recognize the quotes in an image, which can then be used to determine whether they contain fitness-related words. As a result, these visual features from different computer vision tools can be combined to build a prediction model that measures fitspiration-related concepts.

Validating supervised learning models is essential for ensuring their accuracy and reliability. One challenge of training these models is the vast array of methodological choices available. There are multiple supervised learning algorithms like support vector machines and neural networks, each with numerous parameters that need adjusting. To address this issue, the basic machine learning approach requires researchers to split their annotated dataset into three parts: a training set for building supervised learning models, a validation set for comparing methodological choices and fine-tuning these models, and a test set for independently assessing the performance of the selected model (Araujo et al., 2020). The validation set enables researchers to experiment with various parameters and compare model performance. Once a prediction model with optimal performance is identified, the test set is used to assess its accuracy. If multiple classifiers are compared to each other, a statistical test could be adopted to ensure a healthy comparison, and repeated performance measurement on the test set should be avoided (Alpaydin, 2020).

Figure 4. Integrating outputs from multiple computer vision techniques (Face++, Google vision) to predict visual attributes of theoretical interest, using fitspiration images as an example. High definition version is available: https://osf.io/vxbh3. Links to computer vision APIs are provided in table S1.

Applying unsupervised learning to discover visual topics and themes

When scholars do not have pre-defined visual attributes (e.g. objectification) to investigate, it is appropriate to adopt an unsupervised approach to discover potential visual topics or themes and connect them to media effects. For example, research can investigate how different types of smartphone activities affect individuals' sense of belonging and relationship satisfaction, or how different communication strategies from politicians affect viewers' attitudes and voting intention. An unsupervised approach, such as image clustering, groups images into visually similar categories (Zhang & Peng, 2021). One image clustering method that has received attention in communication research is based on transfer learning (in particular, feature extraction with pre-trained models) (Mooseder et al., 2023; Muise et al., 2022; Peng, 2021). In this procedure, a pre-trained model is used to convert images into embeddings or vectors of features, which can then be clustered automatically, using for instance the popular k-means algorithm. Unsupervised image clustering does not automatically assign descriptive labels to clusters of images, meaning that researchers need to interpret these clusters in a post-hoc manner and to determine how they relate to meaningful categories and possible theoretical concepts. Interpreting visual clusters therefore requires a background knowledge and understanding of how these visual categories can be theoretically relevant.

There have been multiple studies that use image clustering to generate meaningful visual categories in social scientific research. For example, Muise et al. (2022) employed a Screenomics framework to capture screenshots from participant smartphones, amassing a vast amount of visual

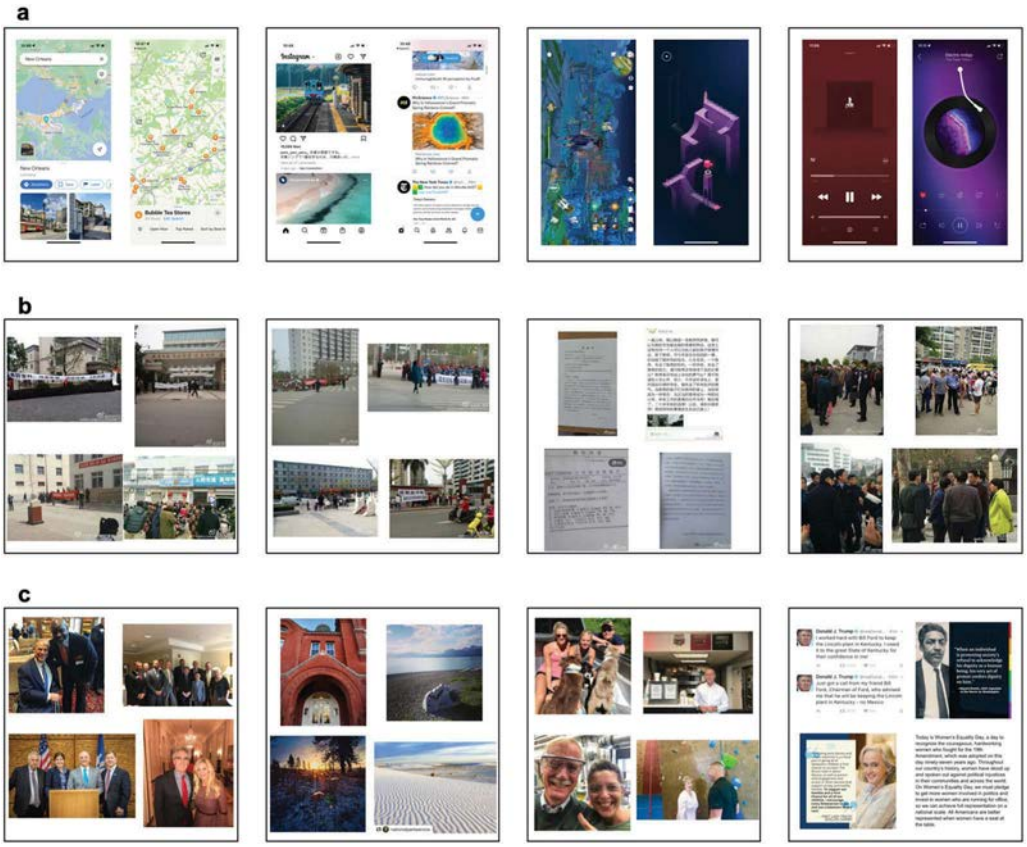

Figure 5. Image clustering results for (a) participants' mobile phone screenshots, (b) protest-related images on Weibo, and (c) Instagram posts from U.S. politicians, with visually similar images being grouped into clusters. Only four clusters from each case are displayed for brevity. Please note that the authors generated the images in the first case for illustrative purposes. High definition version is available at https://osf.lo/kvp67. A guide to image clustering is available at https://github.com/yilangpeng/transfer-learning-clustering.

data. With image clustering, they categorized smartphone usage into experiential categories, such as social media usage, gaming, and using maps (Figure 5a). These activities can be compared across demographic and cultural groups, as well as linked to psychological outcomes such as well-being. Zhang and Peng (2021) studied protest-related images on Weibo and identified various protest tactics employed by demonstrators using image clustering. The clusters revealed tactics such as disruptive strategies used to attract public attention (e.g., blocking streets) and handwritten or printed petition letters used to bypass censorship algorithms (Figure 5b). These categories can be further used to evaluate the effectiveness of different protest tactics in attracting online attention and mobilizing the public.

When adopting image clustering, scholars should choose a pre-trained model suited to their dataset and carefully validate the clustering solution. The pre-trained model will result in a specific feature representation, thus shaping resulting visual categories. For instance, Peng (2021) analyzed politicians' Instagram posts using a model trained on the Places365 dataset, which reflects visual patterns related to scenes. This model leads to the discovery of clusters associated with scenes and settings, which can reflect the concept of personalization (Figure 5c). In addition, the selected number of clusters can impact the granularity of the results, and too many clusters can lead to visually indistinguishable clusters. One approach is to run several candidate solutions, varying the number of clusters within a range, and use human

interpretation and validation to determine a good clustering solution (Zhang & Peng, 2021). Model selection approaches in machine learning (e.g. Akaike Information Criterion) can also guide this process (Alpaydin, 2020).

It is important to recognize the exploratory and unsupervised nature of image clustering methods. This method seeks to identify visually similar images, but may not necessarily result in semantically coherent clusters. To better connect computationally generated clusters to theoretically coherent categories, there are several approaches one can take. For instance, one approach is to generate more granular clusters and then sort them into larger theoretically meaningful categories through human scrutiny (Mooseder et al., 2023; Peng, 2021). Alternatively, one can use image clustering as a preliminary step to identify new visual topics and treat them as new pre-defined categories in a supervised approach. Finally, it is worth considering a multimodal clustering approach that leverages both visual information and contextual data, such as image captions, to more accurately identify contextually meaningful categories.

Challenges and future directions of automated visual analysis

Our review highlights several approaches to integrating computer vision methods into the study of media effects on social media. Each approach offers distinct advantages and limitations (Table 2), and can be tailored to meet specific research objectives. Given the enormous volume of visual content on social media and the rapid development of computer vision tools, automated visual analysis is expected to become increasingly crucial for advancing media effects scholarship in the near future.

Ethical and methodological challenges of automated visual analysis

We present an overview of each approach's possible advantages and limitations in Table 2. First, computer vision programs such as facial detection and facial recognition could contain biases across different demographic groups, demonstrating more erroneous predictions for minority groups (Buolamwini & Gebru, 2018; Phillips et al., 2011). Additionally, computer vision algorithms may learn stereotypical associations between demographic groups and visual attributes. For example, object recognition algorithms may be more inclined to assign certain content tags, such as those related to appearance, to women over men (Schwemmer et al., 2020). The bias in computer vision tools is a matter of concern as it can impede social scientists from making accurate and valid claims while also perpetuating harmful stereotypes.

Social and cultural biases may enter training datasets and computer vision algorithms. One study evaluates the performance of object recognition algorithms on images across cultures: algorithms generally performed worse at recognizing objects in photos from lower income households and certain less developed countries (De Vries et al., 2019). This could be partially explained by the fact that images in large benchmark datasets often come from developed countries and English is often the default language used for data collection (De Vries et al., 2019). Such a gap could pose some challenges for scholars who work on visuals from diverse cultures, for example, object recognition tools may face difficulty in recognizing culturally specific settings such as weddings. Therefore, especially for researchers who work with visual data from non-Western contexts, evaluating the performance of computer vision algorithms is necessary and utilizing supervised learning approaches instead of relying on off-the-shelf tools may be recommended. Additionally, estimating and mitigating cultural biases for computer vision tools should be a major focus in future research programs.

Visual content on social media is particularly sensitive due to its potential inclusion of personal markers, such as faces, and interconnectivity, as a single photo can be linked to a person's location, social contacts, and other personal information. However, some commercial computer vision APIs may keep uploaded data to improve their algorithms, potentially compromising participants' privacy. Especially when working with personal data from study participants, it is crucial to carefully review the terms of service of these commercial options and evaluate the potential privacy risks. Researchers

should consider using open-source tools or creating their own prediction models to ensure that participant data are protected.

To ensure that computer vision research has a positive social impact, researchers must consider how their study affects individuals from different communities. Inferring individual differences through computer vision techniques could lead to social harm, as demonstrated by studies attempting to predict sexual orientation or criminality based on facial photographs. These efforts raise significant ethical concerns, as they may result in inaccurate predictions that harm marginalized groups and reinforce problematic notions that certain attributes are biologically determined and immutable (Agüera y Arcas et al., 2017; Crawford, 2021). Moreover, doubts have also been cast on whether facial expressions accurately reflect an individual's internal emotions. Therefore, equating people's facial emotional expressions with their emotional states through computer vision applications could misclassify individuals and lead to discrimination (Crawford, 2021). Furthermore, it is important to recognize that classification performed by computer vision programs can oversimplify complex human experiences and social realities. For instance, many facial gender analysis programs classify gender in a binary manner, disregarding the full range of gender identities and expressions and potentially perpetuating the marginalization of transgender communities (Keyes, 2018).

Future directions for automated visual analysis in media effects research

To improve the study of media effects, it would be valuable to develop benchmark datasets that better reflect theoretically meaningful categories in visuals. While existing benchmark datasets like ImageNet provide vast resources for computer vision researchers to train more accurate models, they are often annotated for object categories that are not directly relevant to theoretical concepts of interest in communication. Developing benchmarks for visual analysis may be more feasible than for text analysis, as visual media are less language-specific and some visual attributes, such as facial expressions, may be less time-sensitive or context-dependent than textual sentiment analysis. Still, cultural biases in visuals are still an important issue, and a large collaborative effort involving scholars from diverse backgrounds may be necessary to ensure that biases are accounted for. Such benchmark datasets would include annotated images related to various communication domains. For visual politics, images could be annotated for policy topics, issue frames, and political symbols. Similarly, for body image research, visual media could be labeled for different problematic categories, such as the level and type of objectification. This direction offers several analytical advantages. For example, computer vision models pre-trained on these domain-relevant benchmark datasets (potentially on top of other existing datasets such as ImageNet) are likely to produce more accurate results when leveraged for supervised learning tasks relevant to social scientists. Similarly, unsupervised clustering on such benchmark datasets can enable new descriptive analyses. Current techniques use image embeddings to measure image similarity and to find cohesive image clusters (Zhang & Peng, 2022). A model trained on data and tasks that are relevant for social scientists should be able to generate more theoretically meaningful clusters.

In addition, large language models (LLM), through their rich abstraction capabilities, may contribute to media effects research by advancing the semantic understanding of visuals. These models, such as OpenAI's GPT-4,[1] incorporate billions of parameters, and store a large number of facts and general information about the world, with which they can reason in a wide variety of tasks, and generate outputs based on simple descriptions. They accept both image and text as input and emit text outputs, which can further feed text-based image generation models like Midjourney.[2] Their capabilities of extracting high-level semantics from images and text will become very useful for large scale analysis. LLMs also provide high flexibility in concept measurement: for example, existing object recognition tools typically assign predetermined labels to a given image, but with GPT-4, researchers

[1]https://openai.com/research/gpt-4
[2]https://www.midjourney.com/home/

can define their visual labels and use customized text prompts to analyze an image. They can be used to create realistic visual depictions of a certain category of interest like "political demonstration," and help researchers build customized machine learning models for recognizing such content. However, the development of LLMs is computationally costly and may pose negative environmental impacts (Bender et al., 2021). These models may also interpolate and generate fabricated outcomes, making it challenging to accurately assess their outputs. While acknowledging the promise of LLMs in visual analysis, we believe that practices such as validating prediction, mapping theoretical relevance, and identifying biases remain critical.

Multimodal analysis, which combines text, visual, and audio components, presents promising avenues for future research. Integrating information from multiple channels can shed light on the complex interplay among distinct channels and the temporal dynamics of information, which provides insights that visual channels alone cannot reveal. For example, researchers have used features from different channels, including visual, audio, and text, to predict theoretical concepts such as aggression in political debates (Shah et al., 2023) or to categorize videos using unsupervised learning techniques (Lu & Shen, 2023). Furthermore, co-learning approaches (Baltrušaitis et al., 2018) explore how information in one modality can be used to train computational models in a different modality. The improvements in text analysis models will help with improving visual and audio-visual models, because they will provide a way to deal with annotation scarcity in these modalities. Furthermore, learning multimodal representations allows new paradigms of accessing visual content, such as visual question answering (VQA), which seeks to build systems that can take an image and a natural language question about the image to provide accurate natural language answers (Antol et al., 2015).

Conclusion

In conclusion, it is crucial to integrate the analysis of visual media into the content-based social media effects paradigm. Visual analysis can help quantify patterns and themes of visual content on social media, elucidate relationships between exposure to visual media and outcomes, and understand the self-effects of visual production among social media users. We identify multiple approaches to incorporate computer vision tools in studying media effects and hope this article will stimulate further research interest and collaboration in this dynamic field.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Agüera y Arcas, A., Mitchell, M., & Todorov, A. (2017). *Physiognomy's new clothes*. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539. 10.1126/sciadv.aay3539

Alpaydin, E. (2020). *Introduction to machine learning*. MIT Press.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). *VQA: Visual question answering*. Proceedings of the IEEE international conference on computer vision. 2425–2433.

Araujo, T., Lock, I., & van de Velde, B. (2020). Automated visual content analysis (AVCA) in communication research: A protocol for large scale image classification with pre-trained computer vision models. *Communication Methods and Measures*, *14*(4), 239–265. 10.1080/19312458.2020.1810648

Aubrey, J. S., & Frisby, C. M. (2011). Sexual objectification in music videos: A content analysis comparing gender and genre. *Mass Communication and Society*, *14*(4), 475–501. 10.1080/15205436.2010.513468

Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. 10.1109/TPAMI.2018.2798607

Bastos, M., Mercea, D., & Goveia, F. (2021). Guy next door and implausibly attractive young women: The visual frames of social media propaganda. *New Media & Society*, Advance online publication. 10.1177/14614448211026580

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?*. Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 610–623.

Beskow, D. M., Kumar, S., & Carley, K. M. (2020). The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, *57*(2), 102170. 10.1016/j.ipm.2019.102170

Bossetta, M., & Schmøkel, R. (2023). Cross-platform emotions and audience engagement in social media political campaigning: Comparing candidates' Facebook and Instagram images in the 2020 US election. *Political Communication*, *40*(1), 48–68. 10.1080/10584609.2022.2128949

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. 10.1080/10584609.2020.1784327

Brands, C., Kruikemeier, S., & Trilling, D. (2021). Insta (nt) famous? Visual self-presentation and the use of masculine and feminine issues by female politicians on instagram. *Information, Communication & Society*, *24*(14), 2016–2036. 10.1080/1369118X.2021.1962942

Brown, Z., & Tiggemann, M. (2016). Attractive celebrity and peer images on instagram: Effect on women's mood and body image. *Body Image*, *19*, 37–43. 10.1016/j.bodyim.2016.08.007

Buolamwini, J., & Gebru, T. (2018). *Gender shades: Intersectional accuracy disparities in commercial gender classification*. Conference on fairness, accountability and transparency. 77–91. PMLR.

Carrotte, E. R., Prichard, I., & Lim, M. S. C. (2017). "Fitspiration" on social media: A content analysis of gendered images. *Journal of Medical Internet Research*, *19*(3), e6368. 10.2196/jmir.6368

Casas, A., & Webb Williams, N. (2019). Images that matter: Online protests and the mobilizing role of pictures. *Political Research Quarterly*, *72*(2), 360–375. 10.1177/1065912918786805

Cataldo, I., De Luca, I., Giorgetti, V., Cicconcelli, D., Bersani, F. S., Imperatori, C., Abdi, S., Negri, A., Esposito, G., & Corazza, O. (2021). Fitspiration on social media: Body-image and other psychopathological risks among young adults. A narrative review. *Emerging Trends in Drugs, Addictions, and Health*, *1*, 100010. 10.1016/j.etdah.2021.100010

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Return of the devil in the details: Delving deep into convolutional nets*. https://arxiv.org/abs/1405.3531

Chen, K., Kim, S. J., Gao, Q., & Raschka, S. (2022). Visual framing of science conspiracy videos: Integrating machine learning with communication theories to study the use of color and brightness. *Computational Communication Research*, *4*(1). 10.5117/CCR2022.1.003.CHEN

Chen, Y., Sherren, K., Smit, M., & Lee, K. Y. (2021). Using social media images as data in social science research. *New Media & Society*, *25*(4), 849–871. 10.1177/14614448211038761

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dan, V., & Arendt, F. (2021). Visual cues to the hidden agenda: Investigating the effects of ideology-related visual subtle backdrop cues in political communication. *The International Journal of Press/politics*, *26*(1), 22–45. 10.1177/ 1940161220936593

de Vreese, C. H. D., & Semetko, H. A. (2004). News matters: Influences on the vote in the Danish 2000 euro referendum campaign. *European Journal of Political Research*, *43*(5), 699–722. 10.1111/j.0304-4130.2004.00171.x

De Vries, T., Misra, I., Wang, C., & Van der Maaten, L. (2019). *Does object recognition work for everyone?* Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 52–59.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *Imagenet: A large-scale hierarchical image database*. 2009 IEEE conference on computer vision and pattern recognition. 248–255. IEEE.

Dietrich, B. J. (2021). Using motion detection to measure social polarization in the US house of representatives. *Political Analysis*, *29*(2), 250–259. 10.1017/pan.2020.25

Easton, S., Morton, K., Tappy, Z., Francis, D., & Dennison, L. (2018). Young people's experiences of viewing the fitspiration social media trend: Qualitative study. *Journal of Medical Internet Research*, *20*(6), e9156. 10.2196/jmir.9156

Farkas, X., & Bene, M. (2021). Images, politicians, and social media: Patterns and effects of politicians' image-based political communication strategies on social media. *The International Journal of Press/politics*, *26*(1), 119–142. 10.1177/1940161220959553

Fredrickson, B. L., & Roberts, T.-A. (1997). Objectification theory: Toward understanding women's lived experiences and mental health risks. *Psychology of Women Quarterly*, *21*(2), 173–206. 10.1111/j.1471-6402.1997.tb00108.x

Gil de Zúñiga, H., Molyneux, L., & Zheng, P. (2014). Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. *Journal of Communication*, *64*(4), 612–634. 10.1111/jcom.12103

Grabe, M. E., & Bucy, E. P. (2009). *Image bite politics: News and the visual framing of elections*. Oxford University Press.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374–378. 10.1126/science.aau2706

Haim, M., & Jungblut, M. (2021). Politicians' self-depiction and their news portrayal: Evidence from 28 countries using visual computational analysis. *Political Communication*, *38*(1–2), 55–74. 10.1080/10584609.2020.1753869

Hiaeshutter-Rice, D., Neuner, F. G., & Soroka, S. (2021). Cued by culture: Political imagery and partisan evaluations. *Political Behavior*, Advance online publication. 10.1007/s11109-021-09726-6

Hornik, R., Binns, S., Emery, S., Epstein, V. M., Jeong, M., Kim, K., Kim, Y., Kranzler, E. C., Jesch, E., Lee, S. J., Levin, A. V., Liu, J., O'Donnell, M. B., Siegel, L., Tran, H., Williams, S., Yang, Q., & Gibson, L. A. (2022). The effects of tobacco coverage in the public communication environment on young people's decisions to smoke combustible cigarettes. *Journal of Communication*, *72*(2), 187–213. 10.1093/joc/jqab052

Houts, P. S., Doak, C. C., Doak, L. G., & Loscalzo, M. J. (2006). The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling*, *61*(2), 173–190. 10.1016/j.pec.2005.05.004

Joo, J., Bucy, E. P., & Seidel, C. (2019). Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal of Communication*, *13*.

Joo, J., Li, W., Steen, F. F., & Zhu, S. C. (2014). *Visual persuasion: Inferring communicative intents of images. Proceedings of the IEEE conference on computer vision and pattern recognition*. 216–223.

Joo, J., & Steinert-Threlkeld, Z. C. (2022). Image as data: Automated content analysis for visual presentations of political actors and events. *Computational Communication Research*, *4*(1). 10.5117/CCR2022.1.001.JOO

Jungblut, M., & Haim, M. (2021). Visual gender stereotyping in campaign communication: Evidence on female and male candidate imagery in 28 countries. *Communication Research*, Advance online publication. 10.1177/00936502211023333

Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22. 10.1145/3274357

Kleemans, M., Daalmans, S., Carbaat, I., & Anschütz, D. (2018). Picture perfect: The direct effect of manipulated instagram photos on body image in adolescent girls. *Media Psychology*, *21*(1), 93–110. 10.1080/15213269.2016.1257392

Krcmar, M., Ewoldsen, D. R., & Koerner, A. (2016). *Communication science theory and research: An advanced introduction*. Routledge.

Kress, G., & Van Leeuwen, T. (1998). Front pages: (the critical) analysis of newspaper layout. In A. Bell & P. Garrett (Eds.), *Approaches to media discourse* (pp. 186–219). Blackwell.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. 10.1145/3065386

Liu, F., Ford, D., Parnin, C., & Dabbish, L. (2017). Selfies as social movements: Influences on participation and perceived impact on stereotypes. *Proceedings of the ACM on Human-Computer Interaction 1*, 1–21.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), 74–101. 10.1038/s41562-022-01460-1

Lu, Y., & Shen, C. (2023). Unpacking multimodal fact-checking: Features and engagement of fact-checking videos on Chinese TikTok (Douyin). *Social Media+ Society*, *9*(1), 205630512211504. 10.1177/20563051221150406

Manago, A. M., Graham, M. B., Greenfield, P. M., & Salimkhan, G. (2008). Self-presentation and gender on MySpace. *Journal of Applied Developmental Psychology*, *29*(6), 446–458. 10.1016/j.appdev.2008.07.001

Molder, A. L., Lakind, A., Clemmons, Z. E., & Chen, K. (2022). Framing the global youth climate movement: A qualitative content analysis of Greta Thunberg's moral, hopeful, and motivational framing on instagram. *The International Journal of Press/politics*, *27*(3), 668–695. 10.1177/19401612211055691

Mooseder, A., Brantner, C., Zamith, R., & Pfeffer, J. (2023). (Social) media logics and visualizing climate change: 10 years of# climatechange images on twitter. *Social Media+ Society*, *9*(1), 205630512311643. 10.1177/20563051231164310

Muise, D., Lu, Y., Pan, J., & Reeves, B. (2022). Selectively localized: Temporal and visual structure of smartphone screen activity across media environments. *Mobile Media & Communication*, *10*(3), 487–509. 10.1177/20501579221080333

Murashka, V., Liu, J., & Peng, Y. (2021). Fitspiration on instagram: Identifying topic clusters in user comments to posts with objectification features. *Health Communication*, *36*(12), 1537–1548. 10.1080/10410236.2020.1773702

Nesi, J., Choukas-Bradley, S., & Prinstein, M. J. (2018). Transformation of adolescent peer relations in the social media context: Part 1—a theoretical framework and application to dyadic peer relationships. *Clinical Child and Family Psychology Review*, *21*(3), 267–294. 10.1007/s10567-018-0261-x

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, Advance online publication, 1–18. 10.1080/19312458.2023.2181319

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83–110. 10.1007/s10919-009-0082-1

Otto, L. P., Thomas, F., Glogger, I., & De Vreese, C. H. (2022). Linking media content and survey data in a dynamic and digital media environment–mobile longitudinal linkage analysis. *Digital Journalism*, *10*(1), 200–215. 10.1080/21670811.2021.1890169

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. 10.1093/joc/jqy041

Peng, Y. (2021). What makes politicians' instagram posts popular? analyzing social media strategies of candidates and office holders with computer vision. *The International Journal of Press/politics*, *26*(1), 143–166. 10.1177/1940161220964769

Peng, Y., Lu, Y., & Shen, C. (2023). An agenda for studying credibility perceptions of visual misinformation. *Political Communication*, *40*(2), 225–237. Advance online publication 10.1080/10584609.2023.2175398

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, *8*(2), 1–11. 10.1145/1870076.1870082

Pouwels, J. L., Valkenburg, P. M., Beyens, I., van Driel, I. I., & Keijsers, L. (2021). Social media use and friendship closeness in adolescents' daily lives: An experience sampling study. *Developmental Psychology*, *57*(2), 309–323. 10.1037/dev0001148

Riddle, K., & Martins, N. (2022). A content analysis of American primetime television: A 20-year update of the national television violence studies. *Journal of Communication*, *72*(1), 33–58. 10.1093/joc/jqab043

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., & Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius: Sociological Research for a Dynamic World*, *6*, 6. 10.1177/2378023120967171

Shah, D. V., Sun, Z., Bucy, E. P., Kim, S. J., Sun, Y., Li, M., & Sethares, W. (2023). Building an ICCN multimodal classifier of aggressive political debate style: Towards a computational understanding of candidate performance over time. *Communication Methods and Measures*, Advance online publication, 1–18. 10.1080/19312458.2023.2227093

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). *No classification without representation: Assessing geodiversity issues in open data sets for the developing world.*

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1349–1380. 10.1109/34.895972

Steinert-Threlkeld, Z. C., Chan, A. M., & Joo, J. (2022). How state and protester violence affect protest dynamics. *The Journal of Politics*, *84*(2), 798–813. 10.1086/715600

Tiggemann, M., & Zaccardo, M. (2018). 'Strong is the new skinny': A content analysis of #fitspiration images on instagram. *Journal of Health Psychology*, *23*(8), 1003–1011. 10.1177/1359105316639436

Valkenburg, P., Beyens, I., Pouwels, J. L., van Driel, I. I., & Keijsers, L. (2021). Social media use and adolescents' self-esteem: Heading for a person-specific media effects paradigm. *Journal of Communication*, *71*(1), 56–78. 10.1093/joc/jqaa039

Valkenburg, P. M., & Peter, J. (2013). Comm research—views from Europe | five challenges for the future of media-effects research. International Journal of Communication, 7, 19.

Valkenburg, P. M. (2017). Understanding self-effects in social media. *Human Communication Research*, *43*(4), 477–490. 10.1111/hcre.12113

van Atteveldt, W., & Peng, T. Q. (Eds.). (2021). *Computational methods for communication science.* Routledge.

Vendemia, M. A., & DeAndrea, D. C. (2018). The effects of viewing thin, sexualized selfies on instagram: Investigating the role of image source and awareness of photo editing practices. *Body Image*, *27*, 118–127. 10.1016/j.bodyim.2018.08.013

Webb Williams, N., Casas, A., & Wilkerson, J. D. (2020). *Images as data for social science research: An introduction to convolutional neural nets for image classification.* Cambridge University Press.

Wojcieszak, M., de Leeuw, S., Menchen-Trevino, E., Lee, S., Huang-Isherwood, K. M., & Weeks, B. (2021). No polarization from partisan news: over-time evidence from trace data. *The International Journal of Press/politics*, Advance online publication. 10.1177/19401612211047194

Xi, N., Ma, D., Liou, M., Steinert-Threlkeld, Z. C., Anastasopoulos, J., & Joo, J. (2020). *Understanding the political ideology of legislators from social media images.* Proceedings of the international AAAI conference on web and social media. 726–737.

Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *From recognition to cognition: Visual commonsense reasoning.* Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 6720–6731.

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. 10.1177/0081175019860244

Zhang, H., & Peng, Y. (2021). Image clustering: An unsupervised approach to categorize visual data in social science research. *Sociological Methods & Research*, Advance online publication. 10.1177/00491241221082603