A survey of computational methods for iconic image analysis

van Noord, N.

DOI
10.1093/llc/fqac003
Publication date
2022
Document Version
Final published version
Published in
Digital Scholarship in the Humanities

A survey of computational methods for iconic image analysis

Nanne van Noord

Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Digitization and digitalization efforts have led to an explosive growth of the number of images that are published, shared, and made available in collections. In turn, this has resulted in increased awareness of, and interest in, computational methods for automatic image analysis. Despite the tremendous progress made in the development of computational methods, there remains a gap between how a person interprets an image and what can be automatically extracted. By considering iconic images as those images for which this gap is most salient, as their meaning goes well beyond what is represented in the visual data, this article gives an overview of the potential and limitations of computational methods for iconic image analysis. I structure this overview by discussing methods that can be used to analyse the production, distribution, and reception of iconic images. Although the majority of computational methods focus on analysing production aspects, there are promising methods for image distribution aspects, whereas methods for studying image reception have received little attention. By considering the limitations of available methods I argue that computational methods can be of use for studying iconic images, but that comprehensive analysis will require methods that incorporate the plurality of meanings an image can have, and temporal nature thereof.

Correspondence:
Nanne van Noord, Faculty of Humanities, University of Amsterdam, 1012 WX Amsterdam, The Netherlands.
E-mail:
n.j.e.vannoord@uva.nl

1 Introduction

Events entrenched in our collective memory are often represented by an image. Whether it is an image of a young girl severely burned in a napalm attack, a solitary man blocking a column of tanks, or a young boy washed ashore on a Turkish beach, they are all unmistakeably tied to the circumstances in which they occurred and recognized by most people (Perlmutter, 1998; Hariman and Lucaites, 2007). Investigating what makes these images stand out, how they reached this level of fame, and what role they played in the social debate surrounding the historical events they represent are questions central to the study of iconic images (Kroes *et al.*, 2011; Dahmen and Miller, 2012; Kleppe, 2013; Hansen, 2015; van der Hoeven, 2019). The study of iconic images has traditionally relied on close readings of images using methods from a range of perspectives (e.g. semiotics, iconography, (critical) visual studies, and sociopolitically). Yet, the tremendous increase in volume of images being digitized, published, and consumed that accompanied the advent of digital technology, has brought about a data deluge that traditional research methods can no longer cope with (Mortensen *et al.*, 2017; Dahmen *et al.*, 2018). In a world where it is possible for an image to reach an audience of millions in a matter of minutes (i.e. *go viral*), and where

https://doi.org/10.1093/llc/fqac003 Advance Access published on 23 February 2022

image collections of libraries and archives contain millions of digitized images, the study of iconic images can no longer rely on close readings and manual efforts alone.

Criteria for iconicity put forward in literature touch upon the appearance and content of images, but are mainly concerned with how, where, and by whom the images are distributed and observed (Perlmutter, 1998; Dahmen *et al.*, 2018). The emergence of digital media has not only changed the volume of images but it also affects the mechanisms by which images become iconic (Mortensen *et al.*, 2017; Dahmen *et al.*, 2018). Although mass media continue to play a crucial role, images distributed by individual users on social media can now go viral, reach large audiences, and potentially become iconic. Studying iconic images thus involves taking into account both the image itself and the context it emerged from and is used in. But what would studying iconic images at scale look like? Computer Vision (CV) (Forsyth and Ponce, 2012), the field concerned with developing computational methods for analysing large-scale visual corpora, typically does not consider image context (e.g. surrounding data, metadata related to publication and prominence, and metrics related to reception and relatedness). Large-scale visual corpora are commonly constructed by aggregating images without their context and using crowdsourcing to assign labels, which are used to train Machine Learning models (Paullada *et al.*, 2020). A reliance on visual-only corpora appears incompatible with the objective of studying iconic images, for which the context is as important as the image itself, if not more so. In this article, I aim to give an overview of the study on iconic images, to highlight the entry points for the incorporation of computational methods in visual studies, as well as to discuss the limitations of CV for studying iconic images.

A major challenge in CV is described by the Semantic Gap (Smeulders *et al.*, 2000): distinguishing between what can be extracted from an image's visual data and what the image means to a user. For iconic images, this gap takes on social and temporal aspects; the meaning of an iconic image shifts over time and is different across social/cultural groups (Spratt *et al.*, 2005; Dahmen and Miller, 2012; van der Hoeven, 2019). For instance, a common reading of the 'Tank Man' image is its embodiment of protest (Hubbert, 2014). However, an alternative reading put forward by Chinese officials is that it shows the military's restraint (Hernández, 2019), highlighting to what extent the meaning of an iconic image can differ across social groups. As such, it is undesirable to design algorithms that draw conclusions about iconicity independently of human interpretation. As 'interpretation begins with description' (Schroeder, 2006), a collaboration between human and machine can build on the machine's potential to perform description at scale and the human ability to do the interpretation. Moreover, from this collaboration the limitations of what can be achieved through computational analysis become salient, resulting in research questions that can lead to improved computational models and paradigms.

Yet, it is questionable whether the semantic gap is an appropriate guiding principle for the development of methods that consider these temporal and cultural aspects. Although Smeulders *et al.* (2000) do consider culture-based or man-made customs and how they might influence the visual data, they restrict the scope of the semantic gap to interpretation by *the user*. In practice, this results in algorithm developers tailoring to the *average user* (Ryu *et al.*, 2017; Mitchell *et al.*, 2019; Trewin *et al.*, 2019), implicitly assuming that humans are a homogeneous group of users, rather than acknowledging the myriad of meanings that can be derived from (visual) data. Such practice of user-independent interpretation is 'impossible within the critical epistemological approach of the humanities' (Drucker, 2020, p. 5). A user-independent perspective is primarily suitable for building commercial systems intended to generically serve many users. For studying iconic images, the user of the algorithm is not only interested in confirming their own interpretation, but also (and perhaps more so) in interpretations by others (Drew and Guillemin, 2014; van der Hoeven, 2019). In considering *others*, we can distinguish between individuals and groups (i.e. cultures). Obviously, incorporating every individual interpretation is impossible, but by accommodating the perspectives of multiple cultural groups we can work towards a more encompassing interpretation. In some cases, these perspectives might differ as much as they do for the 'Tank Man' image, or the appropriations of the Alan Kurdi image by Ai Weiwei and Charlie Hebdo (Mortensen, 2017). Therefore, to

contextualize images we should not collapse all possible interpretations to a single interpretation.

In not collapsing interpretations we must then also acknowledge the temporal shifts that might occur, whether it is Facebook labelling the 'Napalm girl' as pornographic (Ibrahim, 2017), or shifts in the visual collective memory concerning specific events or the meaning of iconic images (Spratt *et al.*, 2005; Dahmen and Miller, 2012; van der Hoeven, 2019). To be able to historicize iconic images, and to analyse specific temporal shifts it is necessary to account for interpretations across history. Implementations that account for temporal shifts might do this on a highly practical level, by taking into account metadata concerning creation dates, and adjusting the labels assigned with content recognition techniques. For instance, while it is accurate for a location recognition algorithm to place a recent image of the 'Berliner Fernsehturm' (television tower) in Berlin, for an image taken prior to 1990 it can be meaningful to determine (based on distance and direction) whether the picture was taken in East or West Berlin. Similarly, for object recognition, it can be meaningful to note the significance of the presence of 'exotic' fruit in historical paintings, as opposed to in modern photographs (Marks, 2019). Yet, such practical considerations alone are not sufficient. As neither the polyvocal nor the temporal perspectives are currently well-considered in computational methods, I propose a modification of the semantic gap:

The cultural gap is the lack of coincidence between the information that one can extract from the visual data and the interpretations that the same data have for cultural groups across time.

The key differences with the cultural gap are the focus on multiple interpretations, cultural groups rather than a user, and the inclusion of the notion of time. In this article, I will explore a variety of computational methods that may be used for (iconic) image analysis. By focusing on the limitations of these methods, I aim to further illustrate how and why a new objective (i.e. the cultural gap) is necessary to drive progress for computational methods that are informed by visual culture and iconicity.

This article is part of a growing body of computational work that is guided and inspired by theories and challenges related to the study of visual culture (Johnson *et al.*, 2008; Stork, 2009; Crowley and Zisserman, 2013; Elgammal *et al.*, 2018; Impett *et al.*, 2018; Lang and Ommer, 2018; Arnold and Tilton, 2019; Chávez Heras and Blanke, 2020; Münster and Terras, 2020; Wevers and Smits, 2020; Azar *et al.*, 2021). Early works in this area focused on applications for art history (van den Herik and Postma, 2000; Criminisi *et al.*, 2005; Johnson *et al.*, 2008; Yarlagadda *et al.*, 2013), computer art and aesthetics (Noll, 1966; Dietrich, 1986; Manovich, 1994), and archaeology (da Gama Leitao and Stolfi, 2002; Kampel and Melero, 2003; van der Maaten *et al.*, 2006). In recent years, this focus has been expanded to include media studies (Gehl *et al.*, 2017; Arnold and Tilton, 2019; Thomas, 2020; Matud *et al.*, 2021), history (Smits, 2017; Wevers and Smits, 2020), historical maps (Budig *et al.*, 2016; Weinman *et al.*, 2019; Hosseini *et al.*, 2021; Uhl and Duan, 2021), and archives (Chung *et al.*, 2015; van Noord *et al.*, 2021), as well as closer collaborations between computational and Social Science and Humanities (SSH) scholars (Olesen, 2015; Wevers *et al.*, 2018; Bocyte and Oomen, 2020; Masson *et al.*, 2020). This expanded focus has in turn led to an increased awareness that computational research can look to the Humanities for inspiration, for instance, on how to deal with problems concerning ethics and inequality (Jo and Gebru, 2020; Mohamed *et al.*, 2020; Offert and Bell, 2020; Parisi, 2020).

The remainder of this article is structured as follows to guide a discussion of why it is necessary to think in terms of a cultural, rather than a semantic, gap: Section 2 starts with a brief overview of Humanities literature on iconic images, underscoring the role of icons as vehicles of social knowledge and acknowledging the role they play in our (changing) society. Subsequently, I present an overview and discussion of computational methods for analysing (aspects of) iconic images in Section 3. Finally, in Section 4, I conclude and suggest future directions to advance the computational study of iconic images based on insights from visual culture research.

2 Iconic Images

Icons are religious artworks used in various schools of Christianity to depict key religious figures (e.g. Christ, Mary, saints), where the icons are imbued with properties that go beyond what is simply depicted (Kleppe, 2013). This last point is what is central to iconic

images; their meaning and interpretations of their meaning extend beyond what is depicted. Iconic images are often synonymous with events in history, i.e. when remembering the 1989 Tiananmen Square protest, the image that springs to mind is that of a solitary man, on a street leading away from the square, blocking a column of tanks, as opposed to the singing, dancing, and protesting students on the square itself. Iconic images have also been ascribed with the ability to change the course of history, such as the 2015 photograph of deceased 3-year-old, Alan Kurdi, washed ashore on a Turkish beach, which became iconic of the European refugee crisis (Binder and Jaworsky, 2018), and which marked a shift in the public debate on refugees (Vis and Goriunova, 2015). Although the ability of iconic images to bring about societal change (i.e. visual determinism) is debated (Perlmutter, 1998; Hansen, 2015; Binder and Jaworsky, 2018) it is clear that images (and increasingly video) play an important role as not only evidence, but also initiators of news. His role as initiators of news is particularly salient when considering the photographs of Alan Kurdi and Abu Ghraib, or the videos of the beating of Rodney King and the murder of George Floyd. Studying images and the process by which they became iconic can aid in understanding the situation they emerged from, as well as the social knowledge and dominant ideologies they reflect (Lucaites and Hariman, 2001; Hariman and Lucaites, 2007).

Various features of iconic images have been described in literature; here, I follow the categorization by Kleppe (2013) that splits these features into three groups: *production*, *distribution*, and *reception*. These three groups describe the path an image travels along as it is captured by a photographer to being consumed by an observer. These represent key concepts in Media Studies, in particular for understanding media in relation to institutions and audiences. In previous research, these three groups have typically been studied with different methodologies. The production of iconic images concerns the image content and aesthetics, and has predominantly been studied through case studies with an iconographic or semiotics focus (Kleppe, 2013, p. 26; van der Hoeven, 2019, p. 40). Distribution concerns the availability and reproduction of images, and has primarily been studied through labour-intensive efforts that involved manually counting the occurrence of images (Kleppe, 2013; Cohen *et al.*, 2018). The third group describes features related to image reception, and concerns the symbolism, associations, and emotions an image evokes in the recipient, i.e. its 'meaning'. Central to reception is that the meaning of an image is individual to each recipient; an image does not have a universal meaning, nor is the meaning (fully) encoded in the image itself. To study reception previous research has focused on asking participants how they perceive and contextualize images, primarily through (*ad hoc*) interviews (Hariman and Lucaites, 2007; Cohen *et al.*, 2018) or large-scale surveys (Cohen *et al.*, 2018; van der Hoeven, 2019). Of the three groups of iconic image features, reception is arguably the most determining factor for iconicity, but it is also the least formalized and most subjective, making it difficult to study computationally.

A common thread among studies of iconic images is the focus on specific domains (Dahmen and Miller, 2012; Kleppe, 2013; Meuzelaar, 2014) or a small number of images (Dahmen and Miller, 2012; van der Hoeven, 2019). As close viewings of large collections of images are incredibly labour-intensive, researchers tend to focus on traditional media or use images, which are established as being iconic (Hariman and Lucaites, 2007). Yet, from the literature it is apparent that which images are recognized is not consistent across geographical location and social group (Cohen *et al.*, 2018; van der Hoeven, 2019), which makes it unlikely that studies focusing on specific domains, media, or images will be able to cover the full breadth and complexity of iconic images beyond the scope of the images investigated. An additional challenge in this is that what is considered iconic has a temporal dimension as well; iconic images are a reflection of the social knowledge and dominant ideologies when they emerged (Lucaites and Hariman, 2001; Hariman and Lucaites, 2007). But over time, the meaning of an iconic image might shift (Hubbert, 2014; Ibrahim, 2017). It is therefore necessary to include data from diverse geographical locations, time periods, and cultures, which makes it impossible to only consider small-scale datasets.

Until now I have discussed what Perlmutter (1998) describes as *discrete* iconic images: images that concern a specific event, and that have become symbolic of the event. Among discrete iconic images there are

certain 'supericons' (Perlmutter, 1998) which have been propelled to almost instantaneous fame, and that capture important historical events. These supericons are very known and are the first examples we think of when considering iconic images. In addition to discrete icons, Perlmutter also describes *generic* iconic images, which illustrate a common theme by documenting recurring situations with shared characteristics. For instance, while no specific image springs to mind, the images that are conjured up when thinking of a starving child in Africa, a polar bear on melting ice, or a Palestinian stone thrower, share many characteristics. Generic icons are stereotypes or visual tropes that remain consistent across geopolitical context, and encapsulate tropes such as the solitary civilian facing soldiers or policemen or the poverty-stricken mother with children (Zarzycka and Kleppe, 2013). In this sense, generic iconic images are strongly tied to the study of visual rhetoric (Lucaites and Hariman, 2001; Foss, 2005). The manner in which icons refer back to previous icons is used to increase their visual power. Referencing an older icon to support iconic status has been described as *intericonicity* (Hansen, 2015). Interestingly, the act of referencing also reinforces the iconicity status of the older icon, as noted by (Hariman and Lucaites, 2007, p. 12): 'As the image is known for being known, it becomes a technique for visual persuasion.'

As a technique for visual persuasion, iconic images play an active role in public debate and in the writing of history. Historically, the images given this role were chosen by elite media who acted as gatekeepers, determining which images were 'newsworthy' and what prominence to give them (Dahmen and Morrison, 2016). In 'routinizing the coverage of a nonroutine story' the media uses familiar (visual) narratives to make stories easier to understand for the public (Dahmen and Miller, 2012). Through this reliance on familiarity, images are placed in pre-existing and familiar themes, underscoring the role of (generic) icon. However, with the emergence of the digital, the gatekeeping role of traditional media is waning (Dahmen and Morrison, 2016; Mortensen *et al.*, 2017). On the one hand, digital publishing has reduced the need to select single images to reflect events; online articles frequently contain multiple images or even image galleries. On the other hand, the digital has boosted citizen journalism, making it possible for images to reach large audiences without having been selected, or approved, by a news editor.

Besides a shift in gatekeeping, the digital age has also resulted in a tremendous increase in the volume of images seen and distributed, leading to icons with a reduced life-cycle that are quickly replaced, which are so-called 'hypericons' (Dahmen *et al.*, 2018). Sharing is ubiquitous on social media; a message or image can be shared millions of times in the span of minutes, causing it to 'go viral' (Nahon and Hemsley, 2013). Viral images are similar to hypericons in that their fame is almost overnight, and that their fame is typically not long-lasting. However, a key distinction between iconicity and virality is that the former presupposes a representativeness of an event, and that the icon becomes intrinsically linked to the event. In this sense, not all hypericons are true icons: While they might be representative of an event and have—briefly—been famous, they do not persist in our collective memory (Dahmen *et al.*, 2018). Changes in research methodology are necessary to keep up with these changes in the media landscape. For instance, virality can be used as an indicator that an image might be (or become) iconic, but only if the fame persists for an extended period of time, thus requiring longitudinal analysis. To cope with these changes, the increased volume and replacement rate of iconic images, and to evaluate the effects changing role of images in society, we can incorporate computational techniques in iconic images research.

3 Computational Iconic Image Analysis

Although computational techniques and methods can be borrowed from a range of fields, for visual analysis the field of primary interest is CV. One of the earliest works on CV was the Summer Vision Project at MIT (Papert, 1966), which aimed to 'in a single summer' solve automatic visual recognition. Around that time (the 1960s) the prevailing idea was that vision could be solved by simply recognizing which objects were in the environment, and that this solution was sufficient such that a robot could interact with its environment. As it turns out, solving vision is a lot more challenging than previously assumed: Not only is it more

challenging to recognize objects in the environment, but solving vision also requires solving a wider range of problems. This range of problems has been slowly chipped away at since then, and has come to be known as the Semantic Gap, as described in the seminal work by Smeulders *et al.* (2000, p. 1352):

> The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.

The width of this gap is not consistent across images; it can be narrow for literal images, and broad for images which are polysemic or for which the semantics are only partially described. Iconic images clearly fall in this latter category, with much of their meaning not being described by the visual data. I thus argue that for iconic images, the semantic gap does not cover the main challenge: What makes the image iconic is not captured by its visual data, and proper interpretation of the image requires substantial (cultural) domain knowledge and attention to contextual data. The context of an image involves a wide range of information, some of which can be readily expressed as data, such as the location of publication (e.g. website, page, above or below the fold) or adjacent data (e.g. other images, captions, or article text). Other contextual information that influences interpretation has to do with the recipient and their cultural background (van der Hoeven, 2019; Drucker, 2020).

Because of the complexity of iconic images, there is no single method that can be used to model them and their context. Instead, in the following I discuss various methods and approaches that capture some aspect(s) of iconicity. I structure this section by dividing the discussion into three parts centred around the three stages of iconic image-making: production, distribution, and reception. This article is not intended as a complete overview of all computational literature, but rather to highlight the breadth and potential of computational research to study (aspects of) iconic images. A broader perspective on using computational methods to study visual culture is given by Arnold and Tilton (2019), who present *Distant Viewing*, a methodological and theoretical framework for analysing large collections of visual culture. Similarly, distant viewing (and this article) can be placed within the broader framework of cultural analytics (Manovich, 2018), that studies cultural datasets in the broadest sense.

3.1 Production

From a computational perspective, an image is made up of individual pixels. To extract information from an image, it is necessary to look beyond pixels and determine what it represents and what is contained in the image. In terms of production aspects, information can be extracted from visual data (i.e. the pixels), by discovering patterns and recurring elements that form the foundation for further analysis. In addition to content aspects, there are also aesthetic aspects, or as stated by Hariman and Lucaites:

> The iconic image is a moment of visual eloquence, but it never is obtained through artistic experimentation. It is an aesthetics achievement made out of thoroughly conventional materials (Hariman and Lucaites, 2007, p. 30).

This lack of artistic experimentation follows from the origin of iconic images, as they are typically news photographs used to provide a visual grounding in the familiar, to give readers a familiar frame of reference (Dahmen and Miller, 2012). The 'conventional materials' and their aesthetics are the production aspects of iconic images, and interpretation can be guided by representing these as codifiable units. Media Studies, and specifically Production Studies, consider a wider range of production aspects as opposed to only those related to the visual data, such as institutional aspects (e.g. training, reputation, or employer of photographer) (Mayer *et al.*, 2009). Although such institutional aspects may contribute to the iconic status of an image, they are ill-suited for computational analysis and will not be discussed in this article. Instead, I will focus on codifiable units which can be extracted from the visual (or contextual) data.

For computational analysis, it is necessary to use an algorithm to represent an image in a manner such that a specific task can be performed. This process of representation is a mathematical transformation that highlights selected codifiable units, while discarding unrelated image information. The question of which units to highlight and which to discard is based on the task to be performed. Earlier works in CV used

manually defined algorithms, where engineers decided what 'features' (i.e. codifiable units) to select and what to discard; hence, this process is called feature engineering (Zheng and Casari, 2018). With the emergence of 'Deep Learning' the determination of how to represent the visual data (i.e. which features to select and discard) has become encapsulated in the algorithm as part of an optimization process to learn the 'best' representation given the task (Bengio *et al.*, 2013). As part of the shift in CV to deep learning techniques, the algorithms have become black boxes, with even the developers not fully understanding which features are used (Zhang and Zhu, 2018). This lack of transparency is an issue, particularly for fully automated decision-making processes, and has attracted a large amount of research and legislative focus (Wachter *et al.*, 2017; Mitchell *et al.*, 2019; Alameda-Pineda *et al.*, 2020).

In CV the majority of works are focused on content recognition, such as the recognition of objects (Russakovsky *et al.*, 2015), persons (Li *et al.*, 2014; Liu et al., 2018), actions (Wang and Schmid, 2013), and locations (Torii *et al.*, 2013; Zhou *et al.*, 2014), rather than fully automated decision-making. Arguably, 'deciding' what or whom is considered a person by the algorithm is equally an issue that should be investigated under the banner of automated decision-making (Buolamwini and Gebru, 2018). Nonetheless, there are algorithms that extract codifiable units from images which can be used to enable visual analysis at scale. Figure 1 shows an example of an image for which codifiable units are automatically extracted. In the following, I will discuss a variety of algorithms, what codifiable units they extract, and how these are relevant for studying production aspects.

3.1.1 Content recognition

Object recognition is one of the most well-studied topics in CV, with tremendous progress being made on prominent datasets such as ImageNet Large-Scale Visual Recognition Challenge (Russakovsky *et al.*, 2015), and the Microsoft Common Objects in Context (MSCOCO) (Lin *et al.*, 2014). The ImageNet dataset consists of over 14 million images across 20,000 object classes. Yet, most research is done using the reduced set of 1.3 million images from 1,000 classes. The MSCOCO dataset has fewer images and object classes (330K and 80, respectively), but instead has five textual descriptions (captions) per image, spurring on work that connects visual and textual analysis. The objects studied in these datasets are primarily household objects, animals, vehicles, and foodstuffs, as opposed to objects that are commonly used as symbols (e.g. human skulls in visual art to represent death or mortality), or recurring objects in iconic images (e.g. young children, weapons, or soldiers) (Zarzycka and Kleppe, 2013). Broad application

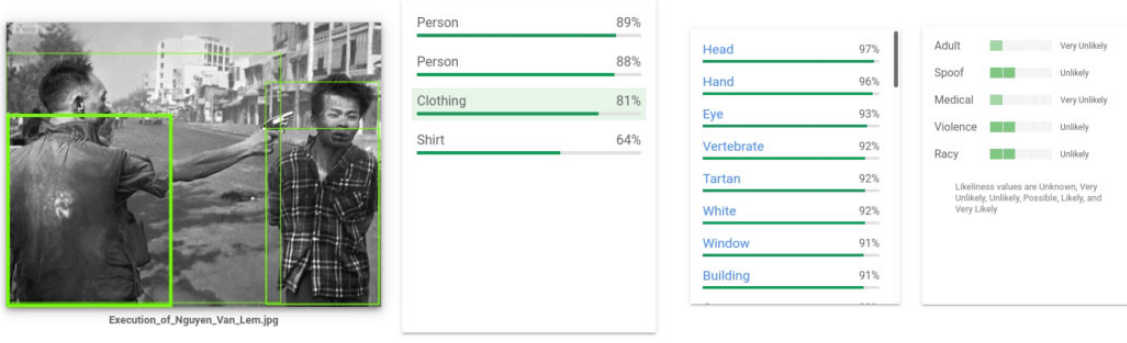

Fig. 1 Result of applying the Google Cloud Vision API to the photo of the execution of Nguyen Van Lém. (Photo of 'the Execution of Nguyen Van Lém' by Eddie Adams (Public domain).) Showing (from left to right) the photo overlayed with green boxes for the four 'objects' found, labels and confidence scores for the objects, a list of labels assigned, and the 'Safe Search' results with likeliness values. The API returns accurate annotations (i.e. codifiable units), but fails to recognize the weapon and hence the nature of the image (as indicated by the 'Safe Search' score for violence). (a) Hong Kong pro-democracy protest gesture. (b) Politicians shaking hands

of these methods is thus inhibited by the objects which can be readily recognized. Moreover, these datasets consist of images from countries with higher income levels and reflect this bias by lacking the ability to recognize these objects in countries with lower income levels (de Vries *et al.*, 2019). However, such limitations do not disqualify object recognition for studying iconic images. Instead, I argue that the potential of these methods is clear and that these limitations can (to a certain extent) be overcome by developing more diverse and richer datasets, for instance by expanding the types of objects considered.

Although persons are sometimes considered to be objects in CV (and thus recognized by object recognition algorithms), there are specialized algorithms that extract codifiable units related to persons. For instance, one of the criteria used by Perlmutter (2005) to determine iconicity is that of 'Fame of Subjects', which could be modelled through face recognition trained on celebrities (Liu *et al.*, 2018). In CV research, a distinction is made between face *recognition* and face *detection*, with the latter focusing on the act of finding faces in an image regardless of whom they belong to. The former, which builds on top of face detection, is aimed at recognizing which exact person the face belongs to (Jain and Li, 2011). Although primarily focused on modern day celebrities, datasets such as CelebA (Liu *et al.*, 2018) and MS-Celeb (Guo *et al.*, 2016) have image data of thousands of celebrities, enabling analysis at scale of the Fame of Subjects criterion. Yet, as noted by Dahmen *et al.* (2018), the image of *Raising the Flag on Iwo Jima* goes against 'Photojournalism 101 classroom standards' for not showing the faces of the subjects. Iconic images often do show faces, and detecting these faces (without recognizing the person) can therefore also be used for visual analysis, as well as to detect *outliers* that do not contain faces.

In addition to recognizing faces, it is also possible to recognize human behaviours such as actions (e.g. handshaking, hugging, running) (Wang and Schmid, 2013), gestures (Freeman and Roth, 1995), or poses (Shotton *et al.*, 2011; Toshev and Szegedy, 2014; Impett and Süsstrunk, 2016). By recognizing these behaviours, we can codify signs expressed by humans to analyse the content of iconic images. Gestures are used to accompany speech to convey information (Beattie and Shovelton, 1999), but humans also use gestures or physical actions to indicate their belonging to certain groups (i.e. law enforcement officials making a 'white power sign')[1] or adherence to cultural conventions (i.e. shaking hands or bowing). Certain images, such as images of handshakes between government officials, or Hong Kong pro-democracy protestors making the 'Five demands, not one less' gesture, as shown in Fig. 2, can be categorized based on codified human behaviours. For this latter example the gesture is often the only information in the visual data to link an image to an event, whereas in the former it could be complimented by other forms of recognition (e.g. face and place recognition).

Another content aspect of images that can be analysed is the physical environment depicted. Similar to the distinction between face recognition and detection, this can be done at two abstraction levels: scene recognition (Zhou *et al.*, 2014) and place recognition (Torii *et al.*, 2013; Lowry *et al.*, 2015). In scene recognition, the aim is to assign a scene category label to an image, where the scene categories include both indoor and outdoor spaces (e.g. bedrooms, airfields, discotheque, greenhouse, hot spring, etc.) (Zhou *et al.*, 2014). Place recognition is concerned with identifying the exact location, by recognizing the street in a city (Torii *et al.*, 2013) or known landmarks (Weyand *et al.*, 2020), or by predicting the (approximate) GPS coordinates (Zamir and Shah, 2010). For press images, as opposed to 'images in the wild' (i.e. on the Internet or social media), the location can most likely be identified more accurately based on contextual data (e.g. image caption) or metadata (e.g. GPS information in EXIF data).[2] A categorization in scenes on the other hand does require inspection of the visual data, as it is typically not captured in the image metadata. Through scene recognition, information about the physical environment an image is taken in can be codified, and thus extracted from image collections at scale.

3.1.2 Aesthetics

CV research into aesthetics can roughly be grouped into two lines: a line akin to stylometry, and a data-driven classification approach. The former focuses on learning visual characteristics indicative of style or authorship (Johnson *et al.*, 2008; Karayev *et al.*, 2014; van Noord *et al.*, 2015; Gatys *et al.*, 2016), while the latter is focused on what is referred to as *Subjective*

Fig. 2 Two examples of images which can be categorized based on human behaviours in the visual data. (Left photo by Studio Incendo (CC BY) and right photo by US Mission Canada (CC BY).) On the left a photo of Hong Kong protesters making the pro-democracy gesture, on the right two politicians shaking hands. The gestures/actions of the persons contextualize and place the images

Attributes. Algorithms to predict subjective attributes are trained on large collections of images where each image is assigned a numerical score. The manner in which these scores are obtained ranges from crowdsourced ratings (Murray *et al.*, 2012; Wilber *et al.*, 2017), to social media statistics (e.g. number of likes, views, or resubmissions) (Khosla *et al.*, 2014; Deza and Parikh, 2015), to using experimental procedures (Khosla *et al.*, 2015). The variety in data collection methods for subjective attributes scores gives rise to a difference in the subjectiveness of the collected scores. Crowdsourced ratings, for instance, represent *perceived* ratings, reflecting their highly subjective nature. This is different for the experimental procedure described in (Khosla *et al.*, 2015), as it directly measures the memorability of images despite also relying on crowd workers. However, for scores based on social media statistics the story is more complex, as there is an interaction with human behaviour and with the specific design and functionality of the platform itself.

Inspired by art historical analyses, attempts have been made to relate visual characteristics of artworks to those of specific artists. The characteristics considered in these approaches range from fine-grained details, such as brushstrokes and material textures (Johnson *et al.*, 2008; Li *et al.*, 2011; Gatys et al., 2016), to approaches using general purpose CV algorithms that focus on iconography and content (Karayev *et al.*, 2014; van Noord *et al.*, 2015). These latter approaches are, on a technical level, highly similar to approaches for content recognition; however, the former are different both on a technical and application level. In recent years, work on neural style transfer (NST) by Gatys *et al.* (2016) has attracted a lot of attention. NST aims to adjust an image *A* to the 'style' of image *B*, while retaining the content of *A*, as shown in Fig. 3. This raises questions about how NST defines style, versus how a human observer would. From the top row in Fig. 3, we can observe that while the stylized image of Che Guevarra does not look like a Van Gogh, it has become visually more similar to the painting. However, when applying the style of one photograph to that of another (Fig. 3, bottom row), it is apparent that the *style* of the Dorothea Lange image was not transferred. Despite the narrow definition of style by NST algorithms, their ability to create highly stylized images that have a clear resemblance to the style image is remarkable. Recent work by Gairola *et al.* (2020) has demonstrated that the NST approach can be extended to group images, which are aesthetically similar, despite having different image content. Being able to group images based on style makes it possible to explore aesthetic themes that would not become apparent by focusing on content only. For instance, the iconic Barack Obama 'Hope' poster[3] has become a source of many parodies or imitations, resulting in images which are similarly stylized but with a different caption and a different person depicted. Although there are no content elements to group variants of the Hope poster together, style

Fig. 3 Example of Neural Style Transfer, which combines the content of the left most image with the 'style' of either the painting by Van Gogh (top row), or the photo by Dorothea Lange (bottom row). (Guerrillero Heroico by Alberto Korda (Public domain) stylized in the 'style' of Self-Portrait with Straw Hat by Vincent van Gogh (image credits Van Gogh Museum, Amsterdam (Vincent van Gogh Foundation)), and in the 'style' of Migrant Mother by Dorothea Lange (Public domain). Stylized using Arbitrary style transfer by Reiichiro Nakano.)

recognition offers a possible solution to find images with a similar style nonetheless. This process could also be used for other iconic styles, such as the Che Guevara poster image, or parodies of Andy Warhol's artworks.

Subjective attribute research has primarily investigated the popularity and virality of images on social media, identifying a variety of factors that go beyond the visual data (Nahon and Hemsley, 2013; Weng *et al.*, 2012; Figueiredo *et al.*, 2014; Goel *et al.*, 2016). For instance, the network size of the sharer plays a role in the popularity of the shared content (Susarla *et al.*, 2012; Weng *et al.*, 2012; Figueiredo *et al.*, 2014; Khosla *et al.*, 2014). Predicting subjective attributes directly from images brings up the question of whether they are intrinsic to images. Isola *et al.* (2011) investigated this question for memorability and found that images containing people in enclosed spaces, with visible faces, correlate more strongly with memorability than images of vistas or peaceful settings. I would thus argue that while subjective attributes cannot ensure whether an image will be popular,

remembered, or go viral (i.e. they do not describe a causal relation), they are useful for measuring the presence of features that correlate with such attributes.

In the scope of subjective attributes, a logical step for the study of iconic images would be to construct a predictive model of iconicity. Surprisingly, as of yet no studies have been conducted in this direction. However, there are a number of works on canonical representations of concepts, which have used iconicity as a synonym for canonical (Berg and Forsyth, 2007; Berg and Berg, 2009; Zhang *et al.*, 2014). While these works touch upon aspects of production that are relevant to the study of iconic images, a vital difference is that works on canonical (or typical (Ehinger *et al.*, 2011)) views are only concerned with visual concepts. This distinction is perhaps most obvious when considering an example by Berg and Forsyth (2007), who rank a collection of images of the Statue of Liberty (in New York, USA) based on how well they match canonical views. Conversely, based on the notion of iconicity I follow in this article, the Statue of Liberty itself could be considered an icon for the USA. Nonetheless, there is an interaction between these two notions, because while the statue itself is an icon for the USA, this obviously does not apply to *every* image of the Statue of Liberty (e.g. a close-up of the statue's nose). As such, for an image of the Statue of Liberty to function as an icon it should match a canonical view.

3.2 Distribution

Analysing production aspects can be considered a bottom-up approach, as it starts from observations about a single image's visual data. Subsequently, observations about single images can be tabulated to make observations about collections. Distribution aspects, on the other hand, are inherent to image collections. In a top-down fashion, a collection can be divided into subgroups with shared properties, such as images which frequently appear together, or images which visually reference each other. For distribution, analysis centres on aspects that are directly related to the circulation of images, and those aspects that emerge from images being circulated, such as intericonicity. Moreover, while the visual data is used in the analysis of distribution aspects, it alone is not sufficient; it is necessary to incorporate contextual information.

Among the iconic image qualities proposed by Perlmutter (1998), three relate strongly to distribution: Prominence of Appearance in media, Frequency of Appearance in media, and Instantaneousness of Fame. As noted by van der Hoeven (2019, p. 44) '[. . .] in our media-filled world with its huge supply of media outlets, a thorough and comprehensive inventory of prominence of photograph appearance [. . .], would be a gargantuan task.' Yet, this type of inventory construction task is one that computational methods excel at. Especially methods for Image Retrieval (Datta *et al.*, 2008), such as near-duplicate image detection (NDID) methods (Ke *et al.*, 2004; Chum *et al.*, 2007), are highly suited for constructing inventories of image appearance. NDID is a retrieval method specifically geared towards images which are duplicates or near-duplicates (i.e. images which are variants of each other but have undergone some transformations, for instance cropping or with text overlay), NDID can be used to tally the frequency of appearance of images by applying it to large collections. Moreover, by incorporating archival metadata about the date of publication and placement (i.e. above the fold, page number, or follower count of the sharer), data can be gathered about image prominence and the instantaneousness of its fame. Additionally, as shown by Moreira *et al.* (2018) such techniques can be used to determine image provenance even after a number of modifications.

Methods for Image Retrieval and NDID are very advanced, but the main challenge in using these methods for measuring the distribution of iconic images is the availability of data. Although a lot of historical material has been digitalized, this is often restricted to specific collections or certain document types—the large costs associated with digitalization forces institutions to make deliberate choices and prioritizations. As such, there are structural omissions in what has been digitalized that would bias the results of NDID (Valeonti *et al.*, 2019; Candela *et al.*, 2020; Jo and Gebru, 2020). Despite these limitations, computational and computer-assisted analyses of Gallery, Library, Archive, Museum (GLAM) collections are becoming more commonplace, demonstrating the new possibilities of such analyses (Arnold and Tilton, 2019; Masson *et al.*, 2020). Nonetheless, such approaches have, as yet, been primarily driven by content recognition, investigating collections in a bottom-up

fashion. Works such as the article on discovering meme genres by Theisen *et al.* (2020), or articles on learning semantic groupings by relying on the data instead of labels (Xie *et al.*, 2016; Chang *et al.*, 2017), do enable new possibilities for exploring image collections, while using the collection itself as the starting point.

The analysis of the recurrence of specific images is primarily of interest when dealing with distinct iconic images. However, when considering intericonicity, there might also be recurring motifs or visual patterns that can only be discovered by analysing a collection as a networked whole, as opposed to analysing images in isolation (Warnke and Dieckmann, 2016). Such patterns touch upon generic iconic images, but also on the concept of denotation from semiotics, as highlighted by van Leeuwen (2004, p. 95) when discussing physiognomic stereotypes. Once known, these patterns or stereotypes can be recognized in an individual image, but establishing them requires studying traditions of representation. In their work on visual link retrieval, Seguin *et al.* (2016) algorithmically produce pairs of (parts of) painting images which are considered visually linked by art historians. The visual links they establish primarily concern notable characters (in specific poses) and landscape elements, but they could also include other iconographic elements (e.g. apples, skulls). While promising, the approach by Seguin *et al.* (2016) requires manual annotation, which creates challenges related to scaling (i.e. large number of possible patterns) and in terms of bias (i.e. only reproducing known or established patterns). More recent work by Shen *et al.* (2019) demonstrates the possibility of discovering similar relationships without annotation, and across datasets of a respectable size.

A different approach to the discovery of shared visual information is presented by Hu *et al.* (2019), who propose a method for localizing the common object across a set of images. Although training their method requires manual annotations, their *few-shot* approach highlights the potential for scaling such a method to many objects, or potentially even previously unseen objects. This latter direction seems especially promising when we consider the progress that is being made in Object Discovery (Arandjelović and Zisserman, 2019), which aims to recognize objects without relying on manual annotations. Conceptually, such works build on the idea that there are image *building blocks* that reoccur, akin to words in language, and that these can be discovered by analysing and comparing images.

Besides fully automatic approaches, a promising direction for the analysis of the distribution of images and visual concepts is the use of Multimedia Analytics (Zahálka and Worring, 2014), which combines visual interfaces with algorithms for multimedia analysis. In Multimedia Analytics, the emphasis is on interactive workflows that enable users to perform analytical tasks on large-scale datasets. Applications in this area rely on the idea of a *similarity space*, a high-dimensional space where the similarity between items is expressed by their distance, i.e. highly similar items are close together, and differing items are apart (Masson *et al.*, 2020). Through interactive interfaces, the metric that is used to establish the similarity can be refined with (inter)Active Learning (Settles, 2009). This is for example employed by Lincoln *et al.* (2020) in the CAMPI project, using an expert-in-the-loop approach to quickly (and manually) assign tags to photo archives, by leveraging automatic similarity judgements to group images so that they can be tagged together. Such tools fit well with a Humanities approach to visual analysis, where 'there is a constant and systematic visual comparison between similarities and dissimilarities, which is the key for noticing relevant phenomena' (Parmeggiani, 2009, p. 75).

3.3 Reception

Studying the reception of iconic images requires an approach that is the furthest removed from existing computational methodologies, which by and large prefer well-defined unambiguous ground truths—even if it requires a formalization that is disconnected from the original meaning of the concepts investigated (Agre, 1997). Reception (and interpretation) on the other hand is modulated by the historical and cultural background of individuals (Perlmutter, 1994; van der Hoeven, 2019; Drucker, 2020). How an image is received is not fully contained in the image itself, and does not follow directly from how an image is distributed. With regard to the iconic image qualities proposed by Perlmutter (1998) the three qualities that relate most directly to reception are: (1) Importance of the Event Depicted; (2) Metonymy; and (3) Primordiality and/or Cultural Resonance. This first quality, the importance of the event depicted, directly

highlights the subjectivity of reception aspects, as what is considered important varies strongly across time and social group. To analyse these attributes, previous research has relied on interviews and surveys to gather data from people in different cultures (Cohen *et al.*, 2018; van der Hoeven, 2019).

Nonetheless, there are ways in which we can gather quantitative clues about the reception of an image. For instance, with respect to the importance of the event depicted, the image itself might contain information about which event it concerns. Additionally, contextual information such as image captions, tags, or the surrounding text can be used to link an image to a specific event. Existing work on (multimodal) event recognition has primarily focused on an abstract notion of events, such as sports matches, birthday celebrations, or wedding ceremonies (Jiang *et al.*, 2011; Jiang, 2012). While this notion of *event* does not align with the notion used by Perlmutter (which refers to historical events), it can be used to determine the importance of the event depicted (i.e. certain events such as weddings might be considered inherently more important than sports matches or birthdays). However, in order to touch upon iconic image quality as described by Perlmutter, it will be necessary to connect images to historical events. As of yet this research direction is largely unexplored, but as illustrated by the work of Yang *et al.* (2011) it is possible to perform topic modelling on historical documents, which on occasion resulted in topics related to historical events. Additionally, event ontologies for historical Linked Data (Hyvönen *et al.*, 2012) might provide entry points to pursue linking images to historical events further. Although most Linked Data of this type has to be assigned manually, progress is being made in automatic linking which might eventually offer a solution for linking images to historical events (Le and Titov, 2019). Linking images to historical events offers two routes for facilitating the understanding of image reception. First, information and metadata can be shared or propagated between images associated with the same event. For instance, images associated with an important event might themselves be considered more important. Secondly, a visual signature can be constructed for an event, allowing for the detection of outliers. For example, in this manner pictures of the WW1 Christmas Truce might stand out from the otherwise often gruesome imagery associated with WW1.

To make it possible to computationally study reception, it is key to embrace that images do not exist in isolation: They are accompanied by information in other modalities (e.g. speech and text) that provide clues about the topic of discussion, sentiments prevalent in the discussion, and sentiments about the image. This thus calls for *in situ* analysis of images, which deviates from standard computational practice that is more akin to *in vitro* analysis: removing images from their context and constructing corpora of images only. Although multimodal datasets are not uncommon, image-centric multimodal datasets primarily contain textual captions which are descriptive of the image content (Lin *et al.*, 2014) or questions related to the image content (Goyal *et al.*, 2017), rather than containing additional context information. Notable exceptions here are the KBK-1m (Kleppe *et al.*, 2017) and Newspaper navigator datasets (Lee *et al.*, 2020), which contain newspaper photographs and the captions or headlines that accompanied them, and the Good News dataset (Biten *et al.*, 2019), which contains the article text in addition to the captions and images. While these latter datasets are promising in that they use press photographs and preserves some of their (textual) context, they still do not match an *in situ* analysis. Arguably, truly modelling an image *in situ* is unachievable, at least with existing technology and methods, which raises questions about what contextual information is necessary to study image reception. Bateman *et al.* (2016) present an annotation module, ICON, for describing images in online communication, consisting of five layers: motif, genre, composition, *consociation*, and context. The first three layers focus on visual aspects, whereas the fourth examines semantic relationships between elements, and the fifth links the verbal to the visual. The ICON module describes the information needed to perform sociopolitical analysis and interpretation, which relates to studying reception but does not fully cover it. Positional relationships (e.g. above the fold, page number), co-occurrences of images, text in addition to captions, and information derived from analysis of the production and distribution can all be used to contextualize an image. A deeply contextualized analysis that takes into account all these factors could perhaps be considered *in situ*.

The second quality related to reception is the metonymy of an image, the extent an image is taken to stand for a wider event (Perlmutter, 1998; Perlmutter and Wagner, 2004). Although the prevalence of this has reduced since the introduction of digital photography (Dahmen *et al.*, 2018), a single image is often still used to represent historical events in, for instance, history textbooks (Kleppe, 2013). As images can function metonymically without actually showing an event, or without showing enough to be able to infer the event, the visual data often cannot be used to determine the metonymy. For instance, 'The Falling Man' a photograph by Richard Drew, shows an unrecognizable man falling from the World Trade Center during 9/11. The photograph is zoomed in, and while we can recognize that there is a building in the background, from the visual data alone it is difficult to determine (computationally) that it is the World Trade Center. Nonetheless, the image is unmistakeably tied to this event, even when recognized as such based on contextual clues. Through frequency analysis of how often an image is used when discussing an event, and not for other events or topics, we can establish the metonymy of an image. However, practically doing this requires an extensive and elaborate index of images and their contexts, which might simply not be available.

The third and last iconic image quality discussed here concerns the primordiality and cultural resonance of images. This quality is by and large encapsulated in the previously discussed notion of intericonicity, and the manner in which images reference (or invoke) existing images or visual tropes. Yet, an image might make references to (non-visual) ideas or notions that are specific to certain cultures (Dahmen *et al.*, 2018). This cultural specificity is at odds with the frequentist approach to intericonicity when considering the distribution aspects of images, as discussed in Section 3.2. A frequentist approach of counting how often a certain reference is made is well-suited for discovering the most prominent references, but it does not touch upon the strength of the resonance a cultural reference has for a specific recipient. Computational tools are not, and will not, be able to determine the cultural resonance of an image for an individual, but rather they reflect the popular or majority vote (van Erp and de Boer, 2020). Thus, a risk of documenting the primordiality and cultural resonance of an image with computational tools is that certain perspectives are favoured over others, or worse, that minority perspectives are not included at all.

To make computational tools that support iconic image research broadly applicable, a polyvocal methodology is necessary, such that the cultural resonance and importance of events can be tailored to a range of cultures and social groups. For instance, due to the prominent role of American media in large parts of the world, images which are iconic in the USA—at least in part—due to their strong cultural resonance, might only be known in other parts of the world because they are 'famous for being famous'. Findings by van der Hoeven (2019) show that for photograph such as the 'Times Square Kiss' by Alfred Eisenstaedt, the 'Migrant Mother' by Dorothea Lange, or 'Raising the Flag on Iwo Jima' by Joe Rosenthal recognition rates differ sharply between the USA and the rest of the world. Further findings show that even when respondents correctly identified a photograph, and the event it was associated with, they frequently did not recognize the 'message' (van der Hoeven, 2019, p. 156). Why and if an image is considered iconic thus varies across country, culture, and time, as yet no suitable computational methodology exists to investigate this at scale.

4 Conclusion

In this article I have explored what computational methods exist for studying iconic images, and where these methods are lacking. I have positioned iconic images as images for which the difference between the interpretation by an observer and what can be (automatically) extracted from the visual data is greatest. Automatically representing and extracting meaning is one of the ambitions of Artificial Intelligence, but as it stands there is a clear lack of humanlike understanding (Mitchell, 2019). Thus, rather than trying to present a be-all and end-all approach for iconic images, I have explored a range of computational methods that can be used to study aspects of iconic images, while describing their limitations. Tackling these limitations, and dedicating effort to solving the issues surrounding polyvocal meanings, will be necessary if computational methods are to meaningfully contribute to the study of visual culture.

Nevertheless, by leveraging computational techniques, we can perform tasks that would otherwise require insurmountable amounts of manual labour, as well as offer new opportunities and insights which can feed back into existing research into iconic images. Computational methods for automatic content recognition are increasingly used to support Digital Humanities research (Arnold and Tilton, 2019). Similarly, Art History has become a common application area for computational stylometry (Johnson *et al.*, 2008; van Noord and Postma, 2017) and motifs discovery (Seguin *et al.*, 2016; Shen *et al.*, 2019). Characteristic to these developments is that the underlying technologies are primarily developed for use in commercial systems, often with the aim of working 'well enough' (Mitchell, 2019, p. 2), rather than providing insight into the wide variety of meanings the visual data might have for different users. Hence, in the introduction I proposed a modified version of the semantic gap, to focus on multiple interpretations, cultural groups, and the inclusion of a notion of time:

The cultural gap is the lack of coincidence between the information that one can extract from the visual data and the interpretations that the same data have for cultural groups across time.

Arguably, the scope of the cultural gap is too narrow in that it still centres around visual data. However, in weighing the importance between including contextual data and not massively expanding the scope of the problem, I chose to stay closer to the existing paradigm for CV. In this paradigm, CV is but a module in a larger system, which for instance can be plugged into a robot which already has functioning systems to move and interact with its environment. Nevertheless, in confronting this gap (or the semantic gap for that matter), contextual data are not explicitly excluded; in fact, their use is to be encouraged. But, as it stands, incorporating both the visual data and the knowledge and background of an observer independent of the context is not possible. Therefore, I consider it overly ambitious to draw up a route towards full interpretation of meaning. Moreover, in building towards a fully fledged human in the loop approach, it seems (initially) desirable to have distinct components that deal with individual modalities and that can be used both *in vivo* and *in situ*. In this perspective, CV is a module that can be plugged into a visual culture research workflow, to support humanists in exploring the cultural gap, but that should be complemented with similar modules for other modalities.

In moving forward, there are several obstacles for exploring the cultural gap. The first and foremost being the lack of open and accessible data from multiple sources. Whilst more GLAM institutions are digitizing and publishing their collections, aggregating and connecting collections across institutions remains challenging. Despite progress in this area with efforts such as Linked (Open) Data (Marden *et al.*, 2013; Dijkshoorn *et al.*, 2018) and the International Image Interoperability Framework (IIIF) (Snydman *et al.*, 2015), that aim to increase interoperability and access, there is a clear lack of accessible, large-scale datasets of visual culture from multiple time periods and domains. Existing large-scale datasets typically focus on (Creative commons) 'Internet images' of objects and places (Lin *et al.*, 2014; Zhou *et al.*, 2014; Russakovsky *et al.*, 2015), *or* on creative works and heritage material (Mensink and Van Gemert, 2014; Wilber *et al.*, 2017; Lincoln *et al.*, 2020). Moreover, datasets, and especially datasets of 'Internet images', are (rightfully) under pressure of meeting ethical standards in terms of which images to include and which to exclude (Prabhu and Birhane, 2020). Yet, ethical standards for datasets for use in commercial applications might (and should) be different from those used to study visual culture. Determining what should (and thus should not) be included in a dataset suitable for training algorithms for visual culture analysis is an open and challenging question. Moreover, as iconic images frequently depict tragedy and human suffering, it is unavoidable that a dataset with iconic images contains images that are not suitable for other purposes.

Another obstacle to exploring the cultural gap, and hence why it is meaningful to make this gap explicit, is the lack of suitable algorithms. With existing technology, it is not possible to automatically provide rich interpretations of iconic images or other culturally complex visual materials, and in all likelihood, this will not be possible in the coming years either. Yet, by defining and exploring the limitations of what is possible we can set a course for future developments. Concrete initial steps for these developments should focus on broadening the historical and geographical scope of existing methods, such that they remain reliable and robust when confronted with data that differs from what is now predominantly available in

visual datasets. As it stands, existing visual datasets used in CV primarily contain modern images taken in wealthy countries scraped from the Internet (de Vries *et al.*, 2019; Paullada *et al.*, 2020; Prabhu and Birhane, 2020). As this bias in the training data extends to trained models, the perspective obtained with such models is highly restrictive when applied to non-standard (from a CV perspective) visual material, as also demonstrated in Fig. 1. A natural progression from this would be to expand the vocabulary of CV models to include terms and concepts that are historical or from different cultures. Once CV models are robust enough to deal with diverse data, and are able to recognize a wide range of visual concepts reliably, it will be key to focus on interactive and human in the loop systems. Rather than expecting that machines will bridge the gap on their own, through interactive systems we can meet them halfway, while still benefiting from their ability to scale.

Acknowledgements

This article has benefited from comments by Julia Noordegraaf, Thomas Smits, and Melvin Wevers. Additionally, the author thanks the anonymous reviewers for their time and highly insightful and constructive comments.

Funding

The research described in this paper was made possible by the CLARIAH-PLUS project (clariah.nl) financed by the Dutch Research Council NWO (Grant 184.034.023).

Conflict of Interest

The author declares no conflict of interest.

References

Agre, P. (1997). Toward a critical technical practice: lessons learned in trying to reform AI. In Bowker, G. C., Star, S. L., Turner, W. and Gasser, L. (eds), *Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide*, Mahwah, NJ: Erlbaum, pp. 131–58.

Alameda-Pineda, X., Redi, M., Otterbacher, J., Sebe, N., and Chang, S.-F.(2020). FATE/MM 20: 2nd international workshop on fairness, accountability, transparency and ethics in multimedia. In *Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, New York, NY, USA, pp. 4761–62. DOI: 10.1145/3394171.3421896

Arandjelović, R. and Zisserman, A. (2019). Object discovery with a copy-pasting GAN. *arXiv:1905.11369 [cs]*.

Arnold, T. and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, **34(Suppl 1)**: i3–i16.

Azar, M., Cox, G., and Impett, L. (2021). Introduction: ways of machine seeing. *AI & SOCIETY*, **36**(4): 1093–1104.

Bateman, J., Tseng, C.-I., Seizov, O., Jacobs, A., Lüdtke, A., Müller, M. G., and Herzog, O. (2016). Towards next-generation visual archives: image, film and discourse. *Visual Studies*, **31**(2): 131–54.

Beattie, G. and Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, **18**(4): 438–62.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8): 1798–1828.

Berg, T. L. and Berg, A. C. (2009). Finding iconic images. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, *2009, pp. 1-8, DOI: 10.1109/CVPRW.2009.5204174.*

Berg, T. L. and Forsyth, D. (2007). *Automatic Ranking of Iconic Images*. Technical Report. Berkeley, CA: University of California.

Binder, W. and Jaworsky, B. N. (2018). Refugees as icons: culture and iconic representation. *Sociology Compass*, **12**(3): e12568.

Biten, A. F., Gomez, L., Rusinol, M., and Karatzas, D. (2019). Good news, everyone! Context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019, pp. 12466–12475.

Budig, B., Van Dijk, T. C., and Wolff, A. (2016). Matching labels and markers in historical maps: an algorithm with interactive postprocessing. *ACM Transactions on Spatial Algorithms and Systems* **2**(4): 1–13.

Buolamwini, J. and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender

classification. In *Conference on Fairness, Accountability and Transparency*, New York, NY, USA, February 2018, pp. 77–91.

Candela, G., Sáez, M. D., Escobar Esteban, M. and Marco-Such, M. (2020). Reusing digital collections from GLAM institutions. *Journal of Information Science*,. doi: 0165551520950246.

Chang, J., Wang, L., Meng, G., Xiang, S., and Pan, C. (2017). Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, Honolulu, HI, USA, July 2017, pp. 5879–87.

Chávez Heras, D. and Blanke, T. (2020). On machine vision and photographic imagination. *AI & SOCIETY*, **36**(4): 1–13.

Chum, O., Philbin, J., Isard, M., and Zisserman, A. (2007). Scalable near identical image and shot detection. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, July 2007, Amsterdam, The Netherlands, pp. 549–56.

Chung, J. S., Arandjelović, R., Bergel, G., Franklin, A., and Zisserman, A. (2015). Re-presentations of art collections. In Agapito, L., Bronstein, M. M. and Rother, C. (eds), *Computer Vision - ECCV 2014 Workshops. Lecture Notes in Computer Science*. New York, USA: Springer International Publishing, pp. 85–100.

Cohen, A. A., Boudana, S., and Frosh, P. (2018). You must remember this: iconic news photographs and collective memory. *Journal of Communication*, **68**(3): 453–79.

Criminisi, A., Kemp, M., and Zisserman, A. (2005). Bringing pictorial space to life: computer techniques for the analysis of paintings. *In A. Bentkowska-Kafel, T. Cashen, H. Gardiner (eds.), Digital Art History: A Subject in Transition*. Brisol, UK: Intellect Books.

Crowley, E. and Zisserman, A. (2013). Of gods and goats: weakly supervised learning of figurative art. In *Proceedings British Machine Vision Conference 2013*, pp. 1–11.

da Gama Leitao, H. C. and Stolfi, J. (2002). A multiscale method for the reassembly of two-dimensional fragmented objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(9): 1239–51.

Dahmen, N. S., Mielczarek, N., and Perlmutter, D. D. (2018). The influence-network model of the photojournalistic icon. *Journalism & Communication Monographs*, **20**(4): 264–313.

Dahmen, N. S. and Miller, A. (2012). Redefining iconicity: a five-year study of visual themes of hurricane Katrina. *Visual Communication Quarterly*, **19**(1): 4–19.

Dahmen, N. S. and Morrison, D. D. (2016). Place, space, time: media gatekeeping and iconic imagery in the digital and social media age. *Digital Journalism*, **4**(5): 658–78.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, **40**(2): 1–60.

de Vries, T., Misra, I., Wang, C., and van der Maaten, L. (2019). Does object recognition work for everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, June 2019, pp. 52–59.

Deza, A. and Parikh, D. (2015). Understanding image virality. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2015, pp. 1818–26.

Dietrich, F. (1986). Visual intelligence: the first decade of computer art (1965–1975). *Leonardo* **19**(2): 159–69.

Dijkshoorn, C., Aroyo, L., van Ossenbruggen, J., and Schreiber, G. (2018). Modeling cultural heritage data for online publication. *Applied Ontology*, **13**(4): 255–71.

Drew, S. and Guillemin, M. (2014). From photographs to findings: visual meaning-making and interpretive engagement in the analysis of participant-generated images. *Visual Studies*, **29**(1): 54–67.

Drucker, J. (2020). *Visualization and Interpretation: Humanistic Approaches to Display*. Cambridge, MA, USA: MIT Press.

Ehinger, K. A., Xiao, J., Torralba, A., and Oliva, A. (2011). Estimating scene typicality from human ratings and image features. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Boston, MA, USA, July 2011, pp. 1–6.

Elgammal, A., Mazzone, M., Liu, B., Kim, D., and Elhoseiny, M. (2018). The shape of art history in the eyes of the machine. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). *AAAI Press, Article 266, pp.2183–2191*.

Figueiredo, F., Almeida, J. M., Gonçalves, M. A. and Benevenuto, F. (2014). On the dynamics of social media popularity: a YouTube case study. *ACM Transactions on Internet Technology (TOIT)*, **14**(4): 1–23.

Forsyth, D. A. and Ponce, J. (2012). *Computer Vision: A Modern Approach*. London, UK: Pearson.

Foss, S. K. (2005). Theory of visual rhetoric. In *Handbook of Visual Communication: Theory, Methods, and Media*.

Freeman, W. T. and Roth, M. (1995). Orientation histograms for hand gesture recognition. In *International Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 1995, pp. 296–301.

Gairola, S., Shah, R., and Narayanan, P. J. (2020). Unsupervised image style embeddings for retrieval and recognition tasks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Snowmass Village, CO, USA, March 2020, pp. 3281–89.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, July 2016, pp. 2414–23.

Gehl, R. W., Moyer-Horner, L., and Yeo, S. K. (2017). Training computers to see internet pornography: gender and sexual discrimination in computer vision science. *Television & New Media*, **18**(6): 529–47.

Goel, S., Anderson, A., Hofman, J., and Watts, D. J. (2016). The structural virality of online diffusion. *Management Science*, **62**(1): 180–96.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the V in VQA matter: elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017, pp. 6325–34.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In Leibe, B., Matas, J., Sebe, N. and Welling, M.(eds), *Computer Vision – ECCV 2016. Lecture Notes in Computer Science*. Cham: Springer International Publishing, pp. 87–102.

Hansen, L. (2015). How images make world politics: international icons and the case of Abu Ghraib. *Review of International Studies*, **41**(2), 263–88.

Hariman, R. and Lucaites, J. L. (2007). *No Caption Needed: Iconic Photographs, Public Culture, and Liberal Democracy*. Chicago, IL, USA: University of Chicago Press.

Hernández, J. C. (2019). 30 Years after Tiananmen, 'Tank Man' remains an icon and a mystery. *New York, USA: The New York Times*.

Hosseini, K., McDonough, K., van Strien, D., Vane, O., and Wilson, D. C. S. (2021). Maps of a nation? The digitized ordnance survey for new historical research. *Journal of Victorian Culture*, **26**(2): 284–99.

Hubbert, J. (2014). Appropriating iconicity: why Tank Man still matters. *Visual Anthropology Review*, **30**(2): 114–26.

Hu, T., Mettes, P., Huang, J.-H., and Snoek, C. (2019). SILCO: show a few images, localize the common object. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, October 2019, pp. 5066–75.

Hyvönen, E., Lindquist, T., Törnroos, J., and Mäkelä, E. (2012). History on the semantic web as linked data-an event gazetteer and timeline for World War I. In *Proceedings of CIDOC*, Helsinki, Finland, June 2012, p. 12.

Ibrahim, Y. (2017). Facebook and the Napalm Girl: reframing the iconic as pornographic. *Social Media+ Society*, **3**(4), 2056305117743140.

Impett, L. and Süsstrunk, S. (2016). Pose and pathosformel in Aby Warburg's Bilderatlas. In *ECCV Workshops*. Amsterdam, The Netherlands, October 2016.

Impett, L. L., Bell, P., Seguin, B., and Ommer, B. (2018). Beyond image search: computer vision in western art history. In *DH*, Mexico City, Mexico, September 2018, pp. 73–75.

Isola, P., Parikh, D., Torralba, A., and Oliva, A. (2011). Understanding the intrinsic memorability of images. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. and Weinberger, K. W. (eds), *Advances in Neural Information Processing Systems 24*. Red Hook, NY, USA: Curran Associates, Inc., pp. 2429–37.

Jain, A. K. and Li, S. Z. (2011). *Handbook of Face Recognition*. Vol. **1**. London, UK: Springer.

Jiang, Y.-G. (2012). Super: towards real-time event recognition in internet videos. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, Hong Kong, Hong Kong, June 2012, pp. 1–8.

Jiang, Y.-G., Ye, G., Chang, S.-F., Ellis, D., Loui, A. C. (2011). Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, Trento Italy, April 2011, pp. 1–8.

Jo, E. S. and Gebru, T. (2020). Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT 27-30 January 2020*, Association for Computing Machinery, New York, NY, USA, pp. 306–16.

Johnson, C. R., Hendriks, E., Berezhnoy, I. J., Brevdo, E., Hughes, S. M., Daubechies, I., Li, J., Postma, E., and Wang, J. Z. (2008). Image processing for artist identification. *IEEE Signal Processing Magazine*, **25**(4): 37–48.

Kampel, M. and Melero, F. J. (2003). Virtual vessel reconstruction from a fragment's profile. In *Proceedings of the*

4th International Conference on Virtual Reality, Archaeology and Intelligent Cultural Heritage, VAST'03, Eurographics Association, Goslar, DEU, pp. 79–88.

Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., and Winnemoeller, H. (2014). Recognizing image style. In *Proceedings of the British Machine Vision Conference*. Nottingham, UK, September 2014, pp. 1–11.

Ke, Y., Sukthankar, R., and Huston, L. (2004). An efficient parts-based near-duplicate and sub-image retrieval system. In *Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA October 10-16, 2004*, Association for Computing Machinery, New York, NY, USA, pp. 869–76.

Khosla, A., Das Sarma, A., and Hamid, R. (2014). What makes an image popular?. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW April 2014, New York, NY, USA, pp. 867–76.

Khosla, A., Raju, A. S., Torralba, A., and Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 2015, pp. 2390–98.

Kleppe, M., Elliott, D., and Faber, W. J. (2017). KBK-1M - Koninklijke Bibliotheek Kranten – 1 Miljoen.

Kleppe, M. (2013). *Canonieke Icoonfoto's. De Rol van (Pers)Foto's in de Nederlandse Geschiedschrijving*. Ph.D. thesis, Erasmus University Rotterdam.

Kroes, R., Orvell, M., and Nadel, A. (2011). The ascent of the falling man: establishing a picture's iconicity. *Journal of American Studies*, **45**(4): E47.

Lang, S. and Ommer, B. (2018). Attesting similarity: supporting the organization and study of art image collections with computer vision. *Digital Scholarship in the Humanities*, **33**(4): 845–56.

Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., and Weld, D. S. (2020). The newspaper navigator dataset: extracting and analyzing visual content from 16 million historic newspaper pages in chronicling America. *arXiv: 2005.01583 [cs]*.

Le, P. and Titov, I. (2019). Distant learning for entity linking with automatic noise detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, August 2019, Florence, Italy, pp. 4081–90.

Li, J., Yao, L., Hendriks, E., and Wang, J. Z. (2011). Rhythmic brushstrokes distinguish van Gogh from his contemporaries: findings via automated brushstroke extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(6): 1159–76.

Lincoln, M., Corrin, J., Davis, E., and Weingart, S. (2020). CAMPI: computer-aided metadata generation for photo archives initiative. Carnegie Mellon University. Preprint. https://doi.org/10.1184/R1/12791807.v2.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: common objects in context. In *European Conference on Computer Vision*, Zurich, Switzerland, September 2014, pp. 740–55.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (Retrieved 15 August 2018).

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 152–59.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., and Milford, M. J. (2015). Visual place recognition: a survey. *IEEE Transactions on Robotics*, **32**(1): 1–19.

Lucaites, J. L. and Hariman, R. (2001). Visual rhetoric, photojournalism, and democratic public culture. *Rhetoric Review*, **20**(1/2): 37–42.

Manovich, L. (1994). The engineering of vision and the aesthetics of computer art. *ACM SIGGRAPH Computer Graphics*, **28**(4): 259–63.

Manovich, L. (2018). The science of culture? Social computing, digital humanities and cultural analytics. Journal of Cultural Analytics 1(1). DOI: 10.22148/16.004.

Marden, J., Li-Madeo, C., Whysel, N., and Edelstein, J. (2013). Linked open data for cultural heritage: evolution of an information technology. In *Proceedings of the 31st ACM International Conference on Design of Communication*, Greenville, NC, USA, September 2013, pp. 107–12.

Marks, T. (2019). Pineapple mania: art history's fixation with an exotic fruit. https://artuk.org/discover/stories/pineapple-mania-art-historys-fixation-with-an-exotic-fruit (accessed 22 January 2021).

Masson, E., Olsen, C. G., van Noord, N., and Fossati, G. (2020). Exploring digitised moving image collections: The SEMIA Project, visual analysis and the turn to abstraction. *Digital Humanities Quarterly*, **14**(4): 1–14.

Matud, M. P., Espinosa, I., and Wangüemert, C. R. (2021). Women and men portrayal on television news: a study of

Spanish television newscast. *Feminist Media Studies*, **21**(2): 298–314.

Mayer, V., Banks, M. J. and Caldwell, J. T. (2009). *Production Studies: Cultural Studies of Media Industries*. New York, USA: Routledge.

Mensink, T. and Van Gemert, J. (2014). The Rijksmuseum challenge: museum-centered visual recognition. In *Proceedings of International Conference on Multimedia Retrieval*, Glasgow, Scotland, April 2014, pp. 451–54.

Meuzelaar, A. (2014). *A History of the Representation of Muslims on Dutch Television*. Ph.D. thesis, Universiteit van Amsterdam.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, Inioluwa D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, January 2019, pp. 220–29.

Mitchell, M. (2019). Artificial intelligence hits the barrier of meaning. *Information*, **10**(2): 51.

Mohamed, S., Png, M.-T., and Isaac, W. (2020). Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*.

Moreira, D., Bharati, A., Brogan, J., Pinto, A., Parowski, M., Bowyer, K. W., Flynn, P. J., Rocha, A., and Scheirer, W. J. (2018). Image provenance analysis at scale. *IEEE Transactions on Image Processing*, **27**(12): 6109–23.

Mortensen, M. (2017). Constructing, confirming, and contesting icons: the Alan Kurdi imagery appropriated by# humanitywashedashore, Ai Weiwei, and Charlie Hebdo. *Media, Culture & Society*, **39**(8): 1142–61.

Mortensen, M., Allan, S., and Peters, C. (2017). The iconic image in a digital age. *Nordicom Review*, **38**: 71–86.

Münster, S. and Terras, M. (2020). The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures. *Digital Scholarship in the Humanities*, **35**(2): 366–89.

Murray, N., Marchesotti, L., and Perronnin, F. (2012). AVA: a large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012, pp. 2408–15.

Nahon, K. and Hemsley, J. (2013). *Going Viral* (1st. ed.). Cambridge, UK: Polity Press.

Noll, A. M. (1966). Human or machine: a subjective comparison of Piet Mondrian's "Composition with Lines"(1917) and a computer-generated picture. *The Psychological Record*, **16**(1): 1–10.

Offert, F. and Bell, P. (2021). Perceptual bias and technical metapictures: critical machine vision as a humanities challenge. *AI & SOCIETY* **36**, 1133–44.

Olesen, C. G. (2015). Formalising digital formalism: an interview with Adelheid Heftberger and Matthias Zeppelzauer about the Vienna Vertov Project. In de Rosa, M. and Fales, L. (eds), *Shifting Layers: New Perspectives in Media Archaeology across Digital Media and Audiovisual Arts*. Milan: Mimesis International.

Papert, S. A. (1966). The Summer Vision Project. https://dspace.mit.edu/handle/1721.1/6125 (accessed 12 October 2020).

Parisi, L. (2020). Negative optics in vision machines. *AI & SOCIETY*, **36**(4): 1–13.

Parmeggiani, P. (2009). Going digital: using new technologies in visual sociology. *Visual Studies*, **24**(1): 71–81.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2021). Data and its (dis)contents: a survey of dataset development and use in machine learning research. Patterns, 2(11): 100336.

Perlmutter, D. D. (1994). Visual historical methods: problems, prospects, applications. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **27**(4): 167–84.

Perlmutter, D. D. (1998). *Photojournalism and Foreign Policy: Icons of Outrage in International Crises*. Westport, CT, USA: Praeger Publishers.

Perlmutter, D. D. (2005). Photojournalism and foreign affairs. *Orbis*, **49**(1): 109–22.

Perlmutter, D. D. and Wagner, G. L. (2004). The anatomy of a photojournalistic icon: marginalization of dissent in the selection and framing of 'a death in Genoa'. *Visual Communication*, **3**(1): 91–108.

Prabhu, V. U. and Birhane, A. (2021). Large image datasets: a pyrrhic win for computer vision?. IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1536–46.

Bocyte, R. and Oomen, J. (2020). Content adaptation, personalisation and fine-grained retrieval: applying AI to support engagement with and reuse of archival content at scale. In *Artificial Intelligence and Digital Heritage (ARTIDIGH)*, Valletta, Malta, February 2020, pp. 506–11.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., and Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, **115**(3): 211–52.

Ryu, H. J., Mitchell, M., and Adam, H. (2017). Improving smiling detection with race and gender diversity. *arXiv: 1712.00193*

Schroeder, J. E. (2006). Critical visual analysis. In Belk, R. (ed.), *Handbook of Qualitative Research Methods in Marketing*. Aldershot, UK: Edward Elgar. pp. 303–21.

Seguin, B. L. A., Striolo, C., di Lenardo, I., and Kaplan, F. (2016). Visual link retrieval in a database of paintings. In Hua, G. and Jégou, H. (eds.), Computer vision – ECCV 2016 workshops. New York, USA: Springer International Publishing, pp. 753–67.

Settles, B. (2009). *Active learning literature survey*. Technical report. Madison, WI: Department of Computer Sciences, University of Wisconsin-Madison.

Shen, X., Efros, A. A., and Aubry, M. (2019). Discovering visual patterns in art collections with spatially-consistent feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019, pp. 9278–87.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. *CVPR 2011*, pp. 1297-1304, doi: 10.1109/CVPR.2011.5995316.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(12): 1349–80.

Smits, T. (2017). Illustrations to photographs: using computer vision to analyse news pictures in Dutch newspapers, 1860-1940. In *Book of Abstracts of DH2017*, Alliance of Digital Humanities Organizations, Montréal, Canada.

Snydman, S., Sanderson, R., and Cramer, T. (2015). The International Image Interoperability Framework (IIIF): a community & technology approach for web-based images. In *Archiving Conference*, Vol. 2015, Society for Imaging Science and Technology, pp. 16–21.

Spratt, M., Peterson, A., and Lagos, T. (2005). Of photographs and flags: uses and perceptions of an iconic image before and after September 11, 2001. *Popular Communication*, **3**(2): 117–36.

Stork, D. G. (2009). Computer vision and computer graphics analysis of paintings and drawings: an introduction to the literature. In Jiang, X. and Petkov, N. (eds), *Computer Analysis of Images and Patterns. Lecture Notes in Computer Science*. Berlin, Heidelberg: pp. 9–24.

Susarla, A., Oh, J.-H., and Tan, Y. (2012). Social networks and the diffusion of user-generated content: evidence from YouTube. *Information Systems Research*, **23**(1): 23–41.

Theisen, W., Brogan, J., Thomas, P. B., Moreira, D., Phoa, P., Weninger, T., and Scheirer, W. (2021). Automatic discovery of political meme genres with diverse appearances. Proceedings of the International AAAI Conference on Web and Social Media, *15(1), 714-26*.

Thomas, C. (2020). Modeling visual rhetoric and semantics in multimedia. Doctoral Dissertation, University of Pittsburgh.

Torii, A.,, Sivic, J., Pajdla, T., and Okutomi, M. (2013). Visual place recognition with repetitive structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 883–90.

Toshev, A. and Szegedy, C. (2014). Deeppose: human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 1653–60.

Trewin, S.,, Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N., and Manser, E. (2019). Considerations for AI fairness for people with disabilities. *AI Matters*, **5**(3), 40–63.

Uhl, J. H. and Duan, W. (2021). Automating information extraction from large historical topographic map archives: new opportunities and challenges. In Werner, M. and Chiang, Y.-Y. (eds), *Handbook of Big Geospatial Data*. Cham: Springer International Publishing, pp. 509–22.

Valeonti, F., Terras, M., and Hudson-Smith, A. (2019). How open is OpenGLAM? Identifying barriers to commercial and non-commercial reuse of digitised art images. *Journal of Documentation*, **76**(1), 1–26.

van den Herik, H. J. and Postma, E. O. (2000). Discovering the visual signature of painters. In Kasabov, N. (ed.), *Future Directions for Intelligent Systems and Information Sciences. Studies in Fuzziness and Soft Computing*. Heidelberg: Physica, pp. 129–47.

van der Hoeven, R. (2019). *The Global Visual Memory: A Study of the Recognition and Interpretation of Iconic and Historical Photographs*. Ph.D. thesis, Universiteit Utrecht.

van der Maaten, L. J. P., Boon, P., Lange, G., Paijmans, J. J., and Postma, E. (2006). Computer vision and machine learning for archaeology. In *Proceedings of CAA-2006. Fargo, ND, USA, April 2006, pp. 1-9*.

van Erp, M. and de Boer, V. (2020). A polyvocal and contextualised semantic web. In: Verborgh R. et al. (eds) *The Semantic Web. ESWC 2021. Lecture Notes in Computer*

Science, vol 12731. Cham: Springer. DOI: 10.1007/978-3-030-77385-4_30.

van Leeuwen, T. (2004). Semiotics and iconography. In van Leeuwen, T. and Jewitt, C. (eds), *The Handbook of Visual Analysis*. London: SAGE Publications Ltd, pp. 92–118.

van Noord, N., Hendriks, E., and Postma, E. (2015). Toward discovery of the artist's style: learning to recognize artists by their artworks. *IEEE Signal Processing Magazine* **32**(4): 46–54.

van Noord, N., Olesen, C., Ordelman, R., and Noordegraaf, J. (2021). Automatic annotations and enrichments for audiovisual archives. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, Virtual, February 2021, pp. 633–40.

van Noord, N. and Postma, E. (2017). Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognition*, **61**: 583–92.

Vis, F. and Goriunova, O. (2015). The iconic image on social media: a rapid research response to the death of Aylan Kurdi. *Manchester, UK: Visual Social Media Lab*.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal Law and Technology*, **31**: 841.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, Australia, April 2013, pp. 3551–58.

Warnke, M. and Dieckmann, L. (2016). Prometheus meets Meta-Image: implementations of Aby Warburg's methodical approach in the digital era. *Visual Studies*, **31**(2): 109–20.

Weinman, J., Chen, Z., Gafford, B., Gifford, N., Lamsal, A., and Niehus-Staab, L. (2019). Deep Neural Networks for Text Detection and Recognition in Historical Maps. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, September 2019, pp. 902–09.

Weng, L., Flammini, A., Vespignani, A., and Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, **2**: 335.

Wevers, M. and Smits, T. (2020). The visual digital turn: using neural networks to study historical images. *Digital Scholarship in the Humanities*, **35**(1): 194–207.

Wevers, M., Smits, T., and Impett, L. (2018). Modeling the genealogy of Imagetexts: studying images and texts in conjunction using computational methods. In *DH*, Mexico City, Mexico, June 2020, pp. 684–85.

Weyand, T., Araujo, A., Cao, B., and Sim, J. (2020). Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Virtual, June 2020, pp. 2575–84.

Wilber, M. J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., and Belongie, S. (2017). Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, October 2017, pp. 1202–11.

Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*, New York, NY, USA, June 2016, pp. 478–87.

Yang, T.-I., Torget, A. and Mihalcea, R. (2011). *Topic Modeling on Historical Newspapers: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, USA, June 2011, pp. 96–104.

Yarlagadda, P., Monroy, A., Carqué, B., and Ommer, B. (2013). Towards a computer-based understanding of medieval images. In Bock, H., Jäger, W. and Winckler, M. (eds) Scientific Computing and Cultural Heritage. Contributions in Mathematical and Computational Sciences, vol 3. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-28021-4_10.

Zahálka, J. and Worring, M. (2014). Towards interactive, intelligent, and integrated multimedia analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Paris, France, October 2014, pp. 3–12.

Zamir, A. R. and Shah, M. (2010). Accurate image localization based on Google maps street view. In *European Conference on Computer Vision*, Crete, Greece, September 2010, pp. 255–68.

Zarzycka, M. and Kleppe, M. (2013). Awards, archives, and affects: tropes in the World Press Photo contest 2009–11. *Media, Culture & Society*, **35**(8): 977–95.

Zhang, Q.-S. and Zhu, S.-C. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, **19**(1): 27–39.

Zhang, Y., Larlus, D., and Perronnin, F. (2014). What makes an image iconic? A fine-grained case study. *arXiv:1408.4325 [cs]*.

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., Newton, MA, USA.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems (NeurIPS Proceedings 2014)*, Montreal, Canada, December 2014, pp. 487–95.

Notes

1 'OK hand sign added to list of hate symbols'. https://www.bbc.com/news/newsbeat-49837898

2 https://www.loc.gov/preservation/digital/formats/fdd/fdd000146.shtml

3 https://en.wikipedia.org/wiki/Barack_Obama_\%22Hope\%22_poster