DHQ: Digital Humanities Quarterly
2020
Volume 14 Number 4

Exploring Digitised Moving Image Collections: The SEMIA Project, Visual Analysis and the Turn to Abstraction

Eef Masson <E_dot_Masson_at_rathenau_dot_nl>, University of Amsterdam / Rathenau Institute
Christian Gosvig Olesen <c_dot_g_dot_olesen_at_uva_dot_nl>, Utrecht University / University of Amsterdam
Nanne van Noord <n_dot_j_dot_e_dot_vannoord_at_uva_dot_nl>, University of Amsterdam / Netherlands Institute for Sound and Vision
Giovanna Fossati <G_dot_Fossati_at_uva_dot_nl>, University of Amsterdam / Eye Filmmuseum

Abstract

In recent years, efforts to unlock digitized moving image collections have focused primarily on the retrieval of collection items through semantic descriptors: keywords or other labels produced either manually, or as (semi-)automatically generated metadata. As a result, access to digital archives is still governed overwhelmingly by a logic of search. In practice, this means that users not only need to know what they are looking for, but are also constrained by the interpretive frameworks informing the materials' labelling. Arguably, this poses restrictions on what they can find, how they can interrelate collection objects, and ultimately, how they can reuse or reinterpret collections. Taking such issues as its starting point, the *Sensory Moving Image Archive* project (SEMIA) investigated how visual analysis tools can help enable more exploratory forms of engaging with digital archives. In doing so, it focused on sensory features, which are essential to users' experiences of audiovisual heritage objects but inadequately captured by verbal description.

In this article, we discuss the project's rationale and its early results. First, we place SEMIA in a recent history of visual analysis for media scholarly research, specifying how it both builds on and departs from this history (also in the epistemic sense). Subsequently, we provide more details about the project's approach to image feature extraction and discuss some analysis results. In our conclusions, we confront those results with what we had initially hoped to gain by applying computer vision methods for enabling access to collections.

1

> *One senior curator said that some of museum staff [sic] were skeptical of the project at first. "We would get an email from Wes asking, 'Do you have a list of green objects? Could you send us a list of everything you have that is yellow?' Our data system does not have these categories."* [Brown 2018]

Introduction

2

Until late April of 2019, visitors of the Kunsthistorisches Museum in Vienna could drop in on the exhibit "Spitzmaus Mummy in a Coffin and Other Treasures",[1] co-curated by filmmaker Wes Anderson and designer Juman Malouf.[2] The exhibit consisted of 430 relatively obscure objects selected from a collection of more than four million, spanning over 5,000 years. In putting the exhibit together, the curators had relied heavily on the museum's curatorial staff, who had helped them navigate the collection [Brown 2018]. Without such assistance, their task arguably would have been impossible to perform. The reason is that while most museums these days work with searchable digital catalogues (or collection management systems), the descriptions those systems contain typically neglect certain aspects of the objects

represented. For example, they usually do not contain specifications of such sensory features as colour – precisely the kind which, as the epigraph to this piece suggests, Anderson and his colleague were interested in.

3 In a more general sense, this holds true also for most moving image archives. Oftentimes, such institutions house collections of many thousands of films or television episodes, composed in turn of millions of discrete images. Typically, the sensory characteristics of those objects barely feature in catalogue descriptions. While some entries contain information, either at the title or the fragment level, about the colour or sound systems used, this information tends to be fragmentary. Moreover, further specifics about the films' or episodes' visual features are usually absent.

4 In recent years, audiovisual heritage institutions have invested much time and resources into digitising their collections, so as to enable various kinds of reuse. Yet in spite of this, the above situation is largely unchanged. So far, attempts to improve usability have focused primarily on the searchability of collections and the retrieval of collection items through (linked) metadata. Therefore, access to digital archives is overwhelmingly governed, even today, by a logic of search – one dominant in practices of information retrieval more in general [Whitelaw 2015]. Search relies on the use of semantic descriptors: keywords or other labels produced either manually, or as (semi-)automatically generated metadata. Apart from being labour-intensive to produce, such descriptors are also highly selective. In the case of audiovisual materials, for instance, they are usually limited to facts about production, or about the people, events and geographic locations they feature. Arguably, they serve the needs of a rather limited range of reuse practices; for instance, the production of documentaries, or scholarship in socio-political history, media production or (to a lesser extent) certain forms of aesthetic analysis. The design of an exhibit like Anderson and Malouf's, but also other kinds of more creative reuse, require different kinds of information.

5 For users, the selectiveness of catalogue descriptions poses two important problems. On the one hand, it forces them to search collections on the basis of prior interpretations, and from the perspective of those who catalogued them – rather than to more freely explore them. On the other, it prevents them from relying in the process on features that are essential to their experience of heritage objects, but inadequately captured through verbal description; for example, visual features such as colour, but also shape or movement. Such characteristics are particularly significant for historic (moving) images, as those are valued not only for the information they hold, but also for their look and "feel" [Delpeut 1999] (cf. [Dudley 2010]).

6 The research project The Sensory Moving Image Archive (SEMIA): Boosting Creative Reuse for Artistic Practice and Research[3] departs from the observation that this situation impedes the work (and play) of a range of potential users. In response to this problem, it raises the question how sensory object features can be mobilised as the driving criterion to explore – rather than search – digitised audiovisual collections. "Users," in this context, are filmmakers or exhibition designers, but also scholars. It has been argued, indeed, that the work of researchers may benefit from modes of access that do not (solely) rely on search and retrieval of single items but afford a more explorative form of browsing [Flanders 2014], ideally also drawing on the sensory relations between discrete items within collections.[4] Beneficiaries of such an approach are scholars already concerned in their work, for instance, with colour palettes, or patterns of movement in historical film – whether considered in terms of their technological preconditions (as in Yumibe [2012]), their relation to film style and aesthetics [Heftberger et al. 2009] [Street 2012] [Street and Yumibe 2013] [Flückiger 2017] [Heftberger 2018] or from a more experiential perspective [Mazzanti 2009], for instance in terms of their haptic or synaesthetic aspects [Catanese et al. 2019]. But arguably, also others can benefit, as it can help reveal previously unanticipated patterns or relations in or between widely divergent materials, that elicit novel research questions of their own.

7 SEMIA, a two-and-a-half year project that ran until late January 2020, was a collaboration between the University of Amsterdam (with contributions from media and audiovisual heritage scholars as well as computer scientists), the Amsterdam University of Applied Sciences (specifically, experts in the domain of data visualisation and interface design), the interaction design company Studio Louter (experienced in the development of museum presentations) and two audiovisual heritage institutions: Eye Filmmuseum (focusing on film and cinematography) and the Netherlands Institute for Sound and Vision (television).[5] The team's overarching aim was to establish whether, and how, repurposing software for analysing and visualising colour, shape, visual complexity and movement might enable

alternative forms of accessing collections of moving images. To this end, it developed a prototype tool that invites users to explore collections on the basis of those features, rather than to search them through (verbal) descriptions resulting from prior interpretations of specific objects in discrete films or film sequences. In doing so, it not only sought to delay the moment in time when significance is assigned – that is, when the meaning of specific sensory features, or of the relations between them, is determined – but also to place this task in the users' own hands (compare Kuhn et al. [2013]). The tool was designed to deal with large numbers of heterogeneous materials (in terms of production date, genre, but also medium) so as to allow for the revelation of potentially surprising connections. The corpus used for testing was made up of fragments from the collections of Eye and Sound and Vision, as featured on the open access platform Open Images.[6]

8 The project consisted of two phases, whose timings partly overlapped: a first, focused on image feature extraction and analysis, and a second, concerned with the development of a "generous" interface [Whitelaw 2015], visualising the relations between fragments on the basis of analysis results. The first phase, which we elaborate on in this article, involved the use of computer vision methods.

9 In computer vision, a subdiscipline of AI, models are developed for extracting key information – so-called visual "features" – from images, so that they can subsequently be cross-referenced. In the analysis process, images are transformed into descriptions that are used in turn to classify them. In the early years of the field, methods were developed that required humans to determine which operations systems had to perform in order to produce the intended analysis results. More recently, however, methods based on machine learning, whereby computers are trained with techniques for automatic feature learning, are becoming more popular.[7]

10 In SEMIA, we used a combination of both types of methods. In what follows, we explain why this is the case, and elaborate on how the computer scientists in our team aligned their work with our overall objective of enabling new forms of exploration. In doing so, we specifically focus on how we changed the preconditions for archival reuse (the scholarly kind in particular). We are motivated by the observation that reliance on visual features and relations in accessing collections not only opens up new avenues for research, but also helps challenge current understandings of how knowledge is produced – in media and heritage studies (traditional as well as digital) and in the digital humanities more broadly.

11 In our contribution, we take a "funnel approach," gradually narrowing our focus to the specific extraction and analysis tasks carried out within the SEMIA project. First, we provide a broad outline, and discussion, of the "landscape" of visual analysis for media scholarly research, and developments in this area over time. We pay attention both to the interests and objectives of those active in the field (along with their epistemic underpinnings) and to their specific approaches or methods. The purpose of this exercise is twofold: to specify the project's place among prior efforts, and to further elucidate our overall motivation in taking it on. Subsequently, we zoom in on what feature analysis means for SEMIA: first, by looking at the general principles behind our approach to feature extraction, and then, by discussing some analysis results. In our conclusions, we confront those results with our initial intent in exploring the affordances of computer vision for providing access to collections.

Visual Analysis in Digital Scholarship, Media Art and Explorative Browsing

12 In developing a tool that supports a more unconstrained browsing of media archives than is currently available, we sought to complement existing approaches to, and methods for, the visual analysis of moving images. Those approaches and methods have emerged primarily in the context of stylometric research of the 1970s and on, and tend to be tailored to the detection of patterns in specific analytical units. In the interpretation of data, stylometric research usually adheres to semantic categories that have traditionally had relevance also for both archives and media historical research (in particular, the above-mentioned categories of director or creator, or production time). For the purposes of the SEMIA project, we needed to let go of the assumptions this implied about what is "meaningful" about collection objects.

To achieve this, we followed the line of reasoning of a recent trend in digital film and media studies scholarship that seeks to reorient visual analysis methods by drawing on artistic practices of archival moving image appropriation. Such strategies are not intent on finding patterns in preselected image units, but are geared instead towards accidental or unanticipated finds that reveal more surprising similarities – or contrasts – in audiovisual materials. Those pioneering scholars, whose work we sample below, are convinced that artistic work can inspire users *not* to approach data from the perspective of specific questions or hypotheses, but to explore them more freely, also letting go in the process of more conventional categories for interpretation. 13

In order to specify the epistemological underpinnings of our own approach, it is helpful to start off with a brief consideration of the foundational assumptions of stylometry. This will help us to subsequently explain how more recent projects in visual analysis in our field draw on this tradition, while also moving it in different directions. We end the section with some further elaboration on the appropriation-indebted trend in film and media studies, explaining how it was inspirational for us. 14

In film and media studies, the visual analysis of moving images was developed as part of the intertwining stylometric research programmes commonly referred to as “statistical style analysis” and “cinemetrics,” initiated with the pioneering work of Barry Salt and Yuri Tsivian respectively. Arguably, these programmes had their very early roots in film theory and criticism from the 1910s and 1920s, attending to the interrelations between film editing, style and perception, and gained a foothold in academic institutions in the 1970s (see Buckland [2008] and Olesen [2017] for more on those historical developments). Their objective was to discern patterns in audiovisual materials, in a way that resembles the analysis of linguistic patterns in literary computing (for instance, for the purpose of authorship attribution, for the dating of films, or for the creation of statistical profiles of directorial styles, periods or genres and their changes over time). Such research often took a deductive approach, producing data that supports stylistic analysis as a more “rigorous” alternative, or complement, to traditional hermeneutic approaches. In its first decades as a scholarly form of research, stylometry pursued its objectives primarily by manually annotating, coding and quantifying data on shot lengths and shot types in films and television materials, to subsequently relate the data thus obtained to known information (for instance production or release date, production company, genre or director) in an attempt to interpret significant patterns. 15

In recent years, as digital humanities methods have proliferated, stylometric research in media studies has become more complex in its methods, but also more varied in its interests. In the past, shot length and shot type were key parameters for analysis; more recently, however, attention is also being paid to colour, motion, (recurring) objects and aspects of visual composition. Projects such as Digital Formalism[8] (2007-2010) and ACTION[9] (2011-2013) are illustrative of this development. Digital Formalism (a collaboration of the University of Vienna, the Austrian Filmmuseum and the Vienna University of Technology) sought to analyse the complex formal characteristics of Soviet director Dziga Vertov’s films. To achieve this, it strongly relied on a logic of feature-learning, whereby relevant image information was extracted with the help of purpose-produced algorithms. This involved the analysis of high-level – that is, complex – semantic features, such as visual composition or motion composition [Zeppelzauer et al. 2012]. The ACTION project at Dartmouth College, resulting in an open-source toolkit, expanded the scope of authorship attribution research by facilitating not only the analysis of motion, but also colour and audio features; moreover, it focused on the films of twenty-four canonical directors, rather than a more homogeneous corpus consisting of work by a single maker.[10] In addition, the project relied less on purpose-produced algorithms, making use instead of existing solutions, including (but not limited to) machine learning tools [Casey and Williams 2013, 4]. This way, it also expanded stylometry’s scope in the technological sense, while it remained true to its foundational drive towards quantitative, empirical research. 16

In this respect, ACTION certainly paved the way for SEMIA. On the one hand, because the project relies to a considerable extent on techniques developed or used in the context of previous stylometric research. And on the other, because it likewise engages in the extraction and quantification of moving image data. In SEMIA, however, such extraction serves rather different purposes. Data analysis, in this case, is not done with the objective of authorship attribution or for the establishment of genre features dominant in a particular corpus or period. As previously explained, the project is focused rather on enabling exploratory browsing, affording (possibly incidental) discovery of similarities that do not neatly align with existing interpretative frameworks. For instance, similarities between collection items that do 17

not have a maker or production time in common, or visual features that can *not* easily be understood as shared stylistic elements.[11]

18 To this end, the project draws inspiration from an emerging approach to visual analysis and data visualisation in digital film and media studies scholarship – an approach that is indebted in turn to media art practice and experimental filmmaking. Kevin Ferguson, a proponent of this trend, explains that there is a tradition of experimental work in media studies that "balances between [...] new media art and digital humanities scholarship" [Ferguson 2016, 279], intent on "deforming" its object of study [Ferguson 2017]. Arguably, such work challenges (especially early) stylometry's version of visual analysis, in pursuit of "a digital humanities project that is more aleatory and aesthetic than it is formal and constrained" [Ferguson 2017]. Instead of rigorously counting and then comparing calculation results to produce historical insights into film form and its development, it highlights the occurrence of highly complex formal systems (which select images features are always part of) that may meaningfully relate to each other in multifarious ways. In doing so, it demonstrates the need to pay attention also to similarities that may not be detected if one sticks to more carefully defined analytical registers.

19 As previously mentioned, film and media scholars who proceed in this way oftentimes seek inspiration in the work of artists, and specifically, those engaged in practices of archival appropriation. History has shown that these practitioners in particular have their own contributions to make to the challenging of preconceptions underpinning scholarly analysis. At times, they even use the same analytical devices for this purpose – but in methodologically less rule-bound ways. In the last few decades, this has led to productive exchanges between academics, archivists and artists – the constellation Thomas Elsaesser once dubbed the "three A's" [Elsaesser 2010, 33] – and as such, produced novel interpretations of audiovisual materials.

20 For instance, in the 1970s, when film historians would use projectors and editing tables to come up with statistics providing insight into developments in film style, artists would use those same devices to visually explore archival films in more idiosyncratic ways. They would focus in the process on particular image details, or dwell on and contemplate specific temporal units by stretching them. Examples of this practice are the 1970s structural films of Ken Jacobs, Al Razutis or Ernie Gehr, who repurposed films from the early 1900s. Their oftentimes rather abstract works highlighted the "different" formal properties of early cinema (compared to the narrative standard of later years). In bringing those to the fore, they challenged prevalent assumptions among contemporary historians, who had in fact largely neglected early cinema in their stylistic accounts to date [Testa 1992, 33]. While scholars may not always be able to make immediate (historiographic) sense of such work – although the contemporaries of Jacobs and others ultimately did – it may invite them to look at specific visual features with fresh eyes, or from different perspectives.

21 Ultimately, the great merit of such artistic work is that it strips archival films of the categories and interpretive frameworks with which they have previously been associated – thus opening up the possibility of applying new ones. Film scholar Michael Pigott, in this context, has credited the practice with "inducing illegibility." In his view, this sort of work serves "the dual purpose (and double tension) of making the image illegible (again) and then attempting to read it" [Pigott 2015, 24]. The potential for inducing illegibility is not exclusive to structural filmmaking (a common reference point for this purpose within film studies) but can also be found in contemporary media art. Currently, there is a small, but important body of artworks that critically explore moving image data, and prove inspirational also to film and media scholars; for instance, the film data visualisation work by such artists as Jim Campbell[12] or Jason Salavon[13] [Habib 2015] [Ferguson 2017]. This work precedes contemporary digital scholarship by fifteen to twenty years, and has used different coding languages and visualisation softwares, but resulting in at times remarkably similar expressions. Likewise, artist and designer Brendan Dawes' Cinema Redux[14] project (2004) experimented with grid visualisations of classic films, inviting gallery visitors to contemplate film data visualisations as visual compositions in their own right, rather than to use them as an empirical basis for establishing patterns along well-known interpretive lines.

22 In setting up SEMIA, the project team, while familiar with the above-mentioned examples, was more directly inspired by the work of Dutch video artist Geert Mul – a long-term collaborator of heritage partner Sound and Vision. Particularly influential for the projects' approach was *Match of the Day* (2004-2008), an early example of an artwork produced with the help of algorithms for visual analysis, made up entirely of stills from satellite television images (see Figure 1). The

piece demonstrates particularly well how artists can productively exploit similarities in image features among widely heterogeneous objects, that are too fuzzy to be meaningful for the rigorous testing of hypotheses.

Figure 1. Geert Mul, *Match of the Day* (2004-2008).

To create his work, Mul used large databases of images, for which he extracted a wide variety of visual features. Those features served in turn as the basis for a matching of images at different levels of similarity. The first part of this process was conducted automatically; however, human intervention occurred when the artist selected approximate rather than identical matches to include in his work [Mul and Masson 2018]. In stylometric research, such “matches” would likely be considered errors, glitches or mismatches. But in the context of an exploratory browse through an archival collection, they are precisely the kinds of results that may yield unexpected connections or patterns, worth investigating further outside of conventional notions of authorship, genre or period. 23

The above observations informed our decision, made early on in the SEMIA project, to radically abandon those kinds of categories, as embedded in archival metadata through semantic descriptions, and to opt instead for a visual analysis approach. We did this primarily by way of experiment, and in the assumption that the explorative options it opened up would eventually prove useful primarily *in combination with* search-based approaches drawing on existing metadata. (Inducing illegibility, after all, is rarely the end of a research process, and primarily makes sense in the early, exploratory phases of study. Further on in the process, existing metadata categories may then prove productive once again.) In what follows, we discuss how we undertook this visual analysis task, paying specific attention to the choices we made in the process – conceptual as well as technical, and in light of the aforementioned principles. 24

Feature Extraction in SEMIA: A Turn towards Abstraction

25 In addition to pursuing a different set of media scholarly objectives, the SEMIA project team also sought to engender a shift in terms of the techniques used for visual analysis. In this section, we discuss the rationale behind our choice for specific feature extraction methods, and why we chose to tweak existing ones in particular ways. The connecting links between those different choices are, first, our wish to extract features that would point to unanticipated – rather than predictable – connections among objects, and second, to do so at a higher level of abstraction than is currently considered "state of the art," in light of the overwhelming focus in computer vision on the recognition of meaningful semantic entities.

26 To a greater extent than other projects so far – ACTION, for instance, or the Zürich-based FilmColors – SEMIA set out to explore the affordances of deep learning techniques for revealing similarity-based patterns in (large) collections of digitised moving images.[15] The assumption was that those patterns would enable users to follow alternative "routes" through the collections, "remixing" them as it were, and that this would elicit new questions about the items and their mutual relations. As we previously explained, we were specifically interested in relations inspired by the material's visual features – rather than the sort of filmographic or technical data that make up traditional metadata categories for film and video.

27 As it happens, such metadata, in archival collections, are often also fragmentary – and therefore, hardly reliable as a starting point for an inclusive form of collection exploration. Early on in the project, we took this as a key argument for looking into the possibilities of computer vision, and specifically deep learning techniques, for the purpose of feature extraction. This approach would help us generate large quantities of new metadata that would invite, if not a more inclusive kind of exploration, then at least one that could complement approaches to access based on search. After all, a lack of metadata in the form of semantic descriptors as encountered in an institutional catalogue may render the objects in a collection invisible, and therefore unfindable. While an approach relying on visual analysis does not solve this problem – as it can create new invisibilities, which we argue elsewhere (see Masson and Olesen [2020]) – it does challenge existing hierarchies of visibility.

28 Initially, the choice for a deep learning approach seemed to fit neatly with the project's intent to refrain as much as possible from determining in advance the route a computer might take in order to identify similarities between collection items. In the alternative scenario, known as "feature engineering," it is humans who design task-specific algorithms, which are used to extract pre-defined features from the images in a database (so that they can subsequently be compared). Deep learning, which relies to an overwhelming extent on the use of Neural Networks (NNs, or "neural nets" for short), involves algorithms trained with techniques for automatic feature learning (and as such, is a particular brand of machine learning). As we mentioned in the introduction, this is a more recent approach, and it entails the learning of specific data representations rather than set analysis tasks. Like feature engineering, deep learning does to some extent rely on the intervention of humans; after all, it is people who, at the training and/or retraining stages, determine which similarities do or do not make sense (see also Masson and van Noord [2019]; in Masson and Olesen [2020], we elaborate on the epistemic implications for users of our tool). However, it does not require them to decide in advance *how* the task of identifying those similarities needs to be performed (that is, on the basis of which features). In principle, this opens the door for image matches unanticipated by people, and therefore, of novel routes through a database or collection.

29 However, we soon decided to only partially rely on such techniques – and the abovementioned role of human knowledge is certainly one of the reasons why. As a rule, deep learning is employed for the recognition of semantic classes, and more specifically, object categories. This is hardly surprising, as the development of such techniques is oftentimes done for purposes that involve the recognition of semantic entities: vehicles, people, buildings, and so on. (One might think here of applications for transportation and traffic control, geolocation, or biometrics; see e.g. Uçar et al. [2017]; Arandjelović et al. [2018]; Taigman et al. [2014]). Within the SEMIA context, however, the use of conventional semantic classes does not make sense, as it is the sensory aspects of collection items – rather than the meanings we may assign to images, or image sections, on the basis of specific content – that are of interest. In fact, semantic classes commonly identified by deep learning approaches partially overlap with the sorts of categories that are used in

descriptive metadata for archival collections, and that are central also to practices of search and retrieve. In performing feature extraction, we had hoped to be able to work instead with more abstract visual categories, which according to computer vision logic, involves extraction at a lower ("syntactic") feature level (a point we elaborate on further below).

30 Another reason why exclusive reliance on a deep learning approach ultimately did not make sense, is that its underlying logic clashed with the requirements we had for interfacing. If our objective was to take sensory features as the point of entry into the collections, then it was imperative that our exploration tool allowed users to also take those features as the basis for digging further into the connections between items. For this purpose, we would need to at least minimally categorise, or re-categorise, those features, from the outset. The most logical choice here was to use the same intuitive classes that had also inspired the project: features such as colour, shape and visual complexity, and, for relations across time, movement.

31 One way of tackling this task with deep learning methods might have been to run successive analyses, whereby each time, the focus would be on one specific set of features, while other features would be cancelled out. For example, in order to extract information about shape, we might have deactivated the neural net's colour 'sensitivity' by temporarily turning all images in the database into black-and-white, so as to focus its attention in the required direction. This type of approach is generally associated with a (fairly new) line of research in computer science, focused on learning so-called "disentangled representations" (see Xiao et al. [2018]; Denton and Birodkar [2017]). So far, however, it has had limited success.[16] But even aside from the issues it currently entails, it also undermines our most fundamental rationale for working with deep learning techniques: the fact that one need not determine in advance how the particular task of similarity detection is carried out, and specifically, which types of features are used in the process. For this reason, we ultimately decided on a more diversified approach, which combined the use of deep learning with feature engineering.

32 A major point of attention was the need to attain a sufficient measure of abstraction in the results of the computer vision part of the project – results that were used in the development of a tool for visualising the sensory relations between the films and fragments in our database (a process we shall discuss elsewhere). We explained that our objective within SEMIA was to inspire users by revealing potentially significant relations between database items; in doing so, however, we sought to relegate the act of assigning such significance – or in Pigott's terms: of attempting to "read" images made illegible, through novel relations – as much as possible to users. For example, while we may want to draw attention to the circumstance that a specific set of database objects covers very similar colour schemes, or that they feature remarkably similar shapes, we leave it to the user to figure out whether, and if so how, this might be significant (that is, what questions it raises about media and their histories, or which alternative ways of researching historical film or television materials it affords). But arguably, we also withhold interpretation at a more basic level. In the above example, for instance, we leave undetermined whether similarity in colour or shape derives from the fact that the images concerned actually feature the same "things." (They might, and they often do – but it is not necessarily so.) In this respect, what we do is entirely at odds with the objectives of much machine learning practice in the field of computer vision.

33 Our search for abstraction is evidenced in a very concrete way by what happened exactly in the feature extraction process. First, the extraction of image information along the lines of colour, shape, visual complexity and movement was not followed in our case by an act of labelling: of placing an image or image section in a particular (semantic) class (we elaborate on this point in Masson and van Noord [2020]). The reason, of course, is that we did not actually seek to identify objects. For the purposes of our SEMIA experiment, the information as such, and the relations it allowed us to infer, were all we were interested in. Second, our search for abstraction is also evident from our application of deep learning methods, which was limited to the extraction of information about shape. Here, we focus on what computer vision experts call "lower-level" features – a notion that requires some further elaboration.

34 In computer vision, conceptual distinctions are oftentimes made between image features at different "levels." From one perspective, these are distinctions in terms of feature complexity. Levels of complexity range from descriptions relevant to smaller units (such as pixels in discrete images) to larger spatial segments (sections of such images, or entire images), whereby the former also serve as building blocks for the latter. From another, complementary perspective, the distinction can also be understood as a sliding scale from more "syntactic" (and abstract) to more "semantic" features

(the latter of which serve the purpose of object identification). Taking the example of shape-related information, we might think of a range that extends from unspecified shapes, for instance defined in terms of their edges (low-level), to more defined spatial segments such as contours or silhouettes (mid-level), all the way to actual object entities (e.g. things, people, faces, etc.) or relations between such entities. In SEMIA, we made use of a neural network trained for making matches at the highest (semantic) level. However, we scraped information at a slightly lower one, which generally contains descriptions of object parts. At this level, it recognises shapes, but without relating them to the objects they are part of.[17]

Arguably, this approach helped us mitigate a broader issue that the use of computer vision methods, and machine learning in general, posed for the project: that its techniques are designed, as Adrian MacKenzie puts it, to "mediate future-oriented decisions" – but by implication, also to *narrow down* a range of options by ruling other decisions out [MacKenzie 2017, 7]. In machine learning, datasets are used to produce probabilistic models, learned rules or associations, that generate predictive and classificatory statements [MacKenzie 2017, 8]. In the case of networks for image pattern recognition, for example, these are statements that lead to conclusions as to how much (or how little) images look alike. However, the valuation of "accurate" identifications at the semantic level as the highest achievable goal within machine learning also imposes limitations, in that it renders meaningless all other similarities – and importantly, dissimilarities – between objects in a database. Anna Munster, therefore, argues that prediction also "takes down potential" (quoted in Mackenzie [2017, 7]). Within the SEMIA context, we expressly tried to bring back some of this potential for the user. Sometimes this required us to deviate from what was 'state of the art' in the field of computer vision. Only in this way, after all, we could leave room for matches that might, within a purely semantic logic, be considered mistakes but still provide productive starting points for unrestrained explorations of patterns that perhaps no one had noticed before. 35

Extraction Results: Lesson Learnt

To round off this account, we now look at the results of our feature extraction efforts, and at what we learnt about the aptness of the approach for our goals. The classes of features the SEMIA project centred on are embedded in a rich history of computer vision research, which, as we previously explained, began with a process of manually designing features for predefined analysis tasks.[18] We also, however, deviated from this history, in that we did not use such algorithms for the purpose they were meant to serve: the assigning of (object) labels. Instead, we only relied on the feature descriptions they produced. Those descriptions are quite general, but still specific enough to bring out the sensory aspects of image elements that we were interested in. In what follows, we very briefly touch upon our methods (further technical details can be found in the notes) and then consider the results we obtained, evaluating their usefulness in light of the project's goals. 36

As mentioned earlier, we chose to focus on four broad sets of image features, commonly understood as instances of shape, colour, visual complexity and movement. Shape, we explained, is the only feature for which we extract information using a neural net. The net we chose was trained for object recognition, but is commonly repurposed for other tasks.[19] To make it meet our demands, we selected an intermediate feature representation rather than the uppermost layer in the net (that is, the highest complexity "level," where, as we explained in the previous section, the prediction probabilities for the semantic classes it was trained on are to be found).[20] This way, we could use its description of object parts and general shape, rather than of specific objects. For colour extraction, we made use of histograms, a common method in image processing.[21] Specifically, we chose histograms in CIELAB colour space (one that aligns closely with human perception) capturing the colour values used in a moving image irrespective of their spatial position. Visual complexity was understood in SEMIA as a measure of how much clutter there is in a visual scene (for instance, a highly textured or very busy scene will have a greater visual complexity than an empty scene, or one with mostly smooth surfaces). For the extraction of information of this kind, we used a method called Subband Entropy, which expresses a scene's visual complexity as a single scalar value.[22] 37

The features used to describe shape, colour, and visual complexity were all extracted with techniques that are applied to still images. In order to apply them to moving image material, we extracted feature descriptions from shots taken from 38

the films and programmes in our corpus. Specifically, we extracted the shape, colour, and visual complexity features from five frames, evenly spaced throughout the shot, and aggregated them to create the final feature descriptions. Movement, however, is a feature specific to moving images. For extracting this kind of information, we relied on an optical flow method, measuring relative motion between two subsequent frames. In each case, we applied it to the same sets of five frames.[23]

For the purpose of the project, we gathered approximately 7,000 videos, which we subsequently segmented into over 100,000 shots with the help of automatic shot boundary detection. Each of those shots was subjected to the four feature extraction algorithms. Altogether, this resulted in four different feature spaces, in which every shot constitutes a datapoint. By measuring the distance between all points, we could determine which other shots are most similar to a given one; the two closest points are known in this context as so-called “nearest neighbours.”[24] 39

In Figure 2, we show three examples with the “Query” shots to the left, represented here by a single still each, and the 16 shots identified as their nearest neighbours in the four different feature spaces to the right. A first possible observation concerns the diversity between the nearest neighbours for the three query shots: while all nearest neighbours share sensory aspects with their respective query image, they are considerably different from those for the other query shots. This at the very least suggests that they are not randomly selected. The second query shot, furthermore, shows a visible similarity between nearest neighbours across the four different feature spaces for each query image. This last pattern logically follows from the nature of nearest neighbours, in that shots that look similar in one sensory aspect, are likely to also look similar in others. Colours in a nature shot (such as the mushroom in the third query shot), for example, are very distinctive, making it likely that its nearest neighbours in terms of colour are also nature scenes. Similarly, the movement of leaves swaying in the wind is very distinctive, making it probable that the nearest neighbours of a shot with this element, in movement terms, also show leaf-rich scenes. 40

At the same time and in spite of other visual similarities, our query images also produce matches that are quite distinct, precisely, in terms of the semantic entities they feature. The movement feature space for the mushroom query image, for instance, features a standing man (presumably, one who moves from left to right or the other way around, in the same way that the mushroom does; however, it would require further inspection to ascertain this or to make sense of this pairing). In instances like these, the matching process has arguably yielded more unexpected or surprising results and variations. Moreover, such matches occur more often if we look beyond the closest of the nearest neighbours. For example, a desert scene is similar to a beach scene in terms of colour, but not in terms of movement; in contrast, a grassy plain has similar movement to a beach scene, but differs strongly in colour. Hence, by exploring similarities in multiple feature spaces, we are still able to uncover such relations that would otherwise remain hidden. 41

Figure 2. Sample stills of query shots from the Open Images platform with four nearest neighbours in the shape, colour, movement, and visual complexity feature spaces

Conclusions

In this article, we have argued for a reorientation of existing visual analysis methods, in response to a need for exploratory browsing of media archives. We explained how we took our cue from a recent line of digital scholarship inspired by artistic strategies in (new) media art, and how we also built on the tradition of exchange between film archives, media history and appropriation art. Historically, artists have used the analytical devices of scholars to different ends, thus engendering shifts in the latter's working assumptions. In a similar vein, the SEMIA project team drew inspiration from the ways in which data artists repurpose existing visual analysis tools. We did so with the specific goal of enabling a transition from searching to browsing large-scale moving image collections. This way, we not only hoped to significantly expand the range of available metadata, but also to allow for the revaluation of the images' sensory

dimensions in the very early stages of research. Ultimately, we think, both approaches to collection access can very well complement each other.

43 Our goal required that for the extraction of data, we adhered to the following general guidelines. In order to reduce the system's reliance on *a priori* interpretations, we first of all sought to avoid direct human intervention in the actual extraction process. As a matter of principle, it should be up to the algorithm to determine "similar," "somewhat similar," or "dissimilar" – even if, as we argue elsewhere, algorithms ultimately always rely on knowledge that originates in humans (see Masson and van Noord [2019]). Furthermore, we tweaked the algorithm to partially prevent it from recognising (human-taught) semantic units. Consequently, it could focus on similarities at a more abstract level. At this stage, some human intervention is ultimately unavoidable, as it is the computer scientist who decides (ideally on the basis of sample testing results) at which feature "level" the extraction takes place.

44 One conclusion that can be drawn from our review of most similar results is that extracting data with a minimum of labelling and human intervention, while also attending to intermediate similarities, never truly cancels out the detection of semantic relations and patterns altogether. In fact, this is hardly surprising, because this relation between low-level feature representations and objects – one that frames objects in terms of its facets; for instance, in the case of an orange, its colour and rounded shape – has been commonly exploited in early work on computer vision to detect semantic relations and objects. Therefore, some feature combinations are simply too distinctive to not be detected with our chosen approach – even if we do our best to block the algorithms' semantic "impulse."[25] Yet our examples show that the analysis of query images also produces nearest neighbour matches that initially seem more "illegible," and therefore, invite further exploration. In this sense, our working method does yield surprising results, or unexpected variations. In the remainder of our project, which we report on elsewhere, our intent has been to further stimulate users in exploring those less obvious connections by extending our interface with the capacity to also browse *dis*similar results.

45 The next step, which we expand on in an upcoming piece, is to assess which kinds of questions and ideas exploratory browsing through the lens of sensory features ultimately yields, and to evaluate how this furthers the efforts of various user groups [Masson and Olesen 2020]. Throughout our research process, we have been wondering about the potential of such browsing for the purpose of what (social) scientists, and more recently also information and media scholars, have termed "serendipitous" discoveries [van Andel 1994] [Foster and Ellis 2014]. The literature uses this term for chance encounters with research objects that engender new ways of explaining or thinking about problems – both known ones, and problems one was perhaps previously unaware of.

Notes

[1] https://www.khm.at/en/visit/exhibitions/2019/wesandersonandjumanmalouf2018/

[2] The exhibit ran from 5 November 2018 until 28 April 2019.

[3] http://sensorymovingimagearchive.humanities.uva.nl/

[4] For further exploration of the relation between searching and browsing, and the explorative affordances of browsing, see Masson (2019), or Masson and Olesen (2020).

[5] Funding was obtained within the SMART Culture scheme of the Netherlands Organisation for Scientific Research (NWO).

[6] See https://www.openbeelden.nl/. Of course, working with digitised versions of originally analogue moving images entails that some of their potentially significant material aspects are already 'erased' in a process that precedes the act of engaging with a collection. In the SEMIA project, we took this to be an inevitability.

[7] In Masson and van Noord (2020), we elaborate on this history.

[8] https://www.ims.tuwien.ac.at/projects/digital-formalism

[9] https://hcommons.org/deposits/item/hc:12153/

[10] ACTION is short for Audio-visual Cinematic Toolbox for Interaction, Organization, and Navigation. http://digitalhumanities.dartmouth.edu/projects/the-action-toolbox/

[11] In this respect, the SEMIA project also differs in its intent from other initiatives that have been reported on since the writing of this piece; for instance, projects by the Distant Viewing Lab (https://www.distantviewing.org/) at the University of Richmond, reported on in Arnold and Tilton (2019) (focusing on narrative patterns and patterns in photographic style) or at the National Library of the Netherlands, by researchers-in-residence Melvin Wevers and Thomas Smits (on stylistic trends in newspaper visuals) [Wevers and Smits 2020].

[12] http://www.jimcampbell.tv/portfolio/still_image_works/illuminated_averages/index.html

[13] http://www.salavon.com/work/Top25/

[14] http://www.brendandawes.com/projects/cinemaredux

[15] The project's full title is: *FilmColors: Bridging the Gap Between Technology and Aesthetics*. It runs until August of 2020. https://filmcolors.org/2015/06/15/erc/

[16] Hitherto, it has primarily been successful when applied to restricted domain and toy problems.

[17] Many thanks to Matthias Zeppelzauer (St. Poelten University of Applied Sciences) for helping us gain a better understanding of these conceptual distinctions. For more on how neural nets specifically "understand" images, see also Olah et al. (2017).

[18] For example, Swain and Ballard, in the early 1990s, used colour information to identify and localise the position of objects [Swain and Ballard 1991].

[19] Specifically, we used ResNet-101; for more information on its repurposing, see He et al. (2016).

[20] The layer we selected was the one located just below the fully connected layers, of 2048 dimensions.

[21] With this approach, each colour dimension is described by 16 bins, resulting in a feature representation of 4096 dimensions.

[22] For more information, see Rosenholtz et al. (2007).

[23] This involved constructing a histogram, for which we separately binned the angle and magnitude for a three by three grid of non-overlapping spatial regions – an approach akin to the HOFM approach described in Colque et al. (2017).

[24] The concept of "nearest neighbour" is also key to the *Neural Neighbours: Pictorial Tropes in the Meserve-Kunhardt Collection* project (https://dhlab.yale.edu/projects/neural-neighbors/) conducted by the Yale Digital Humanities Lab at the Yale Beinecke Rare Book & Manuscript Library. So far, however, this project has focused specifically on (originally) still images.

[25] Exact matches rarely occur, because for the purposes of the project, the detection settings are tweaked in such a way that matches between images from the same videos are ruled out. (Therefore, only duplicate videos in the database can generate such results.)

Works Cited

Arandjelović et al. 2018 Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40.6 (2018): 1437-51. DOI: 10.1109/TPAMI.2017.2711011.

Arnold and Tilton 2019 Taylor, Arnold, and Tilton, Lauren. "Distant viewing: analyzing large visual corpora". *Digital Scholarship in the Humanities,* fqz013 (2019). DOI: 10.1093/digitalsh/fqz013.

Brown 2018 Brown, K. "Wes Anderson's Offbeat Debut as a Curator Drove a Storied Museum's Staff Crazy: The Results Are Enchanting". *Artnet News* (2018). Available at: https://news.artnet.com/exhibitions/wes-anderson-curator-kunsthistorisches-museum-1387429 (accessed 1 March 2020).

Buckland 2008 Buckland, W. "What Does the Statistical Style Analysis of Film Involve?", *Literary and Linguistic Computing*, 23.2 (2008): 219-30. DOI: 10.1093/llc/fqm046.

Casey and Williams 2013 Casey, M., and Williams, M. "ACTION (Audio-visual Cinematic Toolbox for Interaction, Organization, and Navigation): an open-source Python platform" white paper, report ID 104081 (2013). Available at: https://hcommons.org/deposits/item/hc:12153/ (accessed 29 March 2020).

Catanese et al. 2019 Catanese, R., Scotto Lavina, F. and Valente, V. (eds.). *From Sensation to Synaesthesia in Film and New Media.* Cambridge Scholars Publishing, Cambridge (2019).

Colque et al. 2017 Colque, R. V. H. M., Caetano, C., De Andrade, M. T. L., and Schwartz, W. R. "Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos", *IEEE Transactions on Circuits and Systems for Video Technology*, 27.3 (2017): 673-82. DOI: 10.1109/TCSVT.2016.2637778.

Delpeut 1999 Delpeut, P. *Diva dolorosa: Reis naar het einde van een eeuw.* Meulenhoff, Amsterdam (1999).

Denton and Birodkar 2017 Denton, E., and Birodkar, V. "Unsupervised Learning of Disentangled Representations from Video". In E. Guyon et al. (eds.), *Advances in Neural Information Processing Systems 30, Neural Information Processing Systems Foundation* (2017). Available at: https://papers.nips.cc/paper/7028-unsupervised-learning-of-disentangled-representations-from-video.pdf (accessed 29 March 2020).

Dudley 2010 Dudley, S. (ed.). *Museum Materialities: Objects, Engagements, Interpretations.* Routledge, London (2010).

Elsaesser 2010 Elsaesser, T. "Archives and Archaeology: The Place of Non-Fiction Film in Contemporary Media". In V. Hediger and P. Vondereau (eds.), *Films That Work: Industrial Film and the Productivity of Media*, Amsterdam University Press, Amsterdam (2009), pp. 19-34.

Ferguson 2016 Ferguson, K. L. "The Slices of Cinema: Digital Surrealism as Research Strategy". In C. R. Acland and E. Hoyt (eds.), *The Arclight Guidebook to Media History and Digital Humanities*, REFRAME Books, Sussex (2016), pp. 270-299.

Ferguson 2017 Ferguson, K. L. "Digital Surrealism: Visualizing Walt Disney Animation Studios", *Digital Humanities Quarterly*, 11.1 (2017). Available at: http://www.digitalhumanities.org/dhq/vol/11/1/000276/000276.html (accessed 29 March 2020).

Flanders 2014 Flanders, J. "Rethinking Collections". In P. Longley Arthur and K. Bode (eds.), *Advancing Digital Humanities: Research, Methods, Theories*, Palgrave Macmillan, Houndmills (2014), pp. 163-174.

Flückiger 2017 Flückiger, B. "A Digital Humanities Approach to Film Colors", *The Moving Image*, 17.2 (2017): 71–93.

Foster and Ellis 2014 Foster, A. E., and Ellis, D. "Serendipity and its study", *Journal of Documentation*, 70.6 (2014): 1015-38. DOI: 10.1108/00220410310472518.

Habib 2015 Habib, A. *La Main gauche de Jean-Pierre Léaud*. Les Éditions du Boréal, Montréal (2015).

He et al. 2016 He, K., Zhang, X., Ren, S., and Sun, J. "Deep Residual Learning for Image Recognition". In *IEEE Conference on Computer Vision and Pattern Recognition*, Computer Vision Foundation (2016), pp. 770-78. DOI: 10.1109/CVPR.2016.90.

Heftberger 2018 Heftberger, A. *Digital Humanities and Film Studies: Visualising Dziga Vertov's Work*. Springer, Cham (2018).

Heftberger et al. 2009 Heftberger, A., Tsivian, Y., and Lepore, M. "Man with a Movie Camera (SU 1929) under the Lens of Cinemetrics", *Maske und Kothurn* 55.3 (2009): 31-50. DOI: 10.7767/muk.2009.55.3.61.

Kuhn et al. 2013 Kuhn, V., Craig, A., Franklin, K., Simeone, M., Arora, R., Bock, D., and Marini, L. "Large Scale Video Analytics: On-demand, iterative inquiry for moving image research". In *2012 IEEE 8th International Conference on E-Science* (2013). DOI: 10.1109/eScience.2012.6404446.

MacKenzie 2017 MacKenzie, A. *Machine Learners: Archaeology of a Data Practice*. MIT Press, Cambridge, MA (2017).

Masson 2019 Masson, E. "Browsing Moving Image Collections". *The Sensory Moving Image Archive* (2019). Available at: https://sensorymovingimagearchive.humanities.uva.nl/index.php/2019/11/26/browsing-moving-image-collections/ (accessed 1 March 2020).

Masson and Olesen 2020 Masson, E., and Olesen, C.G. "Digital Access as Archival Reconstitution: Algorithmic Sampling, Visualization, and the Production of Meaning in Large Moving Image Repositories". *Signata: Annales des sémiotiques/Annals of Semiotics*, 12 (2020).

Masson and van Noord 2020 Masson, E., and van Noord, N. "Feature Extraction and Classification". *The Sensory Moving Image Archive* (2020). Available at: https://sensorymovingimagearchive.humanities.uva.nl/index.php/2020/01/06/feature-extraction-and-classification/ (accessed 1 March 2020).

Mazzanti 2009 Mazzanti, M. "Colours, Audiences and (Dis)Continuity in the 'Cinema of the Second Period'", *Film History*

21.1 (2009): 67-93.

Mul and Masson 2018 Mul, G., and Masson, E. “Data-Based Art, Algorithmic Poetry: Geert Mul in Conversation with Eef Masson”, *TMG – Journal for Media History*, 21.2 (2018). Available at: https://www.tmgonline.nl/articles/10.18146/2213-7653.2018.375/ (accessed 29 March 2020).

Olah et al. 2017 Olah, C., Mordvintsev, A., and Schubert, L. “Feature Visualization: How neural networks build up their understanding of images”. *Distill* (2017). DOI: 10.23915/distill.00007.

Olesen 2017 Olesen, C. G. “Towards a ‘Humanistic’ Cinemetrics?” In K. van Es and M. T. Schäfer (eds.), *The Datafied Society: Studying Culture through Data*, Amsterdam University Press, Amsterdam (2017), pp. 39-54.

Pigott 2015 Pigott, M. *Joseph Cornell Versus Cinema*. Bloomsbury Academic, London (2015).

Rosenholtz et al. 2007 Rosenholtz, R., Li, Y., and Nakano, L. “Measuring Visual Clutter”, *Journal of vision*, 7.2 (2007): 17. DOI: 10.1167/7.2.17.

Street 2012 Street S. *Colour Films in Britain: The Negotiation of Innovation 1900-1955*. BFI/Palgrave Macmillan, London (2012).

Street and Yumibe 2013 Street, S., and Yumibe, J. “The temporalities of intermediality: Colour in cinema and the arts of the 1920s”, *Early Popular Visual Culture* 11.2 (2013): 140-57. DOI: 10.1080/17460654.2013.783149.

Swain and Ballard 1991 Swain, M. J., and Ballard, D. H. “Color Indexing”, *International Journal of Computer Vision*, 7.1 (1991): 11–32. DOI: 10.1007/BF00130487.

Taigman et al. 2014 Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society/CPS (2014). DOI: 10.1109/CVPR.2014.220.

Testa 1992 Testa, B. *Back and Forth: Early Cinema and the Avant-Garde*. Art Gallery of Ontario, Ontario (1992).

Uçar et al. 2017 Uçar, A., Demir, Y., and Güzeliş, C. “Object recognition and detection with deep learning for autonomous driving applications”, *Simulation*, 93.9 (2017): 759-769. DOI: 10.1177/0037549717709932.

Wevers and Smits 2020 Wevers, M. and Smits, T. “The visual digital turn: Using neural networks to study historical images”, *Digital Scholarship in the Humanities*, 35.1 (2020): 194-207. DOI: 10.1093/llc/fqy085.

Whitelaw 2015 Whitelaw, M. “Generous Interfaces for Digital Cultural Collections”, *Digital Humanities Quarterly*, 9.1 (2015). Available at: http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html (accessed 29 March 2020).

Xiao et al. 2018 Xiao, T., Hong, J., and Ma, J. 2018. "DNA-GAN: Learning Disentangled Representations from Multi-Attribute Images". In *ICLR 2018 – Workshop track*. ICLR, (2018). Available at: https://arxiv.org/pdf/1711.05415.pdf (accessed 29 March 2020).

Yumibe 2012 Yumibe, J. *Moving Color: Early Film, Mass Culture, Modernism*. Rutgers University Press, New Brunswixck NJ/London (2012).

Zeppelzauer et al. 2012 Zeppelzauer, M., Mitrović, D., and Breiteneder, C. “Archive Film Material – A Novel Challenge for Automated Film Analysis”, *Frames Cinema Journal*, 1.1 (2012). Available at http://www.framescinemajournal.com/article/archive-film-material-a-novel-challenge/?format=pdf (accessed 29 march 2020).

van Andel 1994 Van Andel, P. “Anatomy of the Unsought Finding. Serendipity: Origin, History, Domains, Traditions, Appearances, Patterns and Programmability”, *The British Journal for the Philosophy of Science*, 45.2 (1994): 631-48.