volume 7 issue 14/2018

VIEW

JOURNAL OF
EUROPEAN
TELEVISION
HISTORY
& CULTURE

COMPUTER VISION AND THE DIGITAL HUMANITIES

ADAPTING IMAGE PROCESSING ALGORITHMS AND GROUND TRUTH THROUGH ACTIVE LEARNING

Christoph Musik
St. Pölten University of Applied Sciences
Institute of Media Economics
Matthias Corvinus-Straße 15
3100 St. Pölten
Austria
christoph.musik@fhstp.ac.at

Matthias Zeppelzauer
St. Pölten University of Applied Sciences
Institute of Creative Media Technologies
Media Computing Research Group
Matthias Corvinus-Straße 15
3100 St. Pölten
Austria
matthias.zeppelzauer@fhstp.ac.at

Abstract: Automated computer vision methods and tools offer new ways of analysing audio-visual material in the realm of the Digital Humanities (DH). While there are some promising results where these tools can be applied, there are basic challenges, such as algorithmic bias and the lack of sufficient transparency, one needs to carefully use these tools in a productive and responsible way. When it comes to the socio-technical understanding of computer vision tools and methods, a major unit of sociological analysis, attentiveness, and access for configuration (for both computer vision scientists and DH scholars) is what computer science calls "ground truth". What is specified in the ground truth is the template or rule to follow, e.g. what an object looks like. This article aims at providing scholars in the DH with knowledge about how automated tools for image analysis work and how they are constructed. Based on these insights, the paper introduces an approach called "active learning" that can help to configure these tools in ways that fit the specific requirements and research questions of the DH in a more adaptive and user-centered way. We argue that both objectives need to be addressed, as this is, by all means, necessary for a successful implementation of computer vision tools in the DH and related fields.

Keywords: Digital Humanities, Computer Vision, Image Understanding, Machine Learning, Ground Truth Generation, Explainable Artificial Intelligence, Active Learning

1 Introduction

With the constantly growing amount of visual and audiovisual data, the automated and computer-assisted analysis gains more and more attention. Automated computer vision tools that are based on image processing algorithms (IPAs) offer new ways of analysing audio-visual material in the realm of Digital Humanities (DH) and also in Digital Social Sciences. While there are some promising results where CV tools can be applied and are used in DH and beyond[1], there are basic challenges one needs to carefully deal with in order to use these tools in a productive and responsible way. For instance, Heftberger reported that "whenever semantics (image content) comes in, software is not able to detect and annotate it"[2] automatically. A combination of manual and automated annotation and analysis is proposed. Other challenges are algorithmic bias and its implications, or, the lack of transparency of algorithmic decision-making. When it comes to the socio-technical understanding of IPAs and how different kinds of these operatively work in automated computer vision tools, a major unit of sociological analysis, attentiveness, and access for configuration (for both computer vision scientists and DH scholars) is what computer science calls *ground truth*. In short, we define ground truth as an interpretation template that instructs the automated decision of IPAs and computer vision tools during learning. What is specified in the ground truth is the template or rule to follow, e.g. what an object looks like, where an object is located and which area of the image it covers. This paper brings to the fore the sociotechnical construction and basic modes of operation of IPAs and their ground truths in order to understand and work on computer vision tools and their relation to the DH in an interdisciplinary manner.

This article aims at, firstly, providing scholars in the DH with knowledge about how automated tools for image analysis work and how they are constructed, i.e. in the words of Bruno Latour we try to open up the black box[3] of computer vision tools. Secondly, based on these insights, the paper proposes an approach to configure these tools in ways that fit the specific requirements and research questions of the DH in a more adaptive and user-centered way. This especially refers to the approach of active learning[4] and takes up discussions and challenges of explainable artificial intelligence (XAI)[5,6]. We argue that both objectives need to be addressed, as this is, by all means, necessary for a successful implementation of computer vision tools in the DH and related fields. One reason for this is the narrow, inflexible and opaque nature of ready-made and off-the-shelf available computer vision tools, which may restrict their application in the context of DH and may further introduce unwanted bias in the analysis. This contribution represents a position and conceptual paper aiming at both a critical understanding of computer vision tools in their application in the DH and based on this, a constructive approach of how to apply computer vision in the DH in a more meaningful and responsible way. These two objectives are closely linked and do not contradict each other as they both address limitations and opportunities. It is a call neither simply for nor against the deployment of computer vision in the DH but a proposal for how it should be involved in a nuanced and reflexive way. The paper and its objectives are the outcome of the previous experience of the authors (a social scientist specialised in science and technology studies (STS) and media sociology, and a computer vision researcher) in interdisciplinary research at the interface between computer vision and different DH areas. In this regard both refer to practical work and empirical projects where they either worked on or analysed computer vision, IPAs and ground truth construction.

In the following, we first elaborate on the wider sociotechnical developments as well as debates framing the involvement of this paper. This refers to the age of visual big data and the social power of algorithms. In a second step, we provide methodological background information on the practical work and empirical ethnographic research

[1] Maia Zaharieva, Dalibor Mitrovic, Matthias Zeppelzauer, and Christian Breiteneder, 'Film Analysis of Archive Documentaries.' *IEEE Multimedia*, 18, 2, 2017, pp. 38-47.

[2] Adelheid Heftberger, 'Do Computers Dream of Cinema? Film Data for Computer Analysis and Visualisation', in David M. Berry, ed, *Understanding Digital Humanities*, 2012, p.218.

[3] Bruno Latour, *Pandora's hope: essays on the reality of science studies,* Harvard University Press, 1999, p. 304.

[4] Bernard, J., Hutter, M., Zeppelzauer, Fellner, D., Sedlmair, M. (2017) Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study. IEEE Transactions on Visualization and Computer Graphics (TVCG), 24:1(298-308), doi: 10.1109/TVCG.2017.2744818, issn: 1077-2626.

[5] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences.' *Artificial Intelligence*, 267, 2018, pp. 1–38.

[6] Amina Adadi and Mohammed Berrada (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 2018, pp. 52138–52160.

that shaped our position and thereby this conceptual paper. This also includes the example of automatic visual 'fall detection' that we focus on at the end of the main part of the paper. Furthermore, we discuss different conceptual and basic levels of computer vision, including questions of difference between computer and human vision, and, visual expertise. Following this, we address the sociotechnical construction of ground truth and its implications for the DH. This part completes with a discussion of algorithmic bias and error and of how this issue can be dealt with. In the final part of the paper, we propose an approach enabling user participation in the construction of ground truth and highlight the importance of explainable artificial intelligence in this context.

2 The Age of Visual Big Data and the Social Power of Algorithms

In recent years, there has been a worldwide explosion of visual and audiovisual data. These data come from various sources such as images and video clips on social sharing sites (e.g. **Instagram** or **YouTube**), films and series on video-on-demand streaming platforms (e.g. **Netflix**), live coverage on over-the-top services (e.g. **DAZN**), a multiplicity of new digital linear television channels, and the digitalisation of analogue and historic moving image (e.g. film collections).

There is no doubt that with the beginning of this decade we started living in the "Age of Big Data". For example, in an article published in 2012 (Feb 20, "Mining an Information Explosion") the **New York Times** officially welcomed this new age. The welcome call referred to the constantly growing amount of data and stated, quoting an estimation by **IDC**, a technology research firm, that data is currently growing at 50 percent a year, or doubling every two years. The most recent IDC white paper on the Data Age[7], sponsored by the US data storage company **Seagate**, estimates the 2016 amount of worldwide data to be 16,1 ZB (that is 16 100 000 000 000 000 000 000 bytes). IDC forecasts see the amount of data growing up to 163 zettabytes by 2025.

It is important to note that this data explosion is predominantly visual, with data deriving from various sources such as images, videos, and surveillance streams, with the moving image being in pole position when it comes to the quantity of global data. We thus live in the age of "*Visual* Big Data". IDC estimates a constant growth of both entertainment data (image and video content created or consumed for entertainment purposes) and non-entertainment data (image and video content for non-entertainment purposes, such as video surveillance footage or advertising).

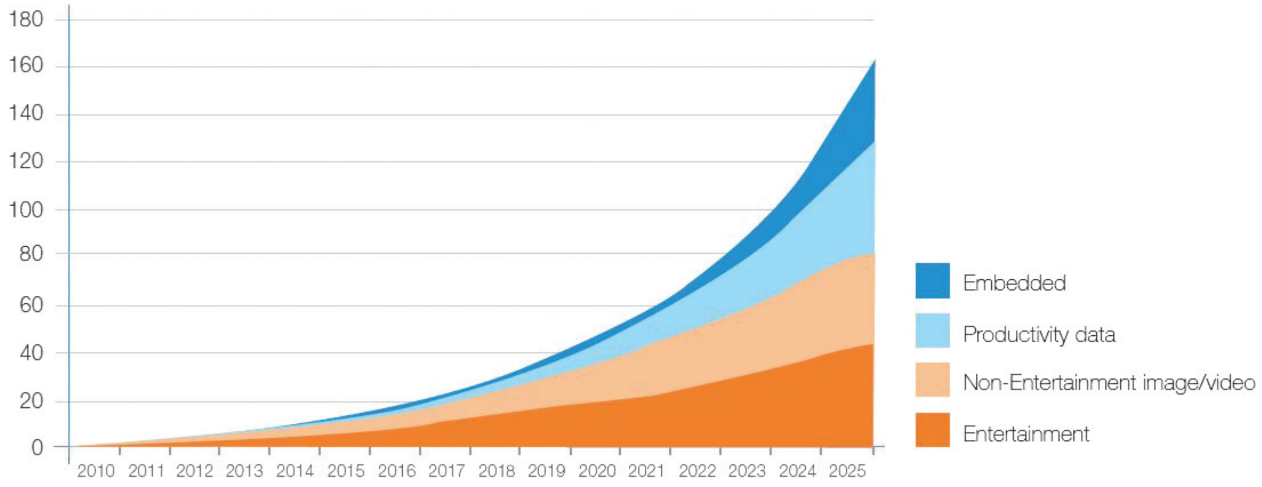

Figure 1. Data creation by type. There is a strong pre-dominance of visual data (Source: IDC's Data Age 2025 study).

[7] David Reinsel, John Gantz, and John Rydning, 'Data Age 2025: The Evolution of Data to Life-Critical,' *An IDC White Paper, Sponsored by Seagate*, April 2017, **https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf**

The great quantity of (audio)visual data and its constant growth implies that coping with, and making sense of data, is not limited to numbers and textual data but does most of all refer to visual data. Against this background, image processing algorithms (IPAs) implemented in computer vision tools for automated image analysis and understanding are essential in making sense of this ever-growing amount of visual data. IPAs can be considered as the basic technology behind all approaches that deal with the automatisation of analysing and understanding visual materials on different levels[8].

Since more and more scholars in the DH work with automated computer vision tools such as **Cinemetrics** and **Videana**, IPAs are becoming increasingly more powerful societal actors in this regard[9]. Thus, it is important to understand exactly and reflect carefully on the production, processing, and interpretation of (digital) images by algorithms on a broader scope and not exclusively in the realm of computer sciences. Algorithms challenge traditional human approaches of analysis in the humanities and other scientific disciplines as they provide a new perspective to the investigated problems and data as well as a new knowledge logic[10]. This logic depends very much on the "proceduralized choices of a machine, designed by human operators to automate some proxy of human judgement or unearth patterns across collected social traces"[11]. According to Gillespie, this logic is competing with the 'editorial logic' in which the subjective choices of trained and certified experts (e.g. registrars, researchers) are dominant. The following questions remain: How do algorithms choose information? How do they select and "know"? How do they "see"? And, is there actually a difference between the 'algorithmic logic' and the 'editorial logic'? The amount of (critical) literature regarding the role of algorithms in recent years in the fields of social sciences and humanities keeps growing constantly (see for example on "the ethics of algorithms" Mittelstadt et al. 2016[12], on the "ethics and politics of algorithms" Matzner 2018[13], or, on "algorithms as culture" Seaver 2017[14]). There is, however, a gap in critical literature reflecting on this topic and its relation to the DH from an interdisciplinary perspective.

3 Notes on Methods, Materials and Background

This conceptual and position paper is the outcome of the research experience of the authors (a social scientist specialised in science and technology studies (STS) and media sociology and a computer vision scientist) in the interdisciplinary engagement on computer vision in the DH. The paper focuses on a conceptual and practical level of computer vision, computer vision tools, and their relation to the DH. This involvement is nevertheless informed by and grounded in different kinds of empirical analyses, case studies and practical interdisciplinary work in the context of computer vision and DH. Therefore, we very briefly provide some basic information on the background of the authors and their work. We are well aware that especially ethnographic fieldwork is in need of proper contextualisation and reflection but the framework of this contribution does not allow for accounting this in-depth.

The first author analysed the field of computer vision and the matter of IPAs in different projects (some of these collaborative interdisciplinary projects) from the standpoint of STS over the course of the last ten years. This means,

[8] Christoph Musik, Computers and the Ability to See. Understanding the negotiation and implementation of image processing algorithms, PhD dissertation, University of Vienna, 2014.

[9] David Beer, 'The social power of algorithms,' *Information, Communication & Society*, 20, 1, 2017, pp.1-13.

[10] Tarleton Gillespie, 'The Relevance of Algorithms,' in Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, eds, *Media Technologies. Essays on Communication, Materiality, and Society,* 2014, p. 192.

[11] Tarleton Gillespie, 'The Relevance of Algorithms,' in Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, eds, *Media Technologies. Essays on Communication, Materiality, and Society*, 2014, p. 192.

[12] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi, 'The Ethics of Algorithms: Mapping the Debate,' *Big Data & Society*, December 2016, pp.1-21.

[13] Tobias Matzner, 'Grasping the ethics and politics of algorithms,' in Ann Rudinow Saetnan, Ingrid Schneider, and Nicola Green, eds, The Politics of Big Data. Big Data, Big Brother?, 2018, pp. 39-45.

[14] Nick Seaver, 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems,' Big Data & Society, December 2017, pp.1-12.

based on approaches of 'laboratory studies'[15,16] and in particular on Forsythe's interpretative cultural anthropology approach[17], both computer scientists' viewpoints and the everyday techno-scientific practice in computer vision laboratories and its surroundings were analysed by means of ethnographic fieldwork[18]. While the paper at hand is able to also draw on many (formal and informal) interviews and collaborative projects with different computer vision scientists from around the world, the main part of this lab-based fieldwork took place in an Austrian computer vision laboratory over a period of two months in 2011. Within this period, the first author was able to observe and be involved in the work on so-called automated fall detection. This involvement is the empirical basis for the example presented in the main part of this paper.

The second author is a computer vision scientist with long-time experience in developing IPAs. A specific focus of his work is the research and application of IPAs in interdisciplinary problem settings, e.g. in cooperation with media scholars and film scholars[19]. In previous joint research, we could identify both limitations[20] and potentials[21] of the application of IPAs in the DH. From the joint research, we could gather in-depth information about the diverse requirements of researchers in the DH employing computer vision approaches in their research. In this paper, we discuss and elaborate on the basic findings and insights gained of the previous interdisciplinary research and develop ideas of how computer vision can be best integrated in DH research and studies.

4 What Does "Seeing" Mean? Between Computer Vision and Human Vision

To understand computer vision we need to understand human vision. Computer vision and human vision are closely connected and intertwined, which stems from the fact that computer vision tries to imitate human vision. Research in STS[22] showed that the boundaries between humans and machines are constant subjects of discussion and negotiation. So are the boundaries between human vision and computer vision. It is crucial to note that questions addressing computer vision are simultaneously always questions addressing human vision, too. To put it briefly, this is because humans are inevitably involved in the design and development of computers that are able to see. But what does "seeing" actually mean? What understandings of seeing, and closely connected to it - recognising - do we have, especially when it comes to teaching computers to achieve sight? Is there one universal way of seeing that is easily transferred to machines, or is seeing rather a diverse "situated" and cultural activity that hampers a simple and smooth transformation of this ability?

From a sociological point of view, seeing is subject to change, both culturally and historically[23]. In order to emphasise different forms of seeing in the context of artificial intelligence, Collins distinguishes a formal (or pattern recognition)

[15] Bruno Latour, and Steve Woolgar, *Laboratory Life. The Construction of Scientific Facts*. Sage Publications, 1979.

[16] Karin Knorr-Cetina, *The Manufacture of Knowledge. An Essay on the Constructivist and Contextual Nature of Science*. Pergamon, 1981.

[17] Diana E. Forsythe, 'Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence,' *Social Studies of Science*, 23, 3, 1993, pp. 445-477.

[18] Christoph Musik, Computers and the Ability to See. Understanding the negotiation and implementation of image processing algorithms, PhD dissertation, University of Vienna, 2014.

[19] Maia Zaharieva, Dalibor Mitrovic, Matthias Zeppelzauer, and Christian Breiteneder, 'Film Analysis of Archive Documentaries.' *IEEE Multimedia*, 18, 2, 2017, pp. 38-47.

[20] Matthias Zeppelzauer, Dalibor Mitrovic, and Christian Breiteneder, 'Archive Film Material - A novel Challenge for Automated Film Analysis' in Catherine Grant,ed, *Frames Cinema Journal*, 1, 2012.

[21] Matthias Zeppelzauer, Maia Zaharieva, Dalibor Mitrovic, and Christian Breiteneder, 'Retrieval of Motion Composition in Film.' *Digital Creativity*, 22, 4, 2011, pp. 219-234.

[22] Lucy Suchman, *Human-Machine Reconfigurations. Plans and Situated Actions. 2nd Edition,* Cambridge University Press, 2007, p.226.

[23] Regula Valérie Burri, and Joseph Dumit, 'Social Studies of Scientific Imaging and Visualization,' in Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, eds, *The Handbook of Science and Technology Studies. Third Edition.* The MIT Press, 2008, pp. 297–317.

model of seeing from an enculturational model[24]. The formal model of seeing “involves recognizing what an object really is by detecting its distinguishing characteristics.” The enculturational model of seeing stresses that the same object may be seen as many different things. In this regard, Goodwin’s analysis[25] of the so-called ‘King Trial’ and the video footage of police violence discussed and transformed in court is a good example for the enculturational model. In the ‘King Trial’, four white police officers were charged with beating Rodney King, an African-American motorist, who had been stopped for speeding in the US in 1992[26]. The incident had been videotaped (see video “RODNEY KING BEATING VIDEO Full length footage SCREENER” below) and for the prosecutor it was clear, objective evidence showing uncontrolled and brutal violence against Rodney King. However, the lawyers defending the police officers did not treat the tape as evidence that spoke for itself. Rather, they were able to transform the perception of the tape from evidence to “a very disciplined and controlled effort to take Mr. King into custody”[27]. With the help of a coding scheme delivered by experts showing how police usually work, a ‘perceptual transformation’ was accomplished. Goodwin concludes that “the perspectival framework provided by a professional coding scheme constitutes the objects in the domain of scrutiny that are the focus of attention”[28].

Video 1. “RODNEY KING BEATING VIDEO Full length footage SCREENER”

5 Who Is the Visual Expert? Human vs. Machine or: Human and Machine?

Reading and analysing audiovisual data in the humanities and in other scientific fields such as the social sciences is first of all about the understanding and interpretation of (moving) images. As such, it is an interesting subject of analysis for the approach of a “sociology of images”[29]. This approach investigates the processes by which (scientific) image interpretation (e.g. what is depicted on an image, painting, X-Ray, etc.) is interactively negotiated in social practices. This means that the focus is not only on the images and their content alone but also on the micro-practices and contexts of image production, interpretation and use (e.g. as evidence for an entity). This focus is similar to

[24] Harry M. Collins, *Tacit and Explicit Knowlegde*. University of Chicago Press, 2010, p.11.

[25] Charles Goodwin, ‘Professional Vision,’ *American Anthropologist,* 96, 3, 1994, pp. 606-633.

[26] Charles Goodwin, ‘Professional Vision,’ *American Anthropologist,* 96, 3, 1994, p. 606.

[27] Charles Goodwin, ‘Professional Vision,’ *American Anthropologist,* 96, 3, 1994, p. 617.

[28] Charles Goodwin, ‘Professional Vision,’ *American Anthropologist,* 96, 3, 1994, p. 622.

[29] Regula Valérie Burri, ‘Visual rationalities: Towards a sociology of images,’ *Current Sociology*, 60, 1, Jan 2012, pp.45-60.

what Burri and Dumit developed in their concept of the Social Studies of Scientific Imaging and Visualisation (SIV)[30]. Here, the sociotechnical negotiation of images and their meanings is central, because next to human interpreters and images there is a wide array of technical actors and influencing factors such as different algorithms and parameters (e.g. scale and resolution) involved in the processes of image production, interpretation, and use. What seems to be clear for techno-scientific fields, one example being the interpretation of magnetic resonance imaging (MRI), is a rather new aspect when it comes to the DH and its use of computational tools whose influence in and on the analysis is often underestimated. A key question referring to both fields concerns the visual expertise: who is actually able to read images and who is allowed to read them because visual expertise is its own form of literacy and specialisation[31]? When it comes to the engagement with images, the focus is on the process of making visual data meaningful. This leads to the following questions: how is meaning addressed to images and by whom? Are IPAs, and if so to what extent, positioned or perceived as visual experts in this regard? These questions can only be answered adequately if we have a clear understanding of how much agency and authority is or can be ascribed to IPAs and how these are integrated within sociotechnical assemblages[32]. From our point of view, the question needs to be how human and machine vision can be *synchronised* in the best possible way, rather than if humans or machines have more visual expertise. A follow-up question is how to combine the mutual strengths of human and computer vision in a way that they attain synergies.

6 Creating Ground Truth in Computer Vision: Image Processing Algorithms as Ground Truth Machines

An important observation in the described fieldwork, interviews and also apparent in the practical work in computer science is that a crucial requirement in the development of IPAs is the creation of what computer scientists call “ground truth”. Ground truth provides the definition of what should be learned by the algorithm, such as categories of objects, behaviour etc., and as such it is a significant societal element as it defines and standardises what is perceived to be real and true in the world. Ground truth in this sense is a specific powerful form of an ‘interpretation template’ or even ‘truth template’. It is applied to visual material to analyse it in terms of specific categories of interest (e.g., people, faces, gender, age, facial expressions). Thus, subject to its assumed authority, ground truth very much predetermines what people and faces look like, who is male, female, or different, how old people are and even how people might feel. In this sense, IPAs are — to refer to the analytic term “truth machines”, introduced by Lynch and colleagues in their analysis of the pattern recognition technology of DNA fingerprinting[33] — “ground truth machines”, i.e. powerful agents that follow ground truth-specific rules and make them applicable in a large scale to massive amounts of data. In short, ground truth is an essential element for IPAs and computer vision.

To become familiar with how a machine learning method actually learns different ground truth categories, we invite the reader to train their own classifier and for their own ground-truth categories by using the following online demo tool: **https://teachablemachine.withgoogle.com/**

Major questions in the context of ground-truth are: how is it actually created, and by whom? With a first basic answer being that it is constructed in a sociotechnical way. That is, humans and technology work together. A more elaborated answer that we frame is: it is a process between the application of tacit and explicit knowledge and between specific computational scripts and experimental imaginations.

[30] Regula Valérie Burri, and Joseph Dumit, ‘Social Studies of Scientific Imaging and Visualization,’ in Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, eds, *The Handbook of Science and Technology Studies. Third Edition.* The MIT Press, 2008, pp. 297–317.

[31] Regula Valérie Burri, and Joseph Dumit, ‘Social Studies of Scientific Imaging and Visualization,’ in Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, eds, *The Handbook of Science and Technology Studies. Third Edition.* The MIT Press, 2008, p. 302.

[32] Lucy Suchman, *Human-Machine Reconfigurations. Plans and Situated Actions. 2nd Edition,* Cambridge University Press, 2007.

[33] cf. the term “Truth Machines“ in Michael Lynch, Simon Cole, Ruth McNally, and Kathleen Jordan, Kathleen, eds, *Truth Machine: The Contentious History of DNA Fingerprinting*. University of Chicago Press, 2008.

7 Between Tacit and Explicit Knowledge

The basic resource for building ground truth is knowledge. A specific ground truth is not there from the beginning. Creating ground truth requires existing knowledge about the entity of interest from the people who generate it (usually computer scientists, students or crowd workers). There is no general "truth" from which to start, i.e. the one and only ground truth in the literal sense of the word that one could draw on universally, but a specific "truth" that has to be negotiated and created by humans. That is, by human selection and interpretation, especially the selection of images, semantic content and the context of the given task. The algorithm has to be taught this knowledge therefore it needs to be fed with example data for training purposes. To give an example: if the task is to automatically recognise individual persons from their face (i.e. face recognition), computer scientists need to show the algorithm images of the corresponding persons. In addition, the scientist has to label this particular person with a name. In this case, the face identity represents the ground truth. A further example is to classify scenes shown in images as either indoor or outdoor scenes, a ground truth for image classification needs to be established that labels images into indoor and outdoor scenes. This means, the correct answer—what is indoor/what is outdoor—is provided explicitly in the training phase of the IPA. In this sense, a ground truth represents the correct or true assignment of input data to well-defined and pre-defined classes and concepts of interest they belong to in the real world.

The crucial question is: who is actually able to give the correct answer? In our previous interdisciplinary research, we have observed different ways to obtain ground truth data. One way is to generate it automatically together with the data, e.g. if synthetic test data is employed. Sometimes ground truth is already available, e.g. from previous manual investigations and it only needs to be converted into a machine-readable format. However, in most cases ground truth needs to be generated manually by annotating the corresponding objects of interest. This process of manual ground truth generation is often referred to as "annotation", a process posing several challenges. Annotating image data is a complex and time-consuming process that in the everyday practice of computer vision is a compulsory basic task. Taking these characteristics into consideration, it is often the case that students or interns annotate images. Sometimes this task is also outsourced to a crowd of people (crowdsourcing). This practice is particularly common in situations where computer scientists assume that no special expert knowledge is required to assign the classes or concepts of interest to the data. This form of "what everybody knows knowledge"[34] corresponds to rather more informal, fluent and changing forms of knowledge that are called 'tacit'[35,36] or 'non-explicit' knowledge. An example from fieldwork observation is as follows: the ability to recognise whether something is machine written or hand written might be clear for most literate people that are used to both types of writing. There might also be tacit agreement about this recognition task, suggesting that expert knowledge for this specific recognition task would not be needed. The ability to recognise and differentiate between machine written and hand written texts does not appear as something specific, but as something self-evident ("what everybody knows"). Another example for "what everybody knows knowledge" might be gender recognition. As long as we assume that there is a clear distinction between male and female persons in a binary gender system, this task is indeed a matter of everyday tacit knowledge. However, as soon as this binary system is challenged by more elaborate systems of gender categories or any other form of disruption (e.g. masquerade in carnival), it might become a task for experts in gender recognition (e.g. human ethologists) or even an infeasible task. Finally, in most cases the developing computer scientists have to make the decision whether to consult experts of a specific field (and if so, what kind).

In any case, in theory there is need to use consistent annotation vocabulary and an annotation protocol must be defined which provides a detailed and explicit guide to the annotating person. The annotation protocol defines the classes or a taxonomy of classes, their corresponding labels, categories, characteristics, and their relationships. However, in practice—and this is a crucial point — different classes and concepts are often ambiguous and assessed

[34] Diana E. Forsythe, 'Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence,' *Social Studies of Science*, 23, 3, 1993, p. 458.

[35] Michael Polanyi, *The Tacit Dimension*, Doubleday & Co, 1966.

[36] Harry M. Collins, *Tacit and Explicit Knowlegde*. University of Chicago Press, 2010.

subjectively by annotators due to cultural influences, different background knowledge and interpretation. Especially for concepts with a certain level of semantic complexity, ambiguities in interpretation arise. An example stemming from previous research in the DH are scenes in a film[37]. The segmentation of a film into scenes allows for different possible interpretations because there is no unique definition of a "scene" that covers all possible types of scenes that may occur in films and thus the boundaries between scenes can often not be clearly defined. Furthermore, even for much simpler concepts, such as shot cuts which have low semantic complexity, ambiguities have been reported when it comes to concrete annotations[38]. The major challenge in the creation of an annotation protocol is the specification of well-defined distinct classes or—sociologically speaking—"islands of meaning"[39] to avoid ambiguities. In addition, experience has shown that it is beneficial for the process of ground truth generation to allow for annotators to highlight ambiguous cases with a predetermined label instead of forcing them to decide for one of the pre-defined categories. This approach helps to identify unclear cases, ambiguities and inaccurate definitions of categories early in the process and thereby fosters knowledge gain.

8 Between the Computational Script and Experimental Imagination

While still most computational tools used in the DH operate on a low-level (e.g. shot cut detection[40]) or mid-level of semantic complexity (e.g. gender recognition), a challenge for future tasks is the visual analysis at higher levels of semantic complexity. On this level, relations of objects and more complex actions of these objects need to be interpreted automatically[41]. The human power of imagination generates an unlimited array of possible tasks to recognise automatically, such as the automated recognition of same-sex couples and their representation in films, the analysis of gait velocity in order to measure the historic development of acceleration in TV-series, or the recognition of kissing and killing and their relation in Hollywood blockbusters. Nevertheless, there might be limits to the realisation of automating these. On the one hand, these limits relate to the *computational script*. The computational script refers to the question of what can be realised with specific computer vision solutions and how does technology influence, restrict and prescribe the creation of a ground truth. On the other hand, these limits connect to the narrow universalisation of specific domains. Every algorithm emerges in a specific context at a specific place and at a specific point of time. In addition to this, many IPAs and therefore many ground truths are custom-made for specific tasks and data. A simple transfer to other contexts is often not possible without loss of scope or depth or this transfer might even create forms of bias. A simple process of universalisation is at least questionable.

To demonstrate this and possible implications, we bring in the example of creating a ground truth for the task of automatically recognising the fall of people (that was observed during ethnographic fieldwork of the first author). The fall of persons might be a result of different events, such as being killed in a violent act or losing one's balance as it frequently is the case with elderly people in their homes. While the latter is connected to environments of "Ambient Assisted Living" (AAL) and in-house surveillance cameras or visual sensors, the challenge remains the same once an IPA needs to recognise falls of people in the context of killing events. Both scenarios make use of visual content for a similar type of action (i.e. falls). In the case of automated fall detection of elderly people in private homes, a challenge for computer vision is to differentiate between critical falls, which need emergency assistance, and other uncritical

[37] Mitrovic Dalibor, Stefan Hartlieb, Matthias Zeppelzauer, and Maia Zaharieva, 'Scene Segmentation in Artistic Archive Documentaries.' *HCI in Work and Learning, Life and Leisure, LNCS*, 6389, 2010, pp. 400-410, Springer, Berlin/Heidelberg.

[38] Anton Fuxjäger, 'Wenn Filmwissenschaftler versuchen sich Maschinen verständlich zu machen - zur mangelden Operationalisierbarkeit des Begriffs "Einstellung" für die Filmanalyse,' *Maske und Kothurn*, 3, 2009.

[39] Eviatar Zerubavel, 'Lumping and Splitting: Notes on Social Classification,' *Sociological Forum*, 11, 3, 1996, pp. 421-433.

[40] Matthias Zeppelzauer, Dalibor Mitrovic, and Christian Breiteneder, 'Analysis of Historical Artistic Documentaries,' in *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services,* Klagenfurt, Austria, 2008, pp. 201-206.

[41] Pavan Turaga, Rama Chellappa,R., V.S. Subrahmanian, and Octavian Udrea, 'Machine Recognition of Human Activities: A Survey,' in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, Nov. 2008, pp. 1473-1488.

actions, such as bending forward or lying down on a bed or couch for resting. This means, while the physical process of falling might look similar in many different cases, there is a wide array of what this physical process really means (e.g. critical fall or a part of lying down for taking a rest). In the AAL case of elderly people falling in their homes, ground truth was constructed within the framework of the computational script of a Microsoft Xbox Kinect sensor as well as through the ingenuity and creativity of the researcher in de-scripting[42], rewriting and making use of this specific hardware-software-data assemblage. The researchers used a Kinect sensor because at that time it was considered a keenly priced and affordable hardware and to keep down future prices of an imagined commercial system. Next to the relatively low price, the Kinect sensor also had the advantage of delivering 3D information using a depth sensor, meaning it was possible to estimate distances. It was further privacy enhancing, since in depth images individual persons can hardly be recognised as only the body shape is outlined. The Kinect sensor promised several advantages for the researchers in this specific case but in order to benefit from these it was necessary to first understand and then "de-script" its specific characteristics. So all that followed in the research process took place within this specific script of the Kinect sensor. As such, the existing product Microsoft Xbox Kinect and its configuration was "inscribed" into the ground truth and thereby implicitly into the entire process of fall detection. This strong dependency on a specific hardware and type of data (depth images) makes the approach specific to the problem at hand but hinders its application to other fall detection problems, for example fall detection in films. Even if there is a solution to tackle the transformation problem of Kinect 3D input data to 2D video data, the respective ground truth would hardly be transferable to the other problem because the nature of the falls, their visual appearance and the contextual embedding is different. The adaption of an algorithm to a certain problem always introduces *bias* with respect to the data and the targeted task.

The other fundamental element observed in the construction of ground truth was an experimental investigation into the dynamics of falling within the context of AAL environments as there was no substantial or applicable knowledge or even images or videos available of "real" or "realistic" falls of elderly people and thereby a lack of suitable training data. Consequently, it was a matter of experimental imagination how "real" or "realistic" falls of elderly people look like. As there were no training sequences showing real falls of elderly people, the training sequences had to be created synthetically by the young computer scientists (by filming them simulating a fall) to train the algorithm. The ground truth was defined within this framework and a mathematical equation was formulated for the detection of falls. This equation was based on the relationship between two central visualized geometric elements; first, the orientation of the floor plane and second, a straight line visually representing the human body (medial axis) in the scene. The assumption was that the more similar the orientation of these two elements was, the more likely it was that a fall had occurred. Once a specific critical threshold of this relation was reached, a fall was declared to be detected. From this observation can be concluded that ground truth, and connected to it the specific decision rules, should be generated for the specific material to avoid ambiguities as best as possible and to increase the likelihood of a correct detection. Similarly, for detecting people in films being killed and hence falling, a specific training dataset and ground truth needs to be established to enable a robust detection. To sum up, the data-centric nature of IPAs makes them less generalisable to other type of material or to the same actions in different contexts. This means, there is always a tradeoff between generalisation/universalisation and the accuracy of an algorithm. For the use of computer vision in the DH, this means that the more specific a given research topic or research question is, the more important it gets to actively participate in the generation of ground truth to improve the adaption of the tools to the actual problem.

9 Emerging challenges: Evaluating Error and Bias

In the sequel, problems of ground truth construction and IPAs such as false negative and false positive detections and systematic biases are discussed. How can we actually know that the delivered results are "correct"; which leads

[42] Madeleine Akrich, 'The De-Scription of Technical Objects,' in Wiebe E. Bijker, John Law, eds, *Shaping Technology/Building Society. Studies in Sociotechnical Change*. The MIT Press, 1992, pp.205-224.

to the follow-up question of how we can assess if the ground truth itself is correct? The crucial point is that evaluation depends on "consistent results no matter who is doing the coding" that help the method gain "…credibility as an accurate measure of reality."[43] Thus, if human coders (the ones who previously annotated the visual material) initiating the evaluation come to an agreement that the tested algorithm was accurate in any random case, then the algorithm is accurate. Therefore, it is a "closed loop" system that works "in terms of the system's internal agreement—that is, whether the classification system consistently agrees with itself"[44].

What is at stake becomes apparent when talking about false negative and false positive detections. False negatives are relevant cases (e.g. falls) that are not detected as relevant (i.e. missed). False positives are irrelevant cases (e.g. a curtsy) that are falsely detected as relevant. Usually the performance of algorithms is measured by computing false negative rates and false positive rates. Here it has to be noted that the concept of gathering false negative and false positive rates does always imply that there is one universal ground truth with which any domains of scrutiny are examined to evaluate accuracy. If in a specific case or context the detection rate of a certain class or object is rather low, this does not automatically mean that the specific class or object does not appear. It might be the case that the object is of course there but not detected because the algorithm has a systematic bias and is not able to detect it in the given context. To give an example: if a ground truth of people being killed (in films) was generated solely based on training images deriving from car accident scenes, sequences of people being killed in a shoot-out may not be recognized correctly. The goal in the development of IPAs is always to achieve a large generalisation ability, i.e. to obtain robustness to different contexts, perspectives, scales, and appearances of the target entity. At the same time the number of false detections should be minimised, which is usually a conflicting goal to generalisability. Thus, a tradeoff has to be found for each concrete task and application.

Furthermore, there is the risk of generating algorithmic bias. For example, Introna and Wood analysed the politics and implications of face recognition technologies[45]. One of their central results was that facial recognition algorithms have a systemic bias: men, Asian and Afro-American populations as well as older people are more likely to be recognised than women, white populations and younger people[46]. Bias in gender and race was also discovered in a more recent evaluation of three commercial gender classification systems. Using another image dataset, it was shown that darker-skinned females belong to the most commonly misclassified group with error rates of up to 34,7%[47] Similarly, in the early years of speech recognition a constant bias of speech recognition systems towards male speech was observed. This bias originated from the fact that it was mainly men who developed those systems at this time and they used their own voices to record their training data (speech sequences) due to a lack of large training corpuses. This inherent tendency to bias bears the risk for a "new type of digital divide"[48] that requires close attention in research. Therefore, Introna and Wood call for "bias studies," especially regarding the question of what can be done to limit biases. While there will be always be forms of algorithmic and human bias and we understand that an unbiased algorithm or world might not exist, it seems to be crucial to reduce bias in terms of transparency and algorithmic explanation possibilities as more and more political decisions are grounded in algorithms. Another example for scholars researching and fighting bias in machine learning is Joy Buolamwini.

[43] Kelly Gates, *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*, NYU Press, 2011, p. 171.

[44] Kelly Gates, *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*, NYU Press, 2011, p. 171.

[45] Lucas Introna, David Wood, 'Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems,' *Surveillance & Society* 2 (2/3): 2004, pp. 177–198.

[46] Lucas Introna, David Wood, 'Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems,' *Surveillance & Society* 2 (2/3): 2004, p. 190.

[47] Joy Buolamwini, Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,' *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR*, 81, 2018, pp.77-91.

[48] Lucas Introna, David Wood, 'Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems,' *Surveillance & Society* 2 (2/3): 2004, p. 192.

Video 2. Ted Talk: Joy Buolamwini on fighting bias in algorithms (TEDxBeaconStreet, November 2016)

10 Conclusions: Towards User Participation and Active Learning

What are the consequences and implications for the Digital Humanities and for the use of automated visual analysis tools in this domain? While there are promising results to extract visual information on a low-level of semantic complexity, the higher-level interpretation of (moving) images is highly dependent on situation and context. Drawing meaningful conclusions from situation and context is a major challenge for computer vision algorithms because it requires a high-level of understanding of objects, relations, diversity, ambiguity, situated actions[49] and cultural local particularities (e.g. greeting rituals, political symbols). Therefore, we argue in support of an integrated and critical approach to the use of computer vision tools, whereby we attach particular importance to transparency and the involvement of (expert) users (e.g. scholars in the DH) in ground truth creation and the training process of the algorithms.

We argue that a major limitation for the application of existing (pre-trained) computer vision approaches in DH is their limitation to a certain and previously defined ground truth. Research questions addressed in the DH often relate to specific high-level concepts, which are not covered by existing ground truths, making the training they received insufficient. Since there is no unique and all-encompassing ground truth, we propose to move away from rigidly pre-defined ground truths to more flexible ground truths that adapt to the specific requirements of the actual expert users. We believe that this is a useful strategy in particular in DH-related research where highly complex and hitherto not analysed concepts are the subject of investigation. Furthermore, the ground truth is often not known a priori and ground truth concepts may evolve or are discovered during the analysis of the visual content under consideration. Such a flexibility should be provided by computer vision tools to support exploration and the establishment of hypotheses.

A promising approach to enable a more flexible learning is “active learning” (AL)[50,51]. In AL the algorithm is not trained in an offline manner from pre-existing ground truth like in most existing approaches today. Instead, the algorithm is

[49] Lucy Suchman, *Human-Machine Reconfigurations. Plans and Situated Actions. 2nd Edition,* Cambridge University Press, 2007.

[50] Burr Settles, *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009.

[51] Devis Tuia, Michele Volpi, Loris Copa, Mikhail Kanevski, and Jordi Munoz-Mari, ‘A survey of active learning algorithms for supervised remote sensing image classification.’ *IEEE Journal of Selected Topics in Signal Processing*, 5, 3, 2011, pp. 606-617.

trained in an online fashion by taking input from the user into account. Thereby the (expert) user and her/his needs are directly integrated into the learning process, which replaces the need to define a ground truth explicitly. This is best explained with an example. Imagine that a DH scholar wants to find political symbols (e.g. flags, graffiti, posters etc.) in a large corpus of videos (e.g. documentaries or news broadcasts). With a very high probability, no algorithm exists which has been trained for detecting such specific symbols. AL can circumvent this problem. In AL the user may provide one or a few examples of symbols that she/he is interested in for initialisation. Starting with these examples, the algorithm tries to train an initial detector for the desired symbols. Concerning symbols that the classifier is not certain about, it can ask the user who can then assign the symbol to a specific category. This queried feedback is used by the algorithm to improve its detection capabilities iteratively. Additionally, the user may provide feedback on the relevance of detection results and communicate them back to the algorithm, i.e. whether a certain detected symbol is relevant or not for the user. This type of feedback mechanism is also referred to as “relevance feedback”[52] and can easily be combined with AL. More recent developments further extend the idea of AL to interactive data exploration methodology (i.e. visual interactive learning, VIAL) to further enable the user to proactively select items for labelling to better guide the training process[53].

The different strategies for incorporating user feedback into the training process makes the algorithm more adaptive to the actual data and thereby enables it to better fit to the actual research questions investigated by the user. AL further circumvents the explicit and a priori definition of a ground truth. Instead, the ground truth is defined implicitly from the expert user’s feedback. This approach makes it possible to adapt directly to the needs of the user and release her/him from the explicit definition of categories and typologies of classes, i.e. the annotation vocabulary. AL can lead to a higher level of transparency and understanding of the algorithms by making the training data explicit to the user. The interactive nature of the AL process empowers the user and enables her/him to learn from the data and to get a higher level of understanding of the problem as well as novel insights from the data. Thereby, the “black box“ of today’s computer vision algorithms can be opened to a certain degree and transformed into a “grey box” which enables a basic level of transparency and documentation. In this regard, the methodology connects well to the research on explainable artificial intelligence (XAI)[5,6]. The goal of XAI is to explain the decisions of a machine-learned algorithm and how the trained model internally works. This enables verifying if the correct patterns (e.g. visual patterns in our case) are learned for a certain class and to answer the question if the algorithm “looks” at the right spots when making a decision (e.g. to eyes and nose when detecting a face). Furthermore, explainability mechanisms enable us to identify biases learned from the data[54]. An online visualisation tool for complex neuronal networks trained from image data is available under: **http://shixialiu.com/publications/cnnvis/demo/**. This demo shows what kind of visual patterns are learned in the individual network layers. Another interactive online tool for the visualisation of the internals of a classifier is available under: **http://scs.ryerson.ca/%7Eaharley/vis/conv/flat.html**. Here the user can generate input data themselves and trace how the decision of the classifier is made. We argue that the combination of XAI techniques with AL is especially promising as both types of approaches are likely to exploit mutual benefits.

By integrating the user’s feedback into the process, the training can be better guided by the needs and intentions of the user. This in turn fosters the generation of more useful and targeted algorithms and tools for expert users[4]. Especially in the DH, where the research questions are often very specific and semantically complex, we believe that such a user-guided approach is a promising solution. The combination with XAI approaches will be essential in the future not only to verify decisions made but also to discover biases learned from the data and to better understand false detections (false positives and false negatives).

[52] Xiang Sean Zhou & Thomas S. Huang, ‘Relevance feedback in image retrieval: A comprehensive review.’ *Multimedia systems*, 8, 6, 2003, pp. 536-544.

[53] Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner, ‘VIAL - A Unified Process for Visual-Interactive Labeling.’ *The Visual ISSN (TVCJ)*, 34, 2018, pp. 1189, Springer: Berlin, Heidelberg.

[54] Brian, Hu Zhang, Blake Lemoine, & Margret Mitchell, ‘Mitigating unwanted biases with adversarial learning.’ *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335-340.

Biographies

Christoph Musik, Dr. phil. is a postdoctoral researcher at the Institute of Media Economics at St. Pölten University of Applied Sciences (Austria). He is the Lecturer in Technology and Society in the Department of Science and Technology Studies at the University of Vienna and a lecturer at the Institute of Sociology at the University of Innsbruck. Since 2015, he has been co-speaker of the section Sociology of Science and Technology of the Austrian Association for Sociology (ÖGS).

Matthias Zeppelzauer is a senior researcher at the Institute of Creative Media Technologies at St. Pölten University of Applied Sciences (Austria). He received his PhD in Computer Science from Vienna University of Technology in 2011 with highest distinction. His research focuses on content-based retrieval, computer vision and machine learning with a special focus on active and user-centered learning. Matthias received several performance scholarships from the Vienna University of Technology and was awarded by the Austrian Computer Society for outstanding achievements in the area of pattern recognition.

DOI: 10.18146/2213-0969.2018.jethc153
Publisher: Netherlands Institute for Sound and Vision in collaboration with Utrecht University, University of Luxembourg and Royal Holloway University of London.

Copyright: The text of this article has been published under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Netherlands License.
This license does not apply to the media referenced in the article, which is subject to the individual rights owner’s terms.