EUROPEAN JOURNAL OF MEDIA STUDIES
www.necsus-ejms.org

How machines see the world: Understanding image annotation

Carloalberto Treccani

NECSUS (7) 1, Spring 2018: 235–254
URL: https://necsus-ejms.org/how-machines-see-the-world-understanding-image-annotation/

Keywords: algorithmic vision, human vision, image annotation, machine vision, resolution

Introduction

In the context of machine vision, image recognition refers to the ability of machines and algorithms to identify people, places, objects, gestures, or other subjects in a given image. Self-driving cars, for instance, use machine vision systems to locate road signs, vehicles, pedestrians, and cyclists, to understand the three-dimensional space in which they are located and avoid collisions. Facebook uses facial image recognition systems to identify photographs in which a person is present but not tagged, and to help visually impaired users identify people in a specific video. However, although it is simple for a human being to make sense of an image and identify its content, these operations are still particularly complex for machines and algorithms. In this paper we will investigate how machines and algorithms are trained for image recognition purposes. To start, the tasks performed by human annotators will be discussed. Part two will outline some problems with this process, and part three will present further thoughts and reflections on the subject. The terms 'machine vision' and 'algorithmic vision', which often appear throughout this paper, replace the term 'computer vision' and are used in a broad sense that seems to better reflect contemporary reality.

Crowdsourcing platforms and annotation

Data have a key role in the progress and improvement of *visually intelligent machines*. Until recently, collecting a large amount of data was a difficult, expensive, and time-consuming task; however, thanks to crowdsourcing platforms the prospect today has radically changed. These platforms 'offer an inexpensive method to capture human knowledge and understanding, for a vast number of visual perception tasks'.[1] Through these platforms, big companies like Amazon (Amazon Mechanical Turk) can hire a large number of digital workers, who manually annotate images presented to them. Working from home at their computers, these digital annotators describe, pigeonhole, mark, segment, and frame images. For example, when a strawberry is shown on the screen, they will label it 'strawberry' (object classification). All tagged images are then organised into semantic areas based on their labelling, and later collected in databases used to train machines and algorithms. But what does 'annotation' mean? To annotate means to define areas in an image and assign them a value. The information, or metadata, can be for instance a series of keywords that attribute a semantic value to the chosen portion of the image. To create a machine vision system able to automatically find a cat and define its location in a picture, for example, a large collection of manually annotated images is required. The tasks digital workers are assigned reflect ones that will subsequently be performed by machines and algorithms. These tasks include:

Object classification (Fig. 1): determining whether an object is present or absent in the image (*Is there a cat in the image? Are human beings present in the image?*).

Object detection (Fig. 2): identifying a particular object and its arrangement in space (*Where is the dog located?*). In this case, the worker is asked to draw a bounding box around a single object.

Scene classification (Fig. 3): classifying a given environment. Questions such as *Is the building a museum or a hospital?* are presented to the annotator, who has to assign the corresponding label.

Image segmentation or pixel-level image segmentation (Fig. 4): determining which object a pixel in the image belongs to. The worker is asked to outline single objects' profiles and annotate every area separately.

Attribute recognition (Fig. 5): defining the visual properties or qualities of objects – how an object looks and not just where is it located. The worker is asked to choose adjectives that describe the object (*Is the scene 'cold' or 'hot'?*).

Three different strategies are later used to ensure annotation quality. One is the creation by an expert of a gold standard,[2] which is secretly inserted into the images to verify the annotators' work. Second, workers may be asked to rate and correct other workers' answers, and third, a large number of annotations per image are collected and compared. The third strategy has become the most popular, and has proved a solid method to acquire high-quality labels and eliminate subjectivity. Although manually labelled images are essential for effective machine vision systems, the methods used for annotating these images are particularly problematic and controversial at the present time. In most cases, annotators are not required to possess special skills or knowledge, and therefore labelling work can be easily outsourced to online marketplaces such as Amazon Mechanical Turk. After workers complete and submit the HITs (Human Intelligence Tasks) following the requester's instructions, they are remunerated a few cents per image. As previously stated, these methods of data collection and labelling have deeply helped the development of machine vision. Nevertheless, they have significant implications and problems. These issues will be further analysed, and the effectiveness and correctness of these methodologies will be investigated from different points of view.

Fig. 1

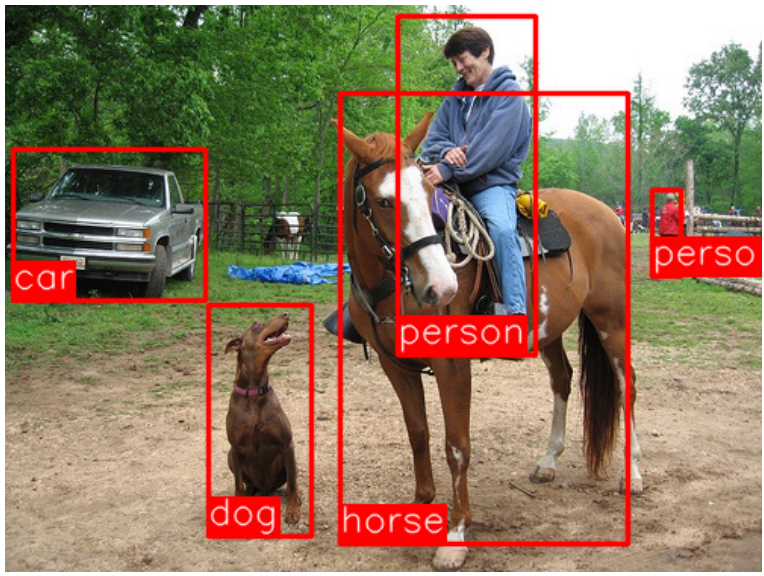

Fig. 2

Fig. 3

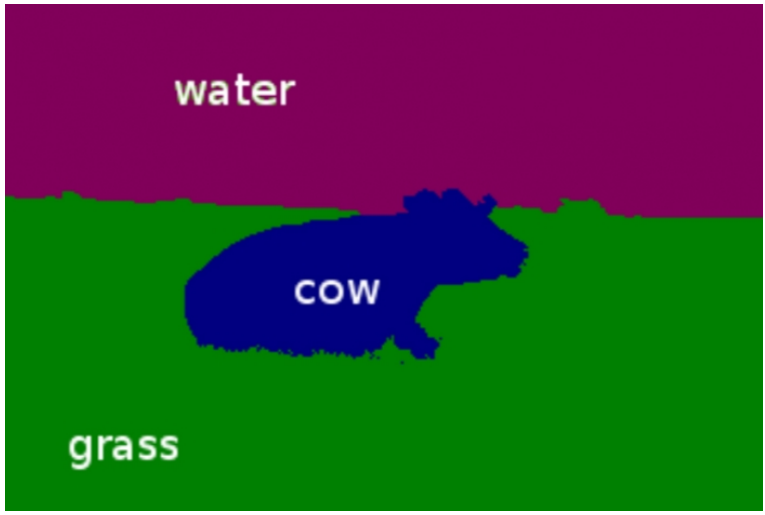

Fig. 4

Fig. 5

Problems related to image annotation

In this section, three different problems related to the training process of machine vision systems will be presented and discussed: experiential problems and the differences between machine and human vision; terminological and

visual problems occurring during the annotation process; and ethical problems related to annotators' work conditions.

Machine, algorithmic, and human vision

> If you consider a child's eyes as a pair of biological cameras, they take one picture about every two hundred milliseconds – the average time an eye movement is made. So by age three, a child would have hundreds of millions of pictures of the real world. That is a lot of training examples. So instead of focusing on solely better and better algorithms, my insight was to give the algorithms the kind of training data that a child was given by experiences, in both quantity and quality. [3]

In his article, Nicolas Malevé quotes Professor Fei Fei Li, Associate Professor at the Computer Science Department at Stanford University, who opens her TED Talk with the words just quoted. Maleve is struck by the equivalences between humans and machines in her speech and the resulting simplifications, commenting,

> eyes and cameras, experience and training, looking and taking pictures. The computer becomes more biological, while the child becomes more robotic. Or, to paraphrase it, the biological and the mechanical flow from one figure to the other, blurring their borders. [4]

This simplification implies that the human experience, in its complexity, could be replicated by a machine. Nevertheless, even if machine vision is constantly and quickly developing, the vision of a human being is much richer and remains difficult to emulate. In the same way that a child builds his visual system and his ability to distinguish and attribute meaning to what he sees, machine vision systems are trained by providing them with a series of images (data) that recreate, or rather try to reproduce, the visual experience of a child. The experience of a machine continues, however, to be extremely limited both in quantitative and qualitative terms. As Li states,[5] the visual experience of a child at the age of three is made up of hundreds of millions of images, while to date *ImageNet*,[6] the largest recorded archive of existing images, consists of just over 14 million images. Therefore, compared to the experience of a small child, machine vision seems inadequate and highly limited. Ultimately, as Beau Lotto explains,[7] what we see is the result of our experience and our interaction with the world. Our vision does not necessarily represent the world as it is, but rather a model of the world with which to interact, which is the result of experiences, behaviours, and past actions. It is an image of the world that is functional for us but not necessarily

true. Depending on the perceiver's context, behaviour, past experience, and aims, objects are recognised differently (e.g. as a cat, a feline, an animal, or living being).[8] Our brain, in fact, creates a sense of the world by physically interacting with it. However, this interaction, this experience of the world, is either completely absent or very limited in the case of machine vision systems.

Terminological and visual problems

During the labelling process, one or more terms are assigned to an image. These terms are used to define the picture, its parts or objects, subjects and entities. The moment an image presents itself to the annotator, however, many problems related to the selection and use of terminology appear. First of all, is it possible to reduce an image, a visual experience, to a mere group of words? Is it possible to translate visual information into language?

> Paradoxically speaking, meta-dating makes us deaf towards images. Western cultural competence and technology of finding, transferring, and processing stored images has been marked by the supremacy of the word as instrument and medium of control and of navigation, such as catchword translation of image contents and the titling of authors and works. Iconography is the essence of a text-based grip on images (comparable to Optical Character Recognition), trying to reduce the informational richness of an image to the clarity of verbal semantics. [9]

A word is thus used to determine an image, to forcibly link it to a field of conventions whose aim is to define the most univocal meaning possible.[10] The effect is then a labelling process that involves an undeniable rigidity of terminological choice and a high degree of simplification, which in turn fails to capture the richness of vision. Adela Barriuso, in 'Notes from an annotator' (2012), in collaboration with Antonio Torralba, describes several issues related to the annotation process. How does an annotator approach and label unknown scenes or objects (e.g. objects in a chemical laboratory)? Barriuso reflects:

> I do not know the name for most of the things in the scene. It looks like a lab, but how would you name the tables? Would you call them 'lab tables', or 'work benches'? I generally skip any picture that I find difficult to label right from the beginning when I open it. [11]

In some cases, the images presented to the annotator do not match her knowledge, and therefore create an obstacle and force the worker to find a solution. The use of synonyms can also be problematic. Although they offer

richness and variety of descriptions, they also represent an issue for algorithms that can have difficulty distinguishing between them. Another detail that is challenging for machines and algorithms is the use of singular or plural forms, or terms in different languages. To solve such problems and limit disagreements between the attributes chosen by annotators, some crowdsourcing platforms establish a list of terms for which models will be trained, called *attribute vocabulary*.[12] However, this method, originally designed to reduce misunderstanding, also places limits on workers, who are induced to prefer specific pre-established terms. Parallel to this solution, other workers choose instead to create their own terminological vocabulary as 'good practice',[13] to reduce terminological 'noise'. Again, although this 'good practice' seems to be an efficient modality, it also represents a particularly restrictive use of terminology. The richness of the visual panorama is thus impoverished by these 'good' practices.

Two additional cases are particularly problematic for annotators: describing an object that is partially hidden by other elements in the image, and objects reflected by surfaces such as mirrors or present in transparent containers (e.g. a food container [Fig. 6]). Barriuso explains in detail the problem and the solutions she has adopted. She states that if an object is partially hidden, she writes 'occluded' on it. However, in the case of a series of books on a shelf, where only the spines are visible, Barriuso labels them as books because these objects are placed as we are used to seeing them. However, this personal choice may not be shared by other annotators, who in turn will choose a different methodology to deal with the matter. Concerning reflected objects, Barriuso again questions how these objects should be labelled:

> […] should I call it a 'cake' or should I call it 'container'? I decided to name it a 'container' because I do not label the things that are visible behind a crystal or something transparent. This is a rule that I always follow. […] I never label the objects that are inside cabinets and that are visible behind glass doors. I also never label anything behind a closed window. I am not sure if this is the right thing to do, but in many cases one has to adopt some criteria (unless somebody corrects me). There are so many open windows that I ask myself: why should I also label the objects that are behind closed windows? [14]

Fig. 6

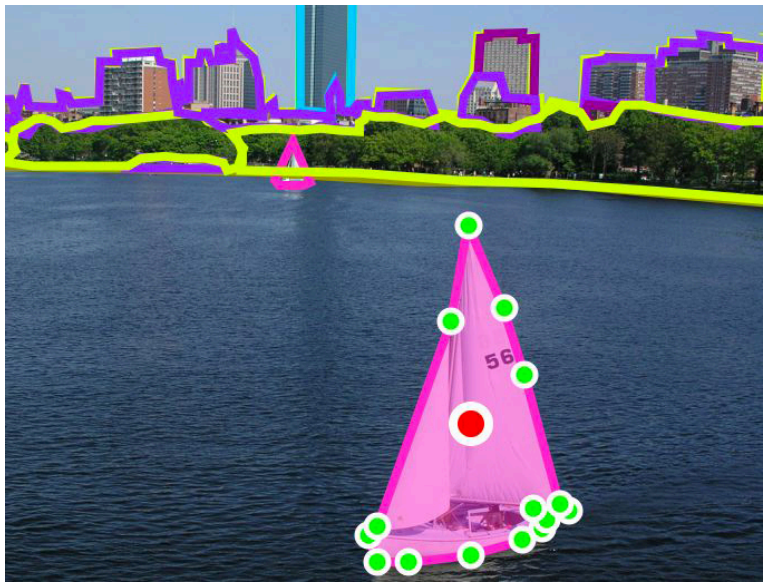

Fig. 7

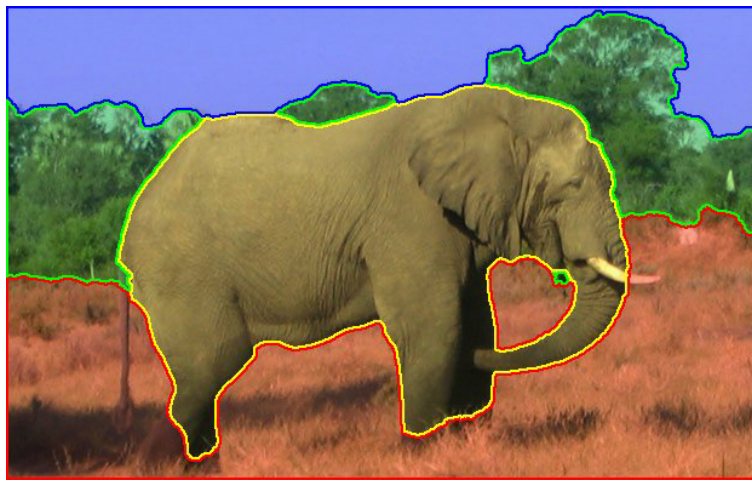

Fig. 8

When one looks at an image, a part of visual discernment surpasses the mere visual configuration; it is for this reason that individuals respond differently to what is depicted. The words chosen to describe it are completely subjective, incorporating the individual worker's ideas, prejudices, and preconceptions. As Baxandall states in *Painting and Experience in 15th Century Italy*, every act of vision can be interpreted as a simplification or distortion, which derives from the historical context and personal life experiences. Such life experiences can only be understood in their completeness and complexity, comprising a plurality of sounds, images, and words that require the individual to actively participate in the interpretation of the world.[15] In the same text Baxandall engages with the terminological problem, writing about the use of specific terms such as 'aria', 'puro', and 'compositione' to describe paintings and pictorial quality.

> We can use them now as a complement and stimulus, and naturally not as a substitute, for our own concepts; they will give us some assurance of not altogether losing sight of what these painters thought they were doing. Quattrocento intentions happened in Quattrocento terms, not in ours. [16]

These words belong to a different era (the 15th century), one that is difficult even for an Italian of the 21st century to comprehend, and they were used in a completely different way compared to those adopted today during the labelling process. Indeed, they were not intended to offer an exhaustive description of a painting or a detail of it, but rather to provide a possible interpretation. These terms fit perfectly within the historical cultural context of

the 15th century, 'embodying in themselves the unity between the pictures and the society they emerged from'.[17] The terminological use in the machine vision context is instead conflicting and not always helpful, but is often a limit for the visual experience of machines and algorithms. This limitation leads to what Baxandall calls a 'systematic rigidity'.[18] The vision is reduced to one or a set of terms which, even if they are detailed, are not able to describe the complexity of the perceptual experience. One further problem is linked to the annotation process. In some cases (e.g. *image segmentation*) the annotator is required to draw a continuous line as accurately as possible, that follows the edges of a specific subject in order to divide or separate an image into significant areas, with the aim of simplifying the representation of an object contained in it, for example into something easier to identify (Fig. 4). Ideally, it should be possible to recognise an object from its contour.[19] This idea, however, implies once again an extreme simplification, and reduces the human ability to understand the world to forms, polygons, and outlines.

> Labeling more than 250,000 objects gives you a different perspective on the act of seeing. After a full day of labeling images, when you walk on the street or drive back home, you see the world in a different way. You see polygons outlining objects, you start thinking about what they are, and you are especially bothered by occlusions. [20]

In addition, the degree of detail with which images are annotated turns out to be a problem, as it is not homogeneous among annotators, and therefore problematic for machines and algorithms.[21] Further questions about how this outlining should be performed can arise. The annotator is often faced with difficult choices about which objects should be traced and labelled, or whether and how a hidden object should be outlined. How precisely, moreover, must the object's profile be traced (Fig. 8)? Taking the image of a tree as an example, it is unclear whether every single leaf should be delineated independently or remain part of the shape of the tree. When a label is assigned, for that matter, should it be identified as a *tree* or an *oak*?

Ethical problems with crowdsourcing platforms

Working on demand from all over the world, web workers involved in the labelling process offer an inexpensive and high-speed solution for a variety of tasks that require the participation of a human being.[18] Nevertheless, ethical concerns about the working conditions involved must be highlighted. The workers, paid mere pennies per image, work in precarious conditions with-

out any labour protection (health insurance, etc.); moreover, if the client considers their work unsatisfactory, payment can be denied without any justification. Workers then sometimes perform substandard work or engage in forms of misbehaviour, doing the bare minimum necessary to receive the HIT payment.[22] Because of these adverse conditions, they often choose to label images and objects even when they lack the competence or knowledge to evaluate the visual information. Furthermore, the worker can be driven to cheat – to click or type randomly or use online resources to give answers that are not accurate, but good enough to ensure remuneration (using online translation services to translate terms in a language unknown to the Turker, or using scripts to solve captcha tests faster).[23] Not infrequently, workers even create programs that automatically complete HITs. Such issues, within the frame of immaterial work, as Hardt and Negri state, 'where labour produces immaterial goods such as a service, a cultural product, knowledge or communication'[24] (with particular attention to the production of knowledge), lead to poor quality or confusing labelling that nonetheless determines the way machines and algorithms understand the world. The problems listed so far lead us to a final consideration. Machines and algorithms can in some cases outperform human vision; they can recognise and remember a large amount of visual information, for instance thousands of types of fish or tree species, beyond any human possibility. They are also able to recognise what they see much faster than a human being. They are therefore capable of a high degree of specialisation and definition. At the same time, they may fail to achieve an overall understanding of complex or particular objects and situations (Fig. 9, Fig. 10),[25] such as relating objects and scenes to create a complete description (*scene description*, e.g. four cats play on a sofa), or fail to understand abstract concepts (e.g. happiness). In this case there is instead a low-grade definition, as there is a wrong or distorted creation of sense of a scene, or more generally of an image. Their visual experience, paradoxically, can be defined as *low resolution* when they must create sense connections, abstract meanings, and manage the complexity and richness of information of a visual perceptive experience (*high-level visual concepts*);[26] while in terms of *high resolution*, they must perform simple, mechanical, and highly specialised tasks. Compared to human vision, that of machines and algorithms seems to be faster, more fragmented and unitarian; a technological vision functional to the logics (of power) that govern these systems and with cultural, political, and epistemological implications (final chapter).

Fig. 9

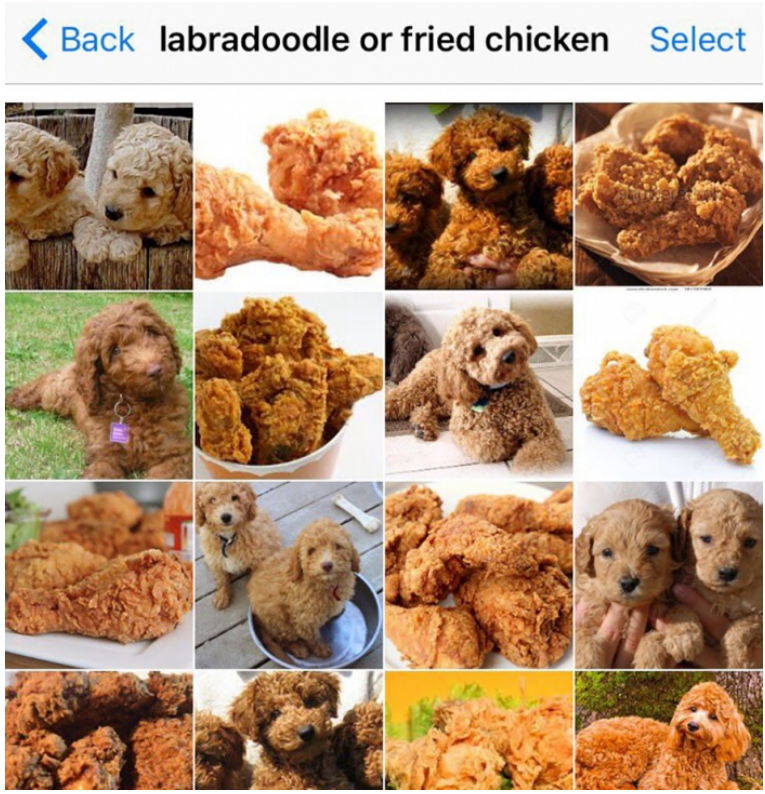

Fig. 10

Vaccari would define this situation as a 'waste dump',[27] where the highest degree of specialisation or definition and the highest degree of homogeneity or nondifferentiation coexist. In the words of McLuhan, they are both cold and hot.[28] Hot (*high-level confidence*),[29] when the ability of 'prediction' of an object in an image (*object classification* and *object detection*, e.g. distinguish different breeds of dogs or locate pedestrians at night) is equal to or higher than human. Cold (*low-level confidence*),[29] when the capacity of a machine or an algorithm to understand an image (*scene understanding*, i.e. recognising and describing a scene) is limited or even wrong and therefore less than human. Differences and implications of this new media reality therefore oblige us to make an effort to further develop the theoretical concepts that have helped human beings to orient themselves in the complex media landscape of the last century, and that, however, cannot simply be superimposed to the contemporary context. This seems to reveal a new type of vision shared with

'intelligent' technological devices.[30] To elucidate this issue, the final part of this paper will present some ideas and proposals.

Conclusions

This paper has described the ways in which data are collected, the tasks assigned to annotators, and problems related to these processes. Even if it has found fault with these methods, its aim is to present different aspects of machine vision to define a starting point for further discussion and development, rather than just offering a critical view of machine learning and image annotation.

As stated earlier, the way machine vision systems see is strongly inspired by the human way of seeing. However, it is possible to reverse the equation and ask whether the opposite is possible: are we outsourcing, externalising our sense of sight? Externalising means delegating, allowing someone or something to perform a certain action on our behalf. We may end up letting our machines, cars, software, and technological devices look for us and take charge of one of the most important human experiences – vision. Delegation also implies trust, abandoning oneself in another's hands and investing someone with a certain power, in this case the power to observe the world in our place. Baxandall has demonstrated the existence of a series of rules that painters of the 15th century were advised to follow. These 'guidelines' explained, for example, how each different hand position depicted in a painting represented a different concept within that cultural context; they were rich and detailed, and helped the painter to address that specific historical and cultural context. These rules both determined the visual habits of Italy in the 15th century and mirrored that society, in a sophisticated game of references and associations. Does a similar link exist today between machines and algorithms and contemporary society? Is it plausible to think of a society where all possible interactions are mediated by machines and algorithms, which thus do not mediate but instead define and constitute a new reality? In this scenario, the significance of human visual sensory experience seems to be increasingly reduced. Through the senses, human beings become aware of the surrounding world, determining it, constructing it and modifying it. However, contemporary visual sensory experience has become mechanical and algorithmically predetermined. What individuals experience is an end in itself, predetermined by software and algorithms that increasingly attribute a new

sense to 'the things of the world'[31] which remain incomprehensible to human beings. Will the anthropocentric vision of the world (a fundamental characteristic of Renaissance thinking) disappear in favour of a new vision shared with machines?[32] A further question presents itself: is it possible to believe that big companies such as Facebook, Google, or Apple are creating a visual monopoly? Again, to quote Nicolas Malevé,

> the computer vision algorithm is immersed in the visual world of millions of people. Nurtured by the Internet, the algorithm has a collective vision tied together by the computer network [… a] vision made of millions of eyes and a collective brain. [33]

From this perspective, machine vision is a 'multiple vision' rich and detailed, the result of the work, actions, and visual experiences of hundreds of thousands of annotators and millions of internet users; this same vision, however, is only summarised in one single vision that is algorithmically driven and in turn mirrors the logics of those who control these machines and algorithms. Therefore it is necessary to reflect on this algorithmic and invisible vision – a different perception that reveals an entirely new panorama that needs further study and analysis. In this unprecedented historical moment, when more machines than human beings analyse and try to make sense of what they see, the challenge is to understand this mechanical and algorithmic vision that influences the way we see the world today and increasingly in the future.[34]

Author

Carloalberto Treccani is a PhD candidate at the School of Creative Media, City University of Hong Kong, and an artist. His research investigates how machine vision is affecting human perception/vision. His artworks have been exhibited in group and solo exhibitions and commissioned by galleries and institutions.

References

Aloimonos, Y. and Rosenfeld, A. 'Computer vision', *Science*, Vol. 253, No. 5025, 1991: 1249-54.

Armitage, J. 'Accelerated Aesthetics: Paul Virilio's the vision machine', *Angelaki: Journal of the Theoretical Humanities*, Vol. 2, No. 3, 1997: 199-209.

Barnard, K. and Johnson, M. 'Word sense disambiguation with pictures', *Artificial Intelligence*, Vol. 167, 2005: 13-30.

Barriuso, A. and Torralba, A. 'Notes on image annotation', retrieved from arXiv:1210.3448, 2012.

Baume, N. *Super vision*. Cambridge: The MIT Press, 2008.

Baxandall, M. *Painting and experience in 15th century Italy*. Oxford: Oxford University Press, 1988.

Beer, D. 'Power through the algorithm? Participatory web cultures and the technological unconscious', *New Media & Society*, Vol. 11, No. 6, 2009: 985-1002.

Bederson, B.B. and Quinn, A.J. 'Web workers unite! addressing challenges of online laborers', Proceedings of the conference *Human factors in computing systems – CHI EA '11*, 2011: 97-105.

Benjamin, W. *The work of art in the age of mechanical reproduction*. London: Penguin Books, 2008.

Bush, P. *Rumor of the true things*: http://www.paulbushlms.com/lms/rumouroftruethings.htm (accessed on 8 February 2018).

Clarke, A. and Tyler, L.K. 'Understanding What We See: How We Derive Meaning From Vision', *Trends in Cognitive Sciences*, Vol. 19, No. 11, 2015: 677-87.

Cowie, R. 'The new orthodoxy in visual perception: Conjectures and doubts about internal processes', *The Irish Journal of Psychology*, Vol.8, No. 2, 1987.

De Rosa, M. 'Poetics and politics of the trace: Notes on surveillance practices through Harun Farocki's work', NECSUS, Vol. 3. No. 1, 2014: 129-149; https://necsus-ejms.org/poetics-politics-trace-notes-surveillance-practices-harun-farockis-work/.

Difallah, D.E., Demartini, G., and Cudré-Mauroux, P. 'Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms', *CEUR Workshop Proceedings*, 2012: 20-25.

Ernst, W. 'Visual im/mediacy: Towards a cultural technology of images': https://www.musikundmedien.hu-berlin.de/de/medienwissenschaft/medientheorien/ernst-in-english/talks-scripts.

Farocki, H. *I believed to see prisoners/Eye* in *Ctrl [Space]: Rhetorics of surveillance from Bentham to big brother*, edited by P. Wiebel, T.Y. Levin, and U. Frohne. Karlsruhe-Cambridge: ZKM/MIT Press, 2002.

Foucault, M. *Discipline and punish: The birth of the prison*. New York: Vintage Books, 1995.

Fuller, M. and Goffey, A. *Evil media*. Cambridge: MIT Press, 2012.

Gallagher, S. and Zahavi, D. *The phenomenological mind*. London: Routledge, 2008.

Gajendar, U. 'Empathizing with the smart and invisible: algorithms!', *Interactions*, Vol. 23, No. 4, 2016: 24-25.

Gray, M.L., Suri, S., Ali, S.S., and Kulkarni, D. 'The Crowd is a Collaborative Network', Proceedings of the *ACM Conference on Computer-Supported Cooperative Work & Social Computing – CSCW '16*, 2016: 134-47.

Hansson, K., Muller, M., Aitamurto, T., Irani, L., Mazarakis, A., Gupta, N., and Ludwig, T. 'Crowd Dynamics: Exploring Conflicts and Contradictions in Crowdsourcing', *2016 CHI Conference on Human Factors in Computing Systems*, 2016: 3604-11.

Hardt, M. and Negri, A. *Empire*. Cambridge: Harvard University Press, 2000.

Hayles, N.K. *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. Chicago: University of Chicago Press, 1999.

Heidegger, M. *The question concerning technology*. London: Garland Science, 1977.

Hjelm, S. 'Visualizing the vague: Invisible computers in contemporary design', *Design Issues*, Vol. 21, No. 2, 2005: 71-78.

Hoelzl, I. and Marie, R. 'From Softimage to Postimage', *Leonardo*, Vol. 50, No. 1, 2017: 72-3.

Horton, J.J. 'The condition of the Turking class: Are online employers fair and honest?', *Economics Letters*, Vol. 11, No. 1, 2011: 10-12.

Johnston, J. 'Machinic Vision', *Critical Inquiry*, Vol. 26, No. 1, 1999: 27-48.

Kelly, K. *What technology wants*. London: Penguin Books, 2011.

Kovashka, A., Russakovsky, O., Fei-Fei, L., and Grauman, K. 'Crowdsourcing in Computer Vision', *Computer Graphics and Vision*, Vol 10, No. 13, 2014: 177-243.

Lash, S. 'Power after Hegemony: Cultural Studies in Mutation', *Theory, Culture & Society*, Vol. 24, No. (3), 2007: 55-78.

Longo, G.O. *Il nuovo Golem: Come il computer cambia la nostra cultura*. Milan: La terza, 2003.

Lotto, B. *Deviate: The science of seeing differently*. London: W&N, 2017.

Malevé, N. '"The cat sits on the bed", Pedagogies of vision in human and machine learning', unthinking photography: https://unthinking.photography/themes/machine-vision/the-cat-sits-on-the-bed-pedagogies-of-vision-in-human-and-machine-learning (accessed on 20 February 2018).

_____. 'The politics of image search – A conversation with Sebastian Schmieg [Part I] and [Part II], unthinking photography: https://unthinking.photography/themes/interviews/interview, (accessed on 15 February 2018).

Manovich, L. *The language of new media*. Cambridge: The MIT Press, 2008.

Martin, D., Hanrahan, B.V., O'Neill, J., and Gupta, N. 'Being a turker', Proceedings of the *17th ACM conference on Computer supported cooperative work & social computing – CSCW '14*, 2014: 224-235.

McLuhan, M. *Gli strumenti del comunicare*. Milano: Il saggiatore, 2015.

Moholy-Nagy, L. *Pitura Fotografia Film*. Milan: Einaudi, 2010.

Mitchell, W.J. *The reconfigured eye: Visual truth in the post-photographic era*. Cambridge: MIT Press, 1992.

_____. *Placing words: Symbols, space, and the city*. Cambridge: MIT Press, 2005.

Neisse, U. 'The Imitation of Man by Machine', *Science*, Vol. 139, No. 3551, 1963: 193-197.

Paglen, T. 'Invisible Images (Your Pictures Are Looking at You)', *The new inquiry*: https://thenewinquiry.com/invisible-images-your-pictures-are-looking- at-you/, (accessed on 10 February 2018).

Panofsky, E. *La prospettiva come forma simbolica*. Milan: Feltrinelli, 1991.

Purves, D., Morgenstern, Y., and Wojtach, W.T. 'Perception and Reality: Why a Wholly Empirical Paradigm is Needed to Understand Vision', *Frontiers in Systems Neuroscience*, Vol. 9, No. November 2015: 1-10.

Purves, D., Wojtach, W.T., and Lotto, B. 'Understanding vision in wholly empirical terms', Proceedings of the National Academy of Sciences, #108 (Supplement 3), September 2011: 15588-95.

Ratto, M. 'Ethics of seamless infrastructures: Resources and future directions', *International Review of Information Ethics*, Vol. 8, 2007: 20-27.

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. 'Inferring ground truth from subjective labelling of venus images', *Advances in Neural Information Processing Systems*, No. 7, 1995: 1085-92.

Somaini, A. and Pinotti, A. *Cultura Visuale: Immagini sguardi media dispositivi*. Milan: Einaudi, 2016.

Steyerl, H. 'In Defense of the Poor Image', *Eflux Journal*, Vol. 10, No. 10, 2009: 1-9.

Vaccari, F. *Fotografia e inconscio tecnologico*. Milan: Einaudi, 2011.

Virilio, P. *Speed and politics*. Cambridge: The MIT Press, 2006.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, KW. 'Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints', Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2979-89.

Notes

[1] Kovashka & Russakovsky & Fei-Fei & Grauman 2014, p. 1.

[2] 'Gold Standard Data is used in many industries to assess the validity of test results. Creating HITs that you already know the answers (aka Gold Standard HITs) is any easy way to assess the accuracy of a Worker. You could create HITs that you know the answers to in Mechanical Turk and compare the results to your known answers to identify Workers that accurately complete your HITs. […] You could use this to measure Worker accuracy over time and determine if a Worker's accuracy is consistent or if the Worker's Qualification should be revoked.' http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf (accessed on 14 February 2018)

[3] Malevé 2016. See https://unthinking.photography/themes/machine-vision/the-cat-sits-on-the-bed-pedagogies-of-vision-in-human-and-machine-learning (accessed on 15 February 2018)

[4] Ibid.

[5] Fei Fei Li, How we teach computers to understand pictures, https://www.youtube.com/watch?time_continue=18&v=40riCqvRoMs (accessed on 5 February 2018)

[6] 'ImageNet is an ongoing research effort to provide researchers around the world an easily accessible image database. […] ImageNet is an image dataset organized according to the WordNet hierarchy. Each meaningful concept in WordNet, possibly described by multiple words or word phrases, is called a "synonym set" or "synset". There are more than 100,000 synsets in WordNet, majority of them are nouns (80,000+). In ImageNet, we aim to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. In its completion, we hope ImageNet will offer tens of millions of cleanly sorted images for most of the concepts in the WordNet hierarchy.' http://image-net.org/about-overview (accessed on 10 February 2018)

[7] Lotto 2017.

[8] Xiao & Hays & Ehinger & Oliva &Torralba 2016.

[9] Ernst 2003 (accessed on 26 February 2018).

[10] Vaccari 2011.

[11] Barriuso & Torralba 2012, p. 6.

[12] Kovashka & Russakovsky & Fei-Fei & Grauman 2014, p. 15.

[13] Barriuso in her paper writes about her strategy to create a personal vocabulary, so as to avoid spelling errors (English is not her mother tongue) and to remain consistent in the choice of terms to be assigned to different objects.

[14] Barriuso & Torralba 2012, p. 11.

[15] Baxandall 1988, p. 29.

[16] Ibid., p. 150.

[17] Ibid., p. 151.

[18] Ibid., p. 150.

[19] Barriuso & Torralba 2012, p. 11.

[20] Ibid.

[21] Kovashka & Russakovsky & Fei-Fei & Grauman 2014, p. 10.

[22] Hansson & Muller & Aitamurto & Irani & Mazarakis & Gupta & Ludwig 2016, p. 3605.

[23] Bederson & Quinn 2011, p. 99.

[24] Hardt & Negri 2000, p. 292.

[25] Chihuahua or muffin? or Labradoodle or fried chicken? is a popular internet meme which shows how image recognition programs can easily make mistakes. See https://www.instagram.com/karenzack/ . See also https://medium.freecodecamp.org/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d (accessed on 25 February 2018)

[26] High-level visual concepts, in the field of computer vision, indicates the ability of a machine or algorithm to extract information from an image that goes beyond the representation itself (e.g. feelings or relations between objects).

[27] Vaccari 2011.

[28] McLuhan 2015.

[29] High-level confidence and low-level confidence are terms used in the field of computer vision to distinguish the ability of a machine or algorithm to predict and recognise an image, or an object present in it, compared to a human being.

[30] Paglen 2017.

[31] Gallagher & Zahavi 2008, p. 160.

[32] Hoelzl & Re[]mi 2017, p. 73.

[33] Malevé 2016. See https://unthinking.photography/themes/machine-vision/the-cat-sits-on-the-bed-pedagogies-of-vision-in-human-and-machine-learning (accessed on 19 February 2018)

[34] Paglen 2017.