Perception as Infrastructure: the Case of Computer Vision

Nicolas Malevé

Nicolas Malevé is a visual artist, computer programmer and data activist who lives and works between Brussels and London. He has recently completed a Phd thesis on the algorithms of vision at the London South Bank University. He is a member of Constant and the Scandinavian Institute for Computational Vandalism. In the Active Archives project, with Michael Murtaugh, he is experimenting with techniques to engage with large collections of visual materials and explore different ways to navigate and question them.

https://constantvzw.org/site/_nicolas-maleve_.html

T.C. During the presentation, you explained that the current move to the cloud implies is a shift regarding Marx's "base determines superstructure" axiom. Could you explain briefly this idea? What are the current dynamics and consequences of that shift in your opinion?

N.M. This is an old debate in the Marxist tradition. To put it simply, initially Marxists were seeing the infrastructure (economy, industry, means of production) as the determining factor in the evolution of society. The superstructure (state, police, education, media) was considered as determined by the infrastructure. With the increasing importance of cognitive labor in the service industry, the attention economy and lately the shadow labor of AI, the relation between infra and super structures becomes more complicated to assess. What traditionally belonged to the superstructure is absorbed in the infrastructure. Through the training of vision algorithms, our ways of seeing become infrastructural.

T.C. As Tyler Reigeluth reminded us during his presentation, the West has cultivated a fetish for automation since the Enlightenment. Machines should work perfectly and humans are often defined as "what is not automated yet". The cloud, Internet of Things and "smart" technologies make the promise to pursue this ideal. However, during the presentation you drew quite a different picture, introducing the notion of *human infrastructure*. What is it?

N.M. The anthropologist and film maker Steffen Kohn uses the term human infrastructure to analyze the situations were humans perform the functions usually attributed to machines. For instance, with the artist Nestor Siré, they map the distribution of digital cultural goods in Cuba. Instead of exchanging digital media via the Internet, delivery men called *paqueteros* bring hard drives with the files to the homes of Cuban citizens. In the presentation, I used this term to talk about the workers of Amazon Mechanical Turk (AMT) who perform the manual labor behind AI. AMT provides an infrastructure without which deep learning algorithms could not work. This infrastructure responds to API calls and delivers outputs that can be integrated in software routines. But those who answer the API calls are poorly paid workers, not software.

T.C. Could you explain what computer vision is and where it comes from? How is it currently developed as an infrastructure? How it is related the cloud? According to you, what is the aim of people promoting this technology?

N.M. Computer vision is the discipline of computer science that attempts to emulate human visual abilities through software. It is present in most of our devices and tools from cell phone, camera, car, search engine, social media. They are also present in various sectors of the industry where they are used to optimize the assembly line. And they are used to monitor citizens behavior and perform social sorting. Andrew Ng says that AI is like electricity, it should offer *cognition on tap*. Computer vision is *vision on tap* for software. Current machine vision algorithms make heavy use of deep learning techniques that require a lot of computational power. Consequently large machine vision models are trained on the cloud. As machine vision is rather polymorphous, it is promoted and used by people with different aims. A photo sharing platform will typically use it to curate its users photographs. A state may promote it to offer more security to its citizens and increase the surveillance of the population. Researchers in oncology may try to convince specialists to use it to facilitate diagnosis. And with the new breakthroughs of Dall-e and Imagen, tech pundits see it disrupting a varied set of industries ranging from stock photography, comic book to video game.

T.C. How are datasets created? How determining is the choice of datasets in Machine Learning generally and computer vision specifically? Are there different approaches?

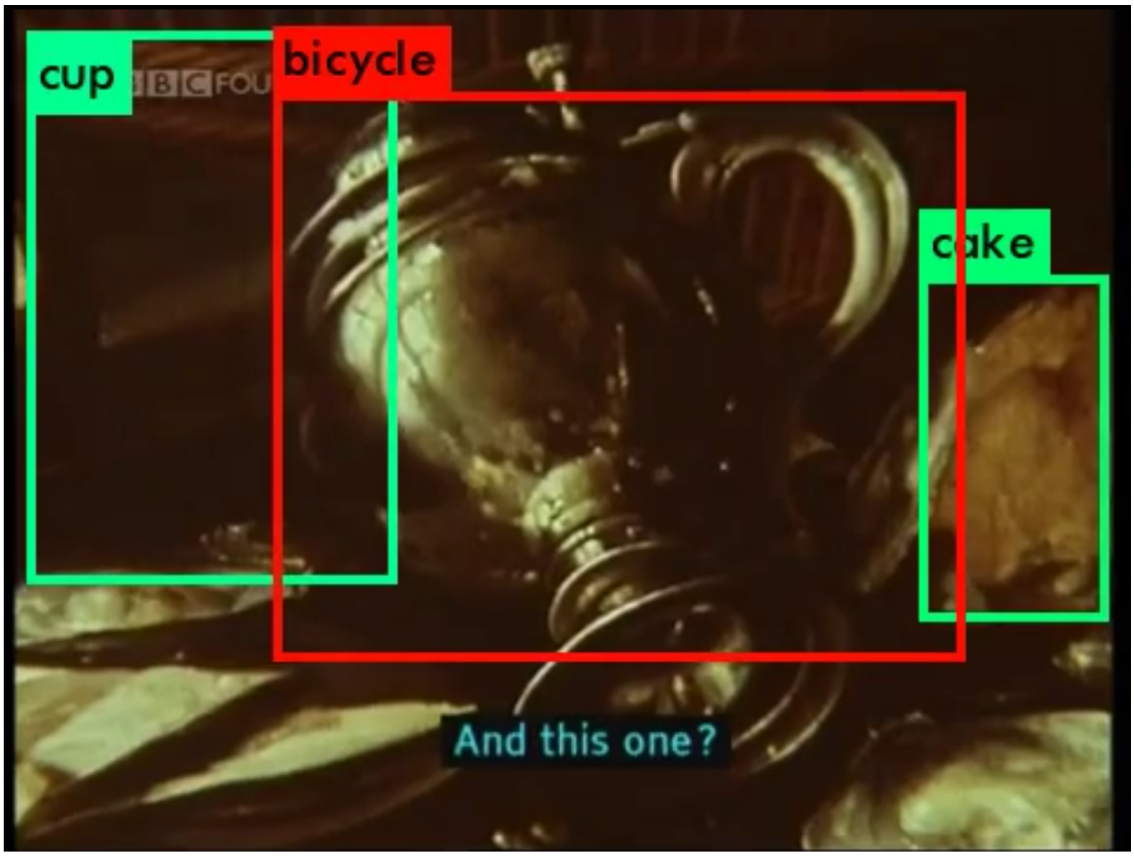
N.M. The dataset offers an interesting entry point to the problem of machine learning and computer vision more specifically. The dataset is a key device in current machine learning. It provides the model of the objects algorithms need to interpret. Let's take the example of a cat. A cat is not reducible to a simple diagram that contains a circle, two dots, a few lines for the whiskers and a chubby body. If you look at cat photos online, the animal takes many shapes, is seen from odd angles, and positions. This makes it impossible to retrofit all the variations to a simple geometrical model. To find enough variations and regularities, one needs to curate large collections of photographs. Current algorithms increasingly rely on huge collections of images, datasets, to learn the regularities of the visual world. Many problems of bias can be traced to the choices made in the creation of these datasets. As algorithms

learn from them, they also encode their worldviews. And as computer scientists acquire their images *en masse* from photo sharing platforms, algorithms learn to see like Flickr photographers. Datasets are inseparable from the question of classification. Dataset = images + taxonomies. For examples the authors of the Common Objects in Context (COCO) claim "Our dataset contains photos of 91 objects types that would be easily recognizable by a 4 year old." Artist Philipp Schmitt has commented: "They voted — among themselves — on the best categories, and even consulted several children in ages from four to eight. What a responsibility for a four-year-old! Soccer ball didn't make the cut, baseball bat did. What is common, of course, depends on who is looking." WordNet, used in the famous ImageNet dataset, counts 117.000 categories with the ambition to order the whole universe from particles to airplanes, planets, spoons, soldiers and microbes. WordNet sadly reproduces numerous cultural biases as it blindly integrates racist tropes, sexist slurs or homophobic clichés.

T.C. How is human labor organized in order to produce datasets? What are the effects of this organization on workers and their perception? How does human cognition fit into the computer vision infrastructure?

N.M. Constructing datasets requires a considerable amount of manual labor to produce the work of annotation: description, tagging, classification, drawing contours of objects The labor of tagging and filtering the images is done through crowdsourcing platforms like Amazon Mechanical Turk. Workers are paid a few cents per task. Workers need to make their choice at an extraordinary pace to earn a minimum amount of money. The ever increasing demand for large datasets from the AI industry forces the annotators to perform their work at speed. To deliver on time in the case of ImageNet, workers were supposed to classify images in less than 500 milliseconds. Leaving the question open as to how to reconcile speed and accuracy. Social violence and economic exploitation are the rule. And we should not forget: the humans who annotate the images also have to learn what to look for in an image. They are also trained on the platforms where they do their work to associate words and images and to see at speed. To sum up, algorithms learn by image and their ways of seeing can be traced back to their datasets with all the tensions it entails. And the process of training machines also involves training the humans who annotate the images.

A side-by-side view of the same video from John Berger's TV series *Ways of Seeing*, using the same AI but trained with different datasets.



A side-by-side view of the same video from John Berger's TV series *Ways of Seeing*, using the same AI but trained with different datasets.

Further readings

Azar, Maryam, Cox, Geoff, and Impett, Laura 2021. Introduction: Ways of Machine Seeing. *AI & Society*, *36*, 1093–1104.

Fei-Fei, Li, Iyer, Asha, Koch, Christof et al. 2007. What do we perceive in a glance of a real-world scene?. *Journal of Vision*, *7*(1), 10.

Irani, Lilly 2015. The Cultural Work of Microwork. *New Media & Society*, *17*(5), 720–739.

Pipkins, Everett 2020. On Lacework: Watching an Entire Machine-Learning Dataset. *Unthinking Photography*. Advanced online publication. Retrieved from https://unthinking.photography/articles/on-lacework

Rubinstein, Daniel and Sluis, Katrina 2013. Notes on the margins of metadata: Concerning the undecidability of the digital image. *Photographies*, *6*(1), 151–158.