GRESHAM COLLEGE

12 FEBRUARY 2019

COMPUTER VISION: MACHINES THAT SEE

PROFESSOR RICHARD HARVEY FBCS

As we look around, we see the world in high-resolution, full-colour detail without any apparent effort at all. Yet our attempts to replicate this in computers have been intriguingly partial and hard-won. Partly the problem is that human vision is an illusion – we know perfectly well that we see in high resolution only in a narrow beam (foveated vision) and that our colour perception is also flawed – recall the dress that broke the internet [1]. Furthermore, newborn babies have such poor vision that we would classify them as par-blind. Hence the benchmark, human vision, is poorly understood which hardly helps to put computer vision on firm theoretical foundation. Despite this, and despite fights for funding and recognition, computer vision is in robust health and real systems that solve real problems are not hard to find. In this lecture I want to give some examples of successful vision systems and to indicate where future progress is possible. Of course it is a highly abbreviated view and computer vision is a highly mathematical subject so I shall tread rather warily over the thin ice of many theoretical issues, but I hope by the end of this talk you will have an appreciation for the main issues,

Computer vision has its early origins in close connection with early work on animal vision systems. For example, David Marr, in his book Vision [2], which is a thoroughly good read and not too technical, argues for a structure to vision systems which he calls the *primal sketch*. The primal sketch is like a rough pencil sketch of a scene with all the detail at the lowest level. The idea is that those details are put together in a tree of information – as we move up the tree, we move from the detection of basic features such as intensities to lines, edges, corners and so and then to objects.

The hierarchy corresponds roughly to the layers in our visual cortex (V1 to V6). The analogy is far from exact because the visual cortex changes its function over time but analogies between human vision and computer vision persist to this day. Indeed, one method of classification, artificial neural networks, are named after one aspect of the human brain.

Computer vision operates on what are known as *images*: matrices of numbers that represent the picture seen by a camera. Images are almost always rectangular arrays of *pixels* (picture elements). Each pixel consists of three numbers that are meant to represent the intensity of the red, green and blue light falling on the camera. The idea is that any display device also has three R,G,B illuminators and hence the colour is replicated. However, all is not as simple as it seems, as this *tristimulus assumption* is only an approximation. Not all colours can be approximated by RGB and, as The Dress That Broke the Internet showed [1], there are fundamental difficulties in identifying colour reliably. The situation is made even more complex by the intricate processing that takes place in a digital camera. It is not uncommon, in small cameras, to have pixels deliberately removed to save space. Hence colours have to be interpolated – typically the red and blue channels might have lower resolution than the green. Or there will have been a whole host of nonlinear operations to make the image "look good". Such operations are usually proprietary so, very few people know the processing that takes place in, for example, a modern mobile phone. For computer vision specialists these can cause additional complexity. In short, compared to audio, the signal processing chain is complex and often secret. A further knottier problem is that unlike sound which can be well approximated by linear models (one signal adds to anther and so on), images generally use a replacement model which further encourages non-linear processing which in itself is more theoretically challenging.

When it comes to video, the situation is even more off-putting. Video is just a sequence of images, or frames, taken at high speed (often 30 frames per second). However, even 30 frames per second is slow compared to everyday motions so it is frequent to find very noticeable motion errors in video. These errors may arise because of a failure to satisfy the Nyquist sampling criterion (see previous lectures) and are sometimes known as the "waggon wheel effect" as in old westerns it was frequent to see the wheels on the wagon turning backwards or they may be simple motion blurring caused by the moving object moving during the capture of a single frame. Nowadays it is possible to film at higher frame rates, but high frame rates imply more illumination (as the frame rate increases, the shutter speed increases, which leads to less light hitting the image sensor). Worse, higher frame rates imply even more staggering amounts of data. A 4K digital movie, without

compression, would generate 6GBytes of data per second. Even with very extensive compression, 25MBits per second is normal, which is very demanding for anything but supercomputers.

These torrents of data create a major conundrum. On the one hand, humans seem to be able summarise activity and objects in video using a modest amount of words (small amounts of information) and on the other, the raw data is enormous (large amounts of information). There is an information pyramid with very compact inference at the top of the pyramid and huge numbers of pixels at the bottom. This paradox was realised very early in the history of human vision and is illustrated nicely with Attneave's cat shown in Figure 1.

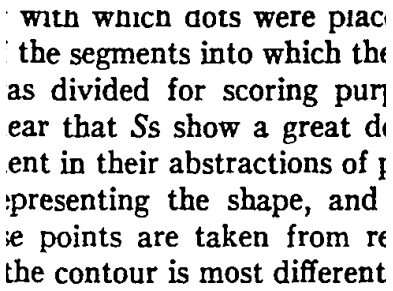

Figure 1: Attneave's cat taken from [3]

Most people perceive a sleeping cat. Figure 1 requires only 64 line segments, which could be encoded in only a few hundred bits. Yet the original image would comprise millions of bits to achieve, possibly, a worse rendition. There is clearly a need to simplify images in a way that allows inference to take place. There is a plethora of such simplifications. Sometimes they are called "coarse-to-fine methods" since they allow the coarse, inference, to be tracked back to the fine-scale of the image but the basic problem is scale – choosing the right scale to operate is always challenging. One formal system that has been theoretically popular is called *scale-space*. Scale-space vision systems start from the observation that objects in images are often darker than or brighter than their immediate background[1]. If one accepts that premise, then any simplification of images has to not introduce any more regional extrema (maxima or minima). This non-enhancement-of-regional-extrema property is very helpful since very few systems satisfy it. For linear, averaging, filters, only one filter meets the criterion namely the Gaussian filter[2] [4]. When this was discovered, this caused great excitement because the human visual cortex also appears to use smoothing filters which are roughly Gaussian. However, the analogy is far from perfect and scale-space mania seems to have died down a bit. Furthermore, if we are allowed to use non-linear systems, then there are several alternatives of which my favourite is a morphological processor called the *sieve* [5][3]. The scale-space purists would eschew any method that does not meet the scale-space axioms but, in reality, there are many systems which attempt to simplify images and the Gaussian blur is quite a popular choice. There are also multiscale systems that do not satisfy the scale-space axioms but do reduce the information with increasing scale. Of these, *wavelets*, are probably the most interesting, especially complex wavelets [6], as they do halve the amount of information for each doubling of scale. In short, unless one can be confident that one is looking for an object at a fixed scale, almost all modern vision systems will have a multiscale front-end.

There are of course, some systems which are spectacularly successful despite operating at one scale. One system which I greatly admire is Iris Scan. The holy grail of biometric identity is a pattern that is unique to every person, does not change as they age, is easy to acquire and is not only random but equally distributed between persons[4]. These desiderata are not east to meet and there have been multiple studies of fundamentally difficult biometrics such as faces, voice, gait, ears and so on. John Daugman realised that the pattern of flecks in the human iris was fundamentally random, and stable with age [7]. The computer vision system to extract the eye pattern is relatively simple because the iris has a nice simple shape and the distance from the camera is controlled. The system is in use by the Indian Government in a project to enrol every Indian citizen, who wishes, so that social security and other benefits can be distributed without fraud. Every volunteer is given an *Aadhaar* identity which they can use to withdraw cash (there are cash machines that use iris codes instead of cards). The size of the program is staggering – there are 36,000 portable roaming enrolment stations and each new enrolment has to be compared against all the others (to prevent duplicate identities). Making a comparison of a single code against 1.2Bn other codes is an impressive feat. The system is currently reported as having achieved 22Bn authentications with, I assume, zero errors [8].

One of the recent triumphs of computer vision is object tracking. The breakthrough paper was by a Canadian scientist, David Lowe [9]. Unfortunately, Lowe's work, which he called the Scale Invariant Feature Transform or (SIFT) was patented which had the usual stifling effect on its adoption, but it still makes interesting reading. In the first step he progressively

[1] Indeed, were they not brighter or darker than the background they would not be visible in greyscale images.
[2] Actually, in discrete images it is an approximation to the Gaussian filter [4].
[3] However, in the interests of fairness, I should point out that I was partly responsible for the work on sieves.
[4] Twins, for example, should not be closer in pattern than other people.

smooths, or blurs, the image with Gaussian filters. He then argues that if there are significant differences between success stages of blurring then this means that something interesting happened at that scale – these points are called *interest points*. At each interest point he generates a list of numbers that describe that point (local gradients and so on). These SIFT points can then be matched from image to image (for tracking or for stereo) or against a database (for object recognition). Partly driven by the excitement of SIFT and partly by the fear of the patent, a number of alternatives were quickly produced. There was also a rediscovery of simpler features such as corners. We gave up on corners many years ago because they were not stable under illuminant variation but the SIFT method encouraged work in more robust tracking and that allowed simple features such as corners back into contention.

SIFT also helped with a long-standing problem of how to infer depth from two-dimensional images. Again, this is something that humans appear to do effortlessly aided greatly by having two eyes a fixed distance apart and on the horizontal whereas computers are often required to infer depth from cameras is less congenial configurations. They key to inferring depth is to find points in the left-hand image that match points in the right-hand image and vice versa. Once these matches have been found, we can work-out how far they are displaced (a quantity known as *disparity*) which in turn maps to depth. But finding those matching points is tricky. In *dense* stereo every pixel on the left is matched to every pixel on the right and vice versa. This uses a lot of processing as each pixel has a potential match along a line so there is much searching for matches. An alternative is *sparse* stereo where feature points are used to find points that are likely to match. The sparse approach is usually preferred in the related problem of tracking – in which the camera and/or object may be also be moving.

So far, the systems I have been discussing have not used machine learning. Although the terms machine learning, or artificial intelligence or pattern recognition, apply to problems that are more general than vision (in the previous lecture we saw how machine learning could be used to recognise speech), it is fair to say that progress in machine learning has been intimately linked with computer vision. Machine learning splits into two models: supervised and unsupervised learning. Unsupervised learning is the business of finding things that seem to be similar without knowing what they are. If I showed you a photograph of 100 chicks of which 99 were female, you could spot the male since male chicks are usually a different colour. You do not need to know that male chicks are darker than females nor do you need to know that colour relates to sex. That is *unsupervised* learning. The majority of vision systems that use machine learning, use *supervised* learning though. In supervised learning we generate some data, called training data, in which we *label* the data with *classes*. Of course, the labels get used in different ways. In the slides I look at a simple system we designed to spot pornography [10] in which we labelled each image with a class label that designated the image as porn, nude, people, misc., graphics. In face detection, images are labelled with boxes showing where the faces are. The trick is then to extract, from the image some numbers, or *features*, which might represent the classes. For example, in pornography detection we discovered that we could build a skin detector quite reliably so the area of skin in an image was one of our features. There are then a host of possible machine learning algorithms which learn the relationship between the features and the classes. To give a simple illustration of this, let's imagine we going to determine the sex of someone by measuring their height and shoe size. Table 1 shows some example data.

Table 1: Example data for explaining machine learning

Height (cm)	Shoe size (UK)	Class
182	12	M
170	6	F
165	4	F
180	10	M

In some classifiers there would be an extensive effort to normalise and summarise these data. Clearly with only four training points that is futile but with 400M points then it would be essential. In testing we take some data with an unknown class, say the feature vector [181, 11] and we ask the classifier to guess the class. In this case I think you would guess M. You might have done this by using your knowledge of what constitutes a male (tall and big feet), your *prior* knowledge, or you might have compared [181,11] with each line of the training matrix, found the closest vector and then guessed the class of the closest. If you followed that latter algorithm then congratulations, you have discovered the nearest-neighbour classification algorithm! If you went through the list looking for the closest vector, the next closest, the next closest and so on and then you took a vote among the k-nearest classes, you have been using the k-nearest neighbour classifier or (kNN). The kNN often works surprisingly well but it is frighteningly expensive – for each classification you have to compute a distance between your test vector and every single training vector. Hence complexity is a big issue for classification. Another problem, which is evident in Table 1 is that height and shoe-size seem to be highly correlated, furthermore height is measured in hundreds, while shoe-size is measured in small numbers. Here the training data need some normalisation and preparation and we would expect a good machine learning system to handle that automatically. And finally we want a machine learning system that not only gives us an answer, but also tells how confident it is about the answer. In the kNN, we not only want the majority vote, but also the share of the vote. Increasingly over the years, there has been an acknowledgement that the

only framework for this activity that makes sense is a probabilistic one. Looking back, it seems surprising that anyone should argue anything else, but there were certainly some dark decades when there were arguments for all sorts of unprincipled algebras[5].

A recent dramatic development in learning has been the so-called deep neural networks. At the start of this lecture I pointed out that the history of computer vision was intertwined with the study of animal vision. One form of classifier that took its inspiration from the human brain is known as artificial neural networks or ANNs. ANNs were extremely popular in the 1980s and 90s but as soon as they became complex enough to solve real problems, they seemed to become impossible to train. So progress in AI turned to alternative forms of learning such as the Support Vector Machine (a fancy sort of projection classifier). However, when those training problems with ANNs got solved there was an explosion – any explosion of performance (suddenly speech recognition made dramatic improvements compared to rather incremental performance of a few years previously) and an explosion of interest. Nowadays every machine learning system had better be using deep learning or that paper will not get accepted or that start-up will not get funded. But is deep learning all that it is cracked up to be? That is the topic of the next lecture.

1. "The dress that BROKE the internet: Millions debate on gold & white or blue & black frock", Sarah Barns, Sunday Express, Published 09:05 Feb 27 2015, Updated 01:51 Sat Mar 28 2015. Online at https://www.express.co.uk/life-style/style/560786/That-dress-white-gold-blue-black-broke-the-internet-Kim-Kardashian
2. David Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francisco: W. H. Freeman and Company. ISBN 0-7167-1284-9. 1982. (In 2010, MIT press re-published the book with a foreword from Shimon Ullmann and an afterword from Tomaso Poggio under ISBN 9780262514620.)
3. "Some informational aspects of visual perception," Fred Attneave, Psychological Review, Vol 61, No 3, 1954, pp183—193.
4. Tony Lindeberg, "Scale-space theory in Computer Vision", Kluwer Academic, Dordrecht, Netherlands, 1994
5. J. Andrew Bangham, Richard Harvey, and Paul D. Ling. "Morphological scale- space preserving transforms in many dimensions". Journal of Electronic Imaging, 5(3):283–299, July 1996.
6. N G Kingsbury: "Complex wavelets for shift invariant analysis and filtering of signals", Journal of Applied and Computational Harmonic Analysis, vol 10, no 3, May 2001, pp. 234-253.
7. Daugman J (2007) "New methods in iris recognition." *IEEE Trans. Systems, Man, Cybernetics B* **37**(5), pp 1167-1175.
8. https://uidai.gov.in/aadhaar_dashboard/
9. David G. Lowe. "Distinctive image features from scale-invariant keypoints". *International Journal of Computer Vision*, 60(2):91–110, 2004.
10. Bosson A., Cawley G.C., Chan Y., Harvey R. (2002) Non-retrieval: Blocking Pornographic Images. In: Lew M.S., Sebe N., Eakins J.P. (eds) Image and Video Retrieval. CIVR 2002. Lecture Notes in Computer Science, vol 2383. Springer, Berlin, Heidelberg

[5] One of the leading offenders was a system called "fuzzy logic". Fuzzy logic was based on little more than a blind faith that by allowing some imprecision in decisions, that the overall decision would be improved.