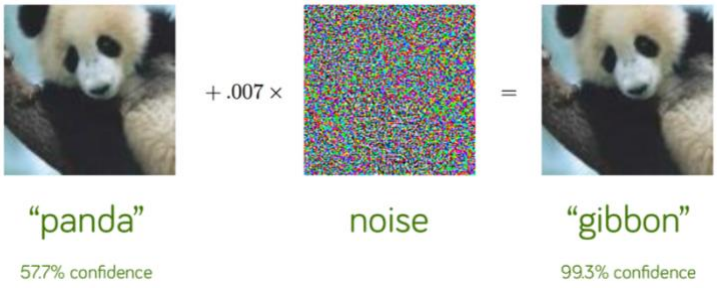
Seeing in 3D: Humans + AI

Paul Linton

In my talk I'm going to give my own take on the subject of 3D vision. But for this pre-paper I want to situate my talk (and work) within a broader movement within the 3D vision literature that is happening in parallel in three fields: Computer Vision, Animal Navigation, and Human Vision. What I intend to do in this pre-paper is give you an overview of the kinds of debates we were having at the recent Royal Society meeting that I organised, "New Approaches to 3D Vision". Over 800 people participated in the meeting, with speakers from DeepMind, Google Robotics, Microsoft Research, and Meta (Facebook) Reality Labs, as well as academics from both basic and applied research. The meeting has a website [link] with recordings, and I'm also going to provide a list of the talks with links at the end.

So the first question (or dare I say 'criterion'!) of the project is "why now"? First, we've seen as massive success in deep learning (artificial intelligence) approaches to identification: identifying objects in 2D photos. So now attention is turning now to the more challenging question of 3D scene understanding. Indeed, this is why I have teamed up with Prof. Niko Kriegeskorte who will be my guest at the seminar. Niko had a revolutionary paper in 2014 [link] showing the similarities between deep learning and human image identification. The purpose of our work together is to now explore if we can do something similar for 3D. Second, even in the 2D case, there are some big problems, e.g. adversarial images. So adding a few specs to an image of a panda makes a neural network think it's a gibbon. So the US government has



funded a number of initiatives to build 'smarter' AI grounded in 3D vision to solve it. Third, people are really excited about the early achievements in 3D vision. For instance, AlphaFold by DeepMind, uses 3D vision to largely solve the protein folding problem – understanding the process by which a polypeptide chain folds in 3D to become a biologically active protein – which is massively important for biology and medicine. Fourth, 3D vision remains the big challenge for robots and autonomous cars, something I will discuss in my talk.

Ok, so why bring Computer Vision, Animal Navigation, and Human Vision together? Well, and this was the argument of my Royal Society meeting, actually all three disciplines are grappling with the same question, and should be helping one another! So what is the question? Well, we typically think of 3D vision as the brain (or computer) creating its own accurate 3D model of the world. So that's an easy idea. But very hard to do efficiently in practice. But the alternative is there's a harder idea to grasp, which is easier to do in practice. And I think a lot of people agree that's the case. We just don't know what that harder idea is! And so that was the purpose of the meeting, to see how different disciplines approach the problem.

So what have been some of the proposals over the years?

1. Pixels-to-Action: So one group say 'look, you're getting it all wrong even thinking about 3D vision'. These tend to be roboticists. And they say, "look, what do we need 3D vision for? We need it to control a robot navigating the environment. So we have an input (the pixels on the camera surface) and an output (the action of the robot in very simple terms – what electrical current to apply to which motors) and that's it. So all we need is a way of linking those two things (*if* pixels a,b,c *then* motor torques x,y,z), and there is no need for this intermediate step you call to as 3D vision". And this is a massively influential position.

It started with Rodney Brooks at MIT in the 1980s and was inspired by insects: "Look at an insect, it can fly around and navigate with just a hundred thousand neurons. It can't be doing this very complex symbolic mathematical computations. There must be something different

going on." Although Rodney Brooks is also the inventor of the Roomba, which does build an internal map of the world – so questionable whether its leading advocate still agrees!

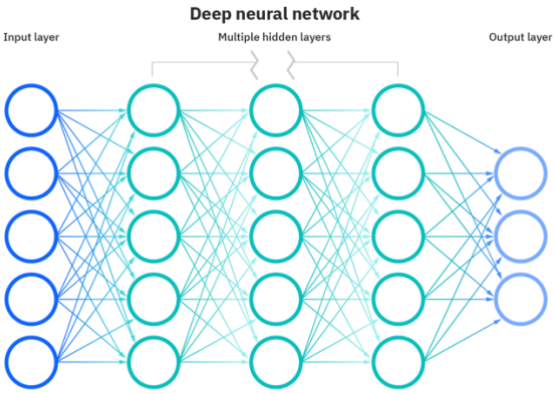
But someone who does certain agree is Google! So Sergey Levine at Google Robotics and Berkeley is working on using reinforcement learning that is based on this 'pixel to action' principle. And at my meeting he described Google's latest efforts in this area.

What is 'reinforcement learning'? Well go back to the example '*if* pixels a,b,c *then* motor torques x,y,z'. How do you get that '*then* motor torques x,y,z' into the computer? In the early days before reinforcement learning, you had to specify it by hand. But think how many different combinations of pixels on the screen you'd have to hand-code! So instead, the computer learns it by seeing pixels 'a,b,c' and then taking an action and seeing if that produces a reward, or not (in classical navigation, 'target reached'). And you start with a random actions, but slowly the robot builds up an idea of which moves are right (lead to target) and which are wrong (don't).

But this requires a lot of training. Why would you do this to yourself (and your robots!)? The answer, according to Levine, is that we don't actually want what 3D vision gives us, which is surfaces and distances. For instance, tall grass looks like a barrier to a robot, when really it is traversable. Whilst mud looks like it isn't a barrier, when actually it is, since it's not traversable. So rather than learning vision, and then traversability, Google's robots just learn traversability.

But this poses a real challenge to computer vision. As (Zhou et al., 2019) note in 'Does computer vision matter for action?': "These models bypass explicit computer vision entirely. They do not incorporate modules that perform recognition, depth estimation, optical flow, or other explicit vision tasks. The underlying assumption is that perceptual capabilities will arise in the model as needed, as a result of training for specific motor tasks. This is a compelling hypothesis that, if taken at face value, appears to obsolete much computer vision research."

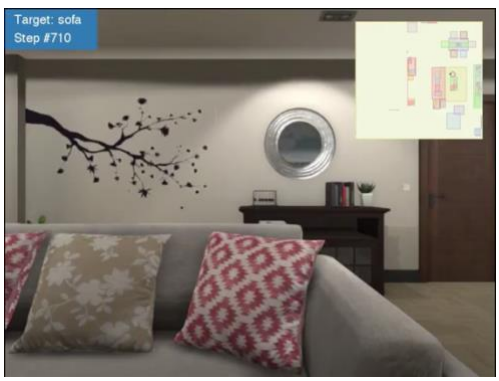
2. Deep Reinforcement Learning: Now to achieve the kind of performance they want, Google use deep neural networks to learn reinforcement learning. What is a neural network?



Here's a very simple neural network. It's deep because it has multiple (more than 1) hidden layers. Each circle is a number – think of it as the output of the arrows pointing to it, and the input to the arrows pointing away from it. Each arrow is a computation – multiplying the input to that arrow by a certain weight. So in our Google robot example, the input would be the pixels on the camera (a,b,c,d,e) and the output would be the amount of current to apply to e.g. (in this case) 3 motors. So in our example, the robot tries some set of weights, and sees what the rewards are. Then it slightly changes the weights, and sees if the rewards are better or worse. And gradually by trial and error it converges on the right weights.

A really good video of this in action is from another group (Yuke Zhu, now at Austin). Watch the following video from 1:38-2:08: https://youtu.be/SmBxMDiOrvs?t=98 (warning, lightly flashing images, so viewer caution). This is the random walk trying to find a sofa. Now, let's look at what happens after training (2:19-2:24): https://youtu.be/SmBxMDiOrvs?t=139

So 710 steps (random walk) vs 22 steps (after training):



But there's a big theoretical discussion here. Now we no longer have a simple mode (pixels → action), we have all these weights in the neural network that it has learnt over millions of frames. And surely buried deep within these weights has to be a model of the 3D world?

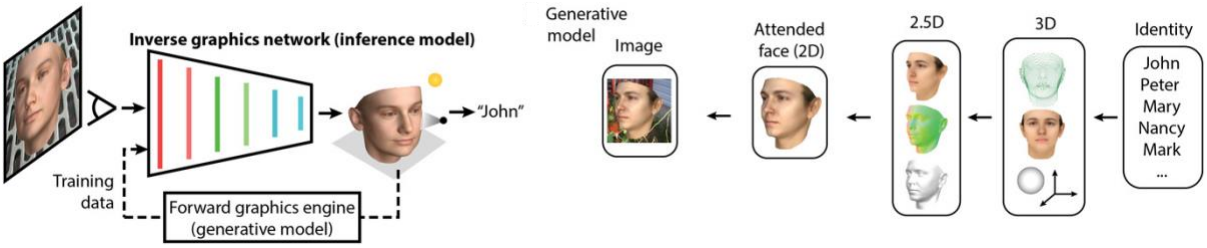
Indeed, Yuke Zhu and colleagues suggest as much: "Our method is considered *map-less*. However, it possesses implicit knowledge of the environment." But what does that mean? It's hard to say, because the model isn't interpretable by humans – it's a bunch of numbers (weights) – what do they mean? So what we have to do is experiments that sort of tease out what the model does. For instance, do small changes in the model correspond to small changes in the world? The Glennerster & Torr labs looked at this with regards to Yuke Zhu's navigation network and found "only a weak correlation between distance in the embedding space and physical distance between observable locations" (Muryy et al., 2020). So something more complicated must be going on under the hood. But it's hard to decode what!

3. Deep Learning: Another approach is simply to train a neural network with "ground truth" images – example 2D images that you know the true 3D value for – with the hope that the network will generalise effectively beyond this training set of 2D images to other 2D images.

But again, we have to be very careful about what is being learned. As (Fleming & Storrs, 2019) explain: "rather than learning the mappings between image quantities ('cues') and physical quantities, we learn to represent the dimensions of variations within and among natural images, which in turn arise from the systematic effects that distal properties have on the image." So it's not as if neural networks are reasoning about 3D space by taking the 2D image we feed it, and thinking 'what kind of 3D scene could produce this'. Instead, all they are learning is a statistical relationship between certain changes in the image, and associating them – purely at the statistical level – with certain 3D properties of the world.

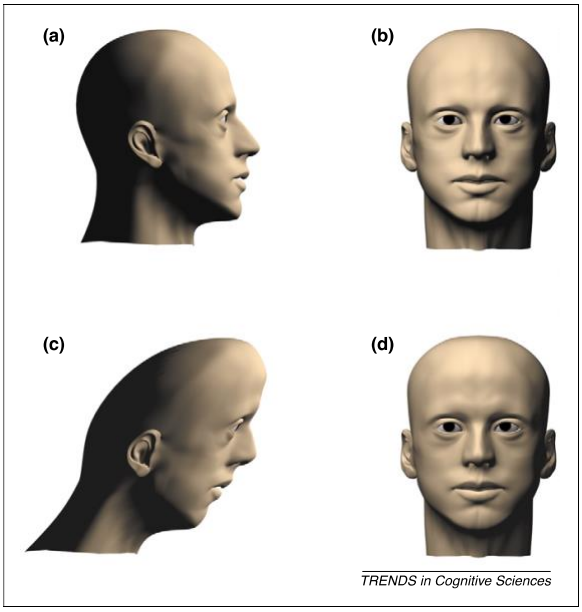
[Skippable aside: But is this all that human learning is? As (Fleming & Storrs, 2019) note, doing this "we may end up with internal representations that are well suited for describing the distal scene factors that have created those images."]

On the other hand, you have other groups saying 'no, we have to go further, and actually teach machines to explicitly reason about 3D space from 2D images'. So an example is Josh Tenanbaum's group at MIT (who collaborates with me and Niko), who has an inverse graphics model of human vision that his group apply to computer vision (Yildirim et al., 2020).



So how do we recognise someone as "John"? On this model, rather than teaching you to identify John by going from a 2D image (certain arrangements of pixels) → an identity ("John"), they build a Generative model (on the right). Basically, if I ask you, "is this John in this photo?", you build in your head a 3D model of the person and then simulate what they would look like in a particular scene. And that's how they create images to train their Inverse Graphics model (on the left). This Inverse graphics model takes a 2D image and then decodes it into its 3D components (3D geometry, material, and lighting), and then it uses this 3D model to identify the person. So it's a very different way of identification (2D → 3D → identity) than simple 2D → identity.

4. Affine Geometry: Another approach that gained a lot of traction in the late 1990s and early 2000s in computer vision, and which is gaining attention in human 3D vision (especially the work of my co-organiser of the Royal Society meeting, Fulvio Domini at Brown) is the 'affine geometry' approach to 3D vision. What does that mean. Well, 'affine' geometry is the normal geometry of the scene subject to a homogenous distortion. Let me give you an example from (Belhumeur et al., 1997) on 'shape from shading'. The bottom head in the image below is an 'affine' (homogenously distorted) version of the top head. But, and this is the key point, both heads produce virtually identical 2D images, and therefore virtually identical shading cues.



(Belhumeur et al., 1997) term this 'the bas relief ambiguity'. There just isn't the optical information in the image to disambiguate these two 3D heads. One option is to say that what the human visual system does is rely on perceptual learning, and having rarely encountered distorted heads, sees both as regular heads. But another argument is, essentially, "who cares?" Whether it's a regular head or a distorted head, I know it's a head shape of some description, and for 90% of tasks that's all I need. So this was a big thought, as I mentioned, in computer vision in the 1990s. So Olivier Faugeras (O. Faugeras, 1995) was a big advocate of this approach:

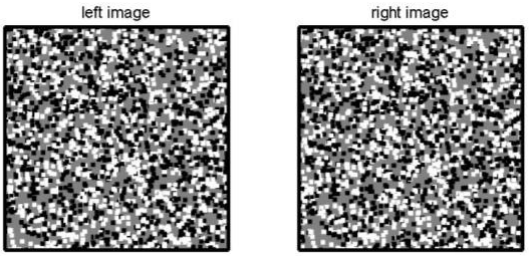
> "We usually think of physical space as being embedded in a three-dimensional Euclidean space, in which measurements of length and angles do make sense. It turns out that for artificial systems, such as robots, this is not a mandatory viewpoint and that it is sometimes sufficient to think of physical space as being embedded in an affine or even a projective space."

And affine accounts of structure from motion (Koenderink & Doorn, 1991), stereo vision (O. D. Faugeras, 1992)(Hartley et al., 1992), navigation (Zeller & Faugeras, 1994)(Beardsley et al., 1995), and object recognition (Jacobs, 1994)(Carlsson, 1998) soon emerged.

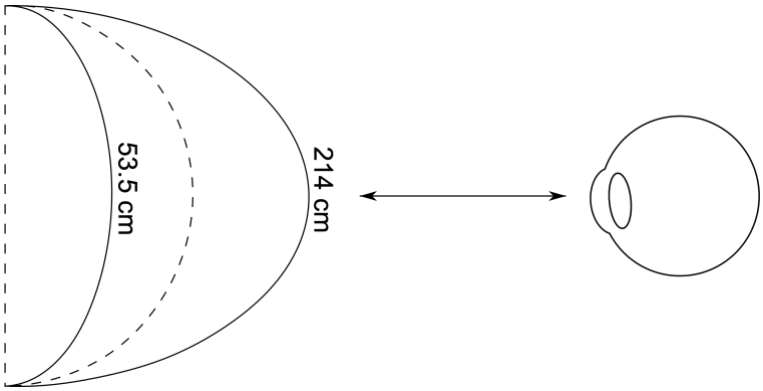
But whilst this is less of a focus in computer vision today, the affine approach has really taken off in human vision. And a good reason is that human vision is subject to considerable distortions. Probably the most well studied, and my favourite one!, is the distortion of stereo 3D

shape with distance. So the experiment I'm going to describe – (Johnston, 1991) – just uses stereo vision, but the same distortions have been replicated with other depth cues present.

How do you just test stereo vision (depth perception with two eyes)? You use random dot stereograms which imbed an image in the difference between the images to the two eyes:



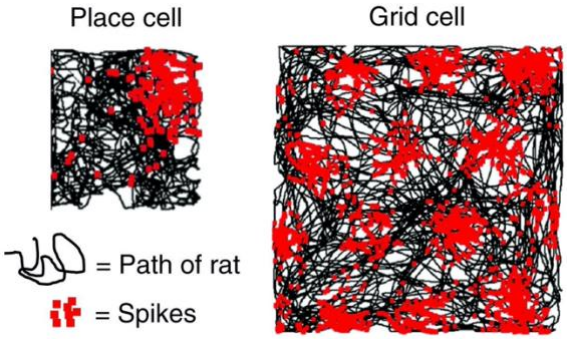
In (Johnston, 1991) a cylinder was viewed side-on and participants could adjust its depth. Their task was simple. Set the cylinder's depth so that its depth was proportional to its height, as indicated by the dotted line below.



Now at a viewing distance of around 1m, participants were pretty good at the task. The problem is when they did the task either at nearer distances or further distances. At near distance (53.5cm), the cylinder they produced was massively compressed in depth, suggested that stereo depth is accentuated at near distances (since participants experienced this as a regular cylinder). By contrast, at further distances (214cm), the cylinder they produced was massively expanded in depth, suggested that stereo depth is reduced at far distances (again, since participants experienced this as a regular cylinder).

For a lot of researchers (though not me!), this sound like human vision just taking an 'affine' approach. Understanding and explaining these distortions is critical for understanding human vision, and this will be a large component of my work at Columbia.

5. Grid Cells: Finally, going back to the question of navigation, I want to briefly address one thing that often comes up. Some of you might know about 'grid cells'. These are cells that have been found in mammals that fire when the animal crosses one of its firing fields, which are locations in space that cover the whole ground and are arranged in a grid. The firing fields for one grid cell for one environment are shown below. Also shown is a place cell that fires when the animal is in a specific location. And together grid cells and place cells won the 2014 Nobel Prize in Physiology or Medicine.

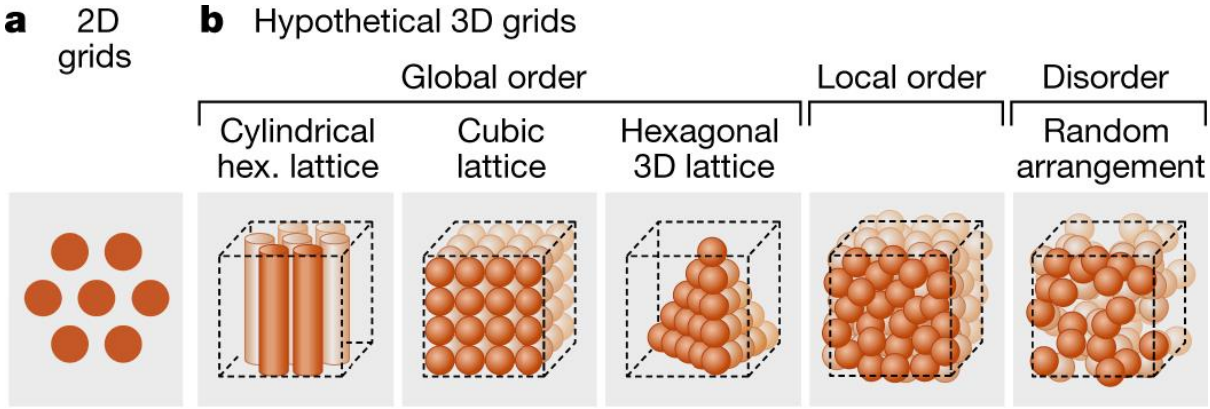


But there's some real debate about just how regular grid cells are. For instance, they can be distorted by the shape of the arena the rat is in, or by the location of rewards. But there's still a strong argument in the literature that they provide a Euclidean (accurate) map of space. And this is very much the approach of DeepMind (Banino et al., 2018), who work on inducing the very same grid cell arrangement spontaneously in AI (deep neural networks):

> "…we argue that grid-like representations furnish [AI] agents with a Euclidean geometric framework – paralleling the proposed computational role in mammals as an early developing Kantian-like spatial scaffold that serves to organize perceptual experience…"

But the big unanswered question going into our Royal Society meeting was how do grid cells work in 3D space. Because up till now they've only really been tested on 2D surfaces.

We had two groups presenting their recently published work exactly on that question. So the '2D grids' in the image below are what were known about grid cells before. Then there were three predictions ('Global Order') of how these might work in 3D, the idea being that the 2D grids are just a 2D cross-section of one of these 3D arrangements. But to everyone's surprise, this didn't turn out to be the case. So in bats flying through the air, they found 'Local Order'. And in rats climbing in a 3D maze, they found 'Disorder'!



This finding is especially striking because rats and bats diverged evolutionarily 65 million years ago, so this seems to be a general finding about mammals. As the authors of the bat study note:

> "…suggestions that grid cells are involved in geometric computations … were motivated by the highly geometric, periodic representation of 2D space by grid cells. Given our findings on the absence of global periodicity in 3D, it seems less plausible that 3D grid cells are involved in general purpose geometric computations…" (Ginosar et al., 2021)

Or, to put the point another way: How does the brain do 3D for vision and navigation?

We still don't really know…

NEW APPROACHES TO 3D VISION
Royal Society, 1-4 Nov 2021

Website / Recordings

DAY ONE (1st Nov) - Seeing Beyond SLAM
Chair: Andrew Fitzgibbon FREng (Microsoft)

Session One: **Neural Scene Representation**
SM Ali Eslami (DeepMind): "Neural priors, neural encoders and neural renderers"
Ida Momennejad (Microsoft Research): "Multi-scale predictive representations and human-like RL"
Session One Discussion (Fitzgibbon / Eslami / Momennejad)

Session Two: **Perception-Action Loop**
Sergey Levine (UC Berkeley and Google): "Generalization in data-driven control"
Andrew Glennerster (University of Reading): "Understanding 3D vision as a policy network"
Session Two Discussion (Fitzgibbon / Levine / Glennerster)

DAY TWO (2nd Nov) – Animals in Action
Chair: Matteo Carandini (University College London)

Session One: **Locating Prey and Rewards**
Jenny Read (Newcastle University): "Stupid stereoscopic algorithms that still work"
Aman Saleem (University College London): "Visual processing in the brain during navigation"
Session One Discussion (Carandini / Read / Saleem)

Session Two: **Navigation in 3D Space**
Kate Jeffery (University College London): "The cognitive map of 3D space: not as metric as we thought?"
Gily Ginosar (Weizmann Institute of Science): "Locally ordered representation of 3D space in the entorhinal cortex"
Session Two Discussion (Carandini / Jeffery / Ginosar)

DAY THREE (3rd Nov) – Experiencing Space
Chair: Mar Gonzalez-Franco (Microsoft Research)

Session One: **Theories of Visual Space**
Dhanraj Vishwanath (University of St Andrews): "Tripartite encoding of visual 3D space"
Paul Linton (City, University of London): "New approaches to visual scale and visual shape"
Session One Discussion (Gonzalez-Franco, Vishwanath, Linton)

Session Two: **Challenges for Virtual Reality**
Sarah Creem-Regehr (University of Utah): "Perception and action in virtual and augmented reality"
Douglas Lanman (Facebook Reality Labs): "Engineering challenges for realistic displays"
Session Two Discussion (Gonzalez-Franco, Creem-Regehr, Lanman)

DAY FOUR (4th Nov) – Grasping the World
Chair: Jody Culham (Western University)

Session One: **One Visual Stream or Two?**
Fulvio Domini (Brown University): "A novel non-probabilistic model of 3D cue integration explains both perception and action"
Irene Sperandio (University of Trento): "Dissociations between perception and action in size-distance scaling"
Session One Discussion (Culham, Domini, Sperandio)

Session Two: **3D Space and Visual Impairment**
Ione Fine (University of Washington): "Do you hear what I see? How do early blind individuals experience object motion?"
Ewa Niechwiej-Szwedo (University of Waterloo): "The role of binocular vision in the development of visuomotor control and performance of fine motor skills"
Session Two Discussion (Culham, Fine, Niechwiej-Szwedo)

Session Three: **Future Directions**
Chair: Michael Morgan FRS (City, University of London)
Panel Discussion by the Chairs (Fitzgibbon / Carandini / Gonzalez-Franco / Culham)

References

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., … Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), Article 7705. https://doi.org/10.1038/s41586-018-0102-6

Beardsley, P. A., Reid, I. D., Zisserman, A., & Murray, D. W. (1995). Active visual navigation using non-metric structure. *Proceedings of IEEE International Conference on Computer Vision*, 58–64. https://doi.org/10.1109/ICCV.1995.466806

Belhumeur, P. N., Kriegman, D. J., & Yuille, A. L. (1997). The bas-relief ambiguity. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1060–1066. https://doi.org/10.1109/CVPR.1997.609461

Carlsson, S. (1998). Geometric structure and view invariant recognition. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *356*(1740), 1233–1250. https://doi.org/10.1098/rsta.1998.0219

Faugeras, O. (1995). Stratification of three-dimensional vision: Projective, affine, and metric representations. *JOSA A*, *12*(3), 465–484. https://doi.org/10.1364/JOSAA.12.000465

Faugeras, O. D. (1992). What can be seen in three dimensions with an uncalibrated stereo rig? In G. Sandini (Ed.), *Computer Vision—ECCV'92* (pp. 563–578). Springer. https://doi.org/10.1007/3-540-55426-2_61

Ginosar, G., Aljadeff, J., Burak, Y., Sompolinsky, H., Las, L., & Ulanovsky, N. (2021). Locally ordered representation of 3D space in the entorhinal cortex. *Nature*, *596*(7872), Article 7872. https://doi.org/10.1038/s41586-021-03783-x

Hartley, R., Gupta, R., & Chang, T. (1992). Stereo from uncalibrated cameras. *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 761–764. https://doi.org/10.1109/CVPR.1992.223179

Jacobs, D. W. (1994). Generalizing invariants for 3-D to 2-D matching. In J. L. Mundy, A. Zisserman, & D. Forsyth (Eds.), *Applications of Invariance in Computer Vision* (pp. 415–434). Springer. https://doi.org/10.1007/3-540-58240-1_22

Johnston, E. B. (1991). Systematic distortions of shape from stereopsis. *Vision Research*, *31*(7), 1351–1360. https://doi.org/10.1016/0042-6989(91)90056-B

Koenderink, J. J., & Doorn, A. J. van. (1991). Affine structure from motion. *JOSA A*, *8*(2), 377–385. https://doi.org/10.1364/JOSAA.8.000377

Muryy, A., Siddharth, N., Nardelli, N., Torr, P. H. S., & Glennerster, A. (2020). Lessons from reinforcement learning for biological representations of space. *ArXiv:1912.06615 [q-Bio]*. http://arxiv.org/abs/1912.06615

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*(10), eaax5979. https://doi.org/10.1126/sciadv.aax5979

Zeller, C., & Faugeras, O. (1994). Applications of non-metric vision to some visual guided tasks. *Proceedings of 12th International Conference on Pattern Recognition*, *1*, 132–136 vol.1. https://doi.org/10.1109/ICPR.1994.576244

Zhou, B., Krähenbühl, P., & Koltun, V. (2019). Does computer vision matter for action? *Science Robotics*, *4*(30), eaaw6661. https://doi.org/10.1126/scirobotics.aaw6661