Response to Commentaries

Roger Watt
Department of Psychology
University of Stirling
Scotland
FK9 4LA
r.j.watt@stirling.ac.uk

Steven Dakin
Institute of Ophthalmology
University College London
London
EC1V 9EL
scdakin@gmail.com

We are very grateful to the authors of the two commentaries on our target article. Our overall response is that there is much to agree with in what they say and little to disagree with. We are gratified to see that both commentaries recognize that the purpose of the paper is to explain a different style of theorising about visual perception.

Our paper describes an account (model or theory, we are not sure which) of vision that takes actual images as inputs – exactly of the form you would take with a digital camera – and produces actual responses of the sort that you could use in an experiment. In principle one could set up a visual task on a computer screen and have a human observer conduct some task, responding on a keypad to the computer, whilst at the same time, a camera placed on their shoulder was also looking at the screen, sending the image to our algorithm to obtain an equivalent response. This is still very unusual for models/theories of vision.

Both commentaries touch on generic issues related to the status of our notion of image description; both commentaries also raise detailed issues about the exact model itself; finally there are some comments about the application of the model to the psychology of reading, both known and unknown.

The commentary by Legge concentrates mainly on aspects of the paper relating to reading and the structure of text, which we will deal with below, but it also makes two important points about the overall theory. First, he raises the question of a single representational outcome as distinct from a system where a perceptual decision can be reached at any of a number of different levels. Second, he notes that we have not said much about the range of ways that image descriptions might be turned into higher level representations. These are fair points, and they are of course inter-linked. Were we pushed to give an answer, we might state that image descriptions are the lowest level at which perceptual decisions are reached, on the grounds that there is little significant extra information available at lower levels but there would be a massive cost (and computational issue) in finding that information. Combinations of image descriptors to generate higher level representations will undoubtedly be important for us in developing this approach. As an aside, we note that the origins of this idea lie in Watt (1991) which explored the logic of and relations between image data structures (such as filtered images), image descriptions, and higher level visual descriptions formed from image descriptions by applying assumptions about scenes.

In their commentary, Dry et al are somewhat more critical of the theoretical enterprise. Perhaps their strongest criticisms concern the issues of comparison with quantitative data and with alternative models. However, they start with a small criticism that we can deal with fairly succinctly. They notice (with eagle eyes) that image descriptors drawn in our figures have an additional parameter, width, not referred to in the text. While it is true that the image descriptions generated by our software include a width parameter, this parameter is not used in the present paper and so was omitted from the description.

Actually, Dry et al are wrong in supposing that the apparently concealed width parameter renders the model underspecified so that others could not replicate our results. We suspect that they are erroneously supposing that our overlap grouping might be based on token overlap (in which case token width would be crucial), whereas it is based on overlap of filter responses themselves.

The next criticism of Dry et al concerns the generalizability of the model to novel stimuli. We have two responses to this, depending on what generalization is taken to mean. First, Dry et al may simply be asking what happens if one takes the study of text images outside of the range of parameters we have used. The simple answer is that we don't know because we haven't done it. We can re-assure Dry et al and other interested readers that the range of stimuli we used and reported are not the result of a careful selection process to find good results – they just are the first thing we did and it worked. We don't see any in principle reason why changing some of the parameters of the text (such as font size or style) should be problematical.

A deeper meaning of generalization could be that there are highly domain specific terms in the model that render it inoperable for other domains. An example of such a domain specific model is actually described later by Dry et al: the use of Delaunay triangulation for symmetry perception (Dry, 2008). Delaunay triangulation can only be applied to data in a specific form: a set of point loci in a Euclidean (or similar) metrical space. Delaunay triangulation can be applied to images made up of patterns of discrete dots, but only once it has been explained how dot positions are computed by the visual system. Delaunay triangulation cannot be applied to an image of a person's face to establish how symmetrical it is. To keep the Delaunay idea, one must bring in additional processes to represent the continuous image by a set of points – and marking up a face image as a set of control points is not trivial. By contrast, we can point our algorithm at any image, set it running and get out image descriptions.

Pursuing the example of symmetry perception a little further, we have previously shown (Dakin & Watt, 1994) that images of mirror-symmetric random dot patterns generate image descriptions that have a very simple and robust property. A symmetric pattern with a vertical axis gives rise to a set of image descriptors in the outputs of horizontal filters whose centroids (ie positions) lie exactly co-aligned along the axis of symmetry. This is shown in Figure 1. We found that this property gives an excellent account of human sensitivity to symmetry. This property is always diagnostic of symmetry, although not the converse: there are two forms of mirror symmetry that do not generate the pattern. The pattern only arises when the symmetry crosses the axis: patterns where the central area is not symmetric do not generate this property – and they are not seen as symmetric Bruce and Morgan, (1975). Similarly, this property only arises when patterns are restricted to orientations perpendicular to the axis of symmetry – also supported by human psychophysical data (Dakin & Hess, 1996; Rainville & Kingdom, 2000).

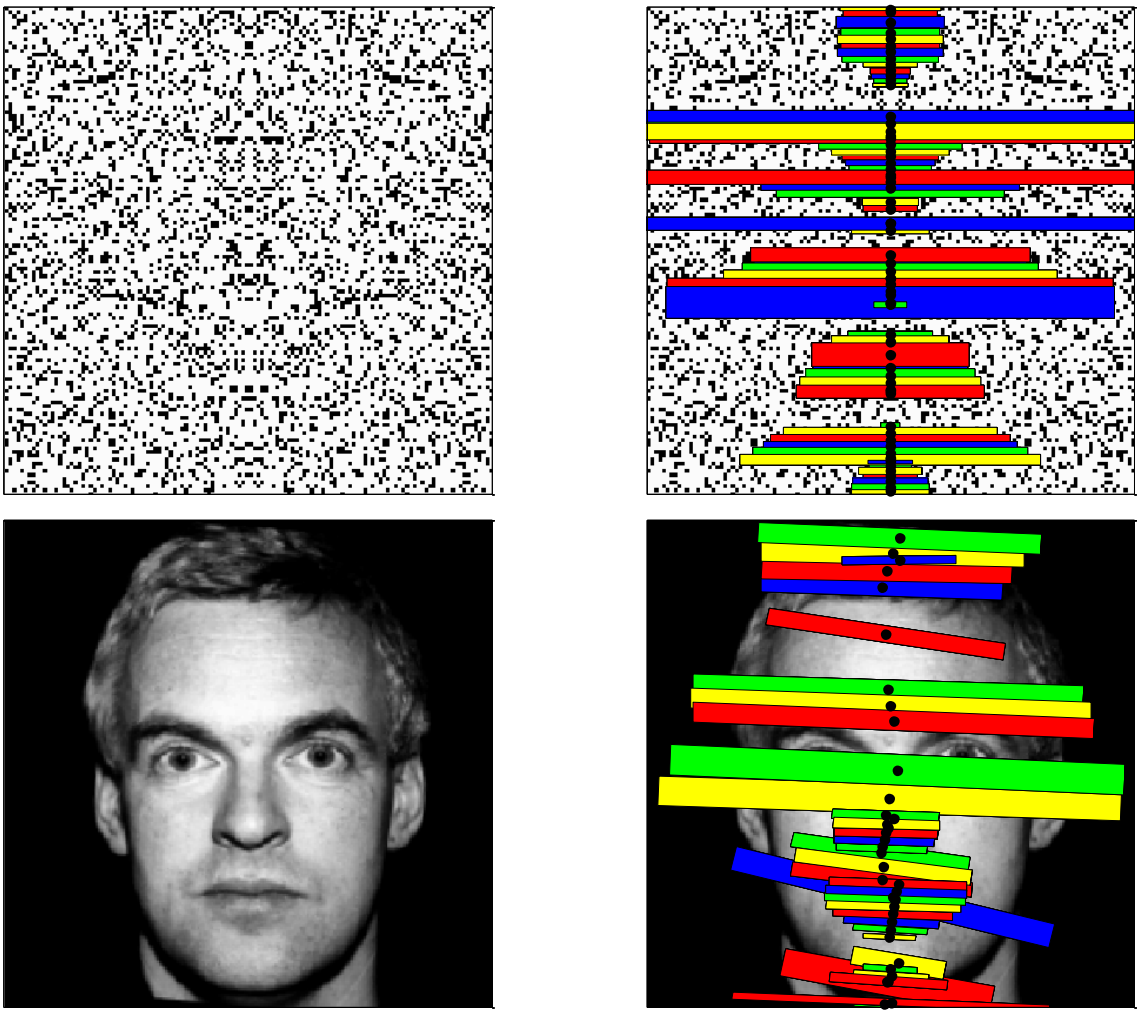

Figure 1 about here

This figure shows two example images (left) and their image descriptions (right). The top image is a pattern of randomly placed dost where the left half of the image is exactly a mirror image of the right half. Notice how the consequence of this is that there is a set of image descriptors whose centres all lie exactly on the axis of symmetry. The lower image is a sample image of a face. The same pattern is seen in the image description: centres closely aligned with the midline of the face which is the axis of approximate symmetry.

We cannot resist the temptation here to conclude this brief consideration of symmetry by mentioning the image description arising from an image of a human face using work we have recently published (Dakin and Watt, 2009). Because faces are roughly mirror symmetric, they also generate this image descriptor alignment property. This is shown in Figure 1.

Thus, the image description approach does not have a problem in operating with other images or indeed with modelling other visual tasks: no additional pieces of theory or software code are needed. Since all inputs to vision are images, there is no input or input domain for which our model is inapplicable. We repeat that this is still unusual for a model of vision.

In order to keep with this theme we skip ahead, for a moment, to the final criticism of Dry et al – the question of how image descriptions compare with other models. We argue there are currently no other models of comparable scope – using actual images and tasks ranging from contrast detection right through to reading – we allow ourselves the luxury of a quick response here. Image descriptions are a language that can be used to describe any image, but also to specify any number of tasks on that image. A trivial further example might help. Imagine that we have a task where an observer is shown a small set of lines, some black, some white and asked whether the longest line is black or white. Each image shown to the observer generates a set of image descriptors. The task can be specified as a sequence of operations in the same language: select image descriptor with largest length parameter; set response to equal sign of mass parameter. In this sense, the theory does more than most – at the very least the number of qualitatively different tasks it can do (and therefore for which it can make predictions) is uncountable. That said, we acknowledge that there is an equally uncountable number of other image description languages.

This point illustrates the approach we are pursuing with image descriptions. In the previous paragraph, we have introduced the idea that an image description might be subjected to a “select descriptor with property xxx” operation. In the target article we used, without comment, an operation “compare sequence of length and y-position values with the sequence of values yyy”. The issue of what visual perception can and cannot do – specified entirely in terms of the scope of such operations - is opened up by this approach.

The penultimate criticism of Dry et al is the hardest to reply to. The authors are irritated that we haven’t done the traditional thing of generating experimental data and testing the model against it, iteratively refining en route. This is exacerbated by their clear sense that our approach is not sufficiently specific. Once again we have several broad responses. First, we would dispute that experimental data, so often carefully set up to give the best chance of justifying a model, is any better than our observational data (in our case, pages of randomly-selected text). So the fact of our data being observational does not strike us as being a problem.

Our next response to this point is that the degree of effort that has gone into understanding the properties of early vision justifies our hijacking of existing findings-such as those relating to contrast transduction (the example they offer) – to derive, for example, the properties of the filters the model employs. Since the image description account is about how the information processed by these mechanisms is read out for subsequent purposes, we suspect that it is entirely safe to do this. We think it very unlikely that a change in the image contrast transduction function that precedes filtering (in the paper we didn’t apply one) will make a noticeable difference to the behaviour we have described in this paper.

Dry et al, consider the Kanisza figure (see Figure 2) as type of phenomenon that requires a more sophisticated model – with components of border ownership, depth differentiation, 2D spatial integration being combined to generate an emergent macroscopic property – the illusory square. They even take the risk of saying that this cannot be achieved by a simple proximity principle. Figure 2 also shows an image description for the stimulus, using just one scale and 2 orientations. Note the presence of elongated structure at the edges of the illusory square. Need we say more?

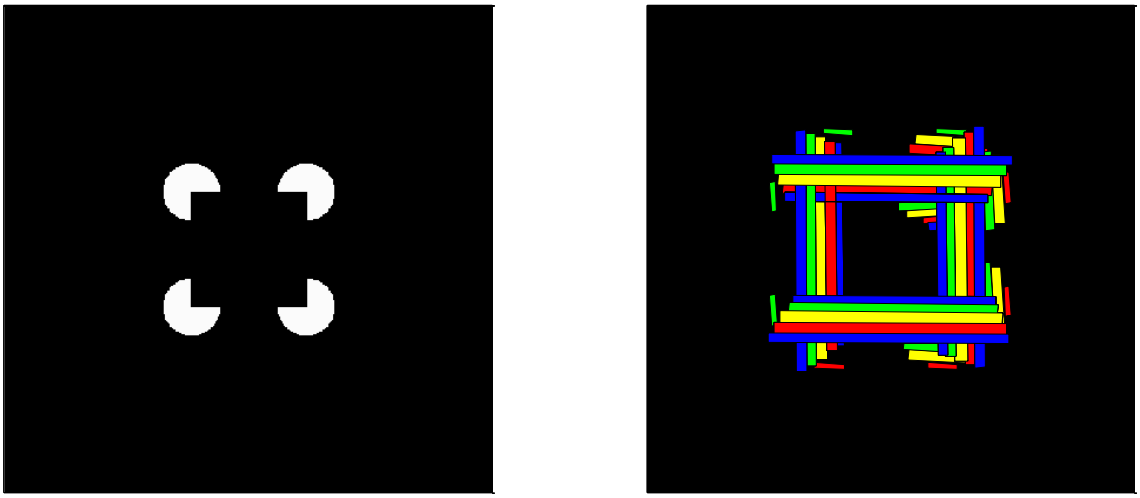

Figure 2 about here

A final example image and description. The figure on the left is the Kanisza triangle, named after the Italian Gestalt psychologist from Trieste. He noticed that the figure appears to be that of a square overlapping 4 circles placed at its corners. This leads to the percept of illusory contours: the apparent edges of the square between the circles have no luminance difference to cue them as edges. The image description on the right (at an appropriate spatial scale) has structures that correspond to the perceived but not present edges of the square.

Finally Dry et al spell out the type of model that they prefer: "…quantified via a direct comparison with empirical data." We agree entirely with their taste in these matters which is why we subjected the image description theory to a direct comparison with empirical data (pages of text). Moreover we would also point out that our account didn't start out as an attempt to explain the quantity of white-space on paper in a typical library, we just got that outcome for free. We felt and still do feel that this unexpected outcome is somehow exciting.

We now make some brief comments about the text and reading issues raised by Legge. Rather than arguing the details, our preference is to ask how much of our theory might fall if any of the issues they raise were to go against our account. So for example, they remind us that people do succeed in reading even when white space between words is simply removed. Had we claimed that our method was the only way that visual information is made available for word recognition and hence reading, we would now admit to having wasted everyone's time. However, as Legge notes, we weren't so incautious. All we claim is that reading will be based on image descriptions.

The other cases that Legge describes, of large letter spacing and of words constructed so that there is no filter that responds to the whole word, are all indeed cases where the type of image descriptions that we have been using in the present paper would not be grouped appropriately into words. Different scales (and orientations) yield different image descriptions and amongst the wide range of possible scales, orientations, grouped only by overlap, there will be various different patterns that can be used to identify a word. As an example, Legge notes that a finer scale would allow letter by letter identification. We think it interesting that at the scales we have inspected, this rather simple pattern which is rather like a bar-code emerges, and we found it rather exciting

that the bar-code turns out to have such high specificity for the word that generates it. This suggests to us an opportunity for great efficiency in word recognition when done in this manner. From that point, it is perhaps disappointing that reading rate drops rather little when forced into other methods of word recognition, but of course reading rate is not just determined by visual factors but also by linguistic factors which can combine to generate reasonable hypotheses about the next word to be read.

We are tickled by the aside about typography in the late classical Greek era and the early common era: that western civilization managed in the early days without white space. They also managed without nylon, petrol and Facebook. We can only conclude that the mind was simply very different in those days.

Acknowledgement
RJW acknowledges the support of a Leverhulme Research Fellowship in preparing both the target article and this reply to commentaries.

References
Bruce, VG and Morgan, MJ (1975) Violations of symmetry and repetition in visual patterns. Perception 4, 239 - 249

Dakin, SC & Hess, RF (1996) Spatial-frequency tuning of visual contour integration. JOSA A, 15, 1486-1499

Dakin, SC & Watt, RJ (1994) Detection of bilateral symmetry using spatial filters. Spatial Vision. 8, 393-413.

Dakin SC and Watt RJ (2009) Biological “bar codes” in human faces. Journal of Vision, 9(4):2, 1-10, http://journalofvision.org/9/4/2/, doi:10.1167/9.4.2.

Dry, M (2008) Using relational structure to detect symmetry: A Voronoi tessellation based model of symmetry perception. Acta Psychologica 128, 75-90

Rainville, SJM & Kingdom, FAA (2000) The functional role of oriented spatial filters in the perception of mirror symmetry — psychophysics and modelling. Vision Research 40, 2621-2644

Watt RJ (1991) Understanding Vision. Academic Press, London