The utility of image descriptions in the initial stages of vision: a case study of printed text

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

Abstract

Vision research has made very substantial progress towards understanding how we see. It is one area of psychology where the three-way thrust of behavioural measurements (psychophysics), brain imaging, and computational studies have been combined quite routinely for some years. The purpose of this paper is to demonstrate a relatively unusual form of computational modelling which we characterise as involving image descriptions. Image descriptions are statements about structures in images and relationships between structures. Most modelling in vision is either conceived in fairly abstract terms, or is done at the level of images. Neither is entirely satisfactory, and image descriptions are a simple formulation of age-old ideas about a vocabulary of image features that are detected and parameterized from actual digital images.

For our example, we use the domain of the visual perception of printed text. This is an area that has been characterized by thorough, robust psychophysical experiments. The fundamental requirements of visual processing in this domain are: grouping of some parts if the image into words; at the same time segmenting words from each other. We show how these are readily understood in terms of our model of image descriptions, and show quantitatively that typographical practice, refined over centuries, is about optimum for the visual system at least as represented by our model. In addition, we show that the same notion of image descriptions could, in principle, support word recognition in certain circumstances.

Introduction

Look at the image shown in Figure 1. It is a wild flower that is relatively common in upland areas of Scotland. If this was a real image formed in your eye, what would your visual system have to say about the stimulus?

Figure 1 about here

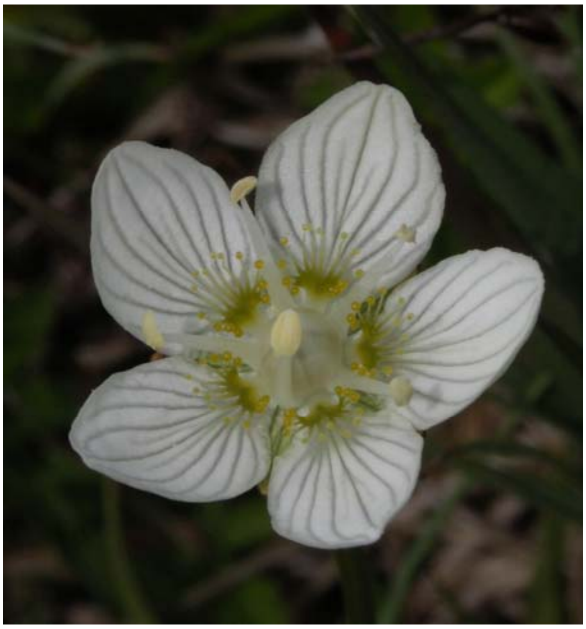


Figure 1: A sample image

There are plenty of things that we can agree, a priori, that vision does not have to say. First, and most emphatically, vision does not have much if anything to say about the stimulus as an image. So it does not tell you how much light is to be found in such-and-such a visual direction (which would correspond to the grey-level in some specified pixel in the figure). The image certainly holds information in the form of how-much at such-and-such a point, but the purpose of vision is to tell you what-type of thing there is and where.

However, vision does not go as far as telling you that the stimulus contains a flower of the saxifrage family, Grass of Parnassus (Parnassia palustris). Vision will deliver information about the contents of the stimulus that you can use, if you have the necessary botanical expertise, to reach this conclusion. The idea that vision delivers information about current stimulation that can be used to make decisions, guide actions and formulate plans is not new and not difficult. It is, alas, rather bland and points out the obvious doing little more than identifying vision with seeing.

Since Hubel and Wiesel (1968, 1977) and Campbell and Robson (1968), vision science has had a powerful theoretical construct to deal with the question of what information is extracted from an image. The receptive field is an information selection mechanism localized at a point in the image: a specific receptive field can tell you whether the input has a particular pattern at a particular place. A set of identical receptive fields, one for each point in the visual field, is a visual filter. The output of a visual filter is a map showing which parts of the input stimulus have pattern that resembles the receptive field. The early stages of vision have many such maps, each dealing with one type of elementary pattern. These maps contain all the information that vision delivers: they are the sum total of what we can see. They have done very little towards seeing, however: all they tell you is how-much pattern there is in such-and-such a visual direction.

The transition to the next type of representation is a crucial one. Where a specific point in an image sufficiently resembles a particular pattern, we could say that it belongs to a specific instance of that pattern, such as an elongated edge. We will use the word structure to refer to a specific instance of a pattern. A significant number of points in an image map will belong to any given structure in this way, the information they hold can be replaced by a description of the structure. This leads us to think of the visual field, not as a fixed array of points each holding how-much of some pattern, but as variable list of localized structures.

In almost a repeat of the previous paragraph, we now say that these primitive structures, which will be simpler than whole objects, can each belong to a larger structure. The key to this level of representation lies in the process of grouping: deciding what belongs with what, and most importantly also deciding what does not belong with what.

To summarize: we want to explore the notion that vision delivers a symbolic description of stimuli, using image primitives rather like nouns to refer to elementary structures in the image. Primitives have properties: lengths, orientations, positions and so on. These properties are rather like adjectives: they serve to distinguish between different instances of primitives. Finally the symbolic description also expresses relationships between image primitives, forming groups. This grouping is rather like using verbs to describe the way things (referenced by nouns) interact.

There is a very important consequence of taking the process of vision to this level of a symbolic language-like description. The full set of possible images is colossal, much larger than the number of neurones inside the human cranium: even just the set of discriminable images is colossal. Moreover, every image we form in our eyes is completely novel. Only a linguistic style of image description will provide the necessary generative power to be able to describe such a large number of images.

There is nothing new in the notion of a symbolic stage in vision. Perhaps in recent times it has been explained most persuasively by Marr (1976, 1982). Marr’s Primal Sketch was particularly exciting, 30 years ago, because it suggested that one could take an image and then actually simulate vision to produce a description of the scene imaged. It offered the prospect of actual models of vision rather than hypothetical and abstract formulations of vision.

One might suppose that 30 years would be enough to have fulfilled this promise: that there would be computational models of biological vision that produce descriptions, at least of images if not of scenes. The technology has been available – you can readily obtain digital images, and desktop computers now have formidable power (actually the graphics card in your home computer has orders of magnitude more power than was available to Marr).

Marr supposed that it is an important goal for vision to find and describe luminance edges in images, since they often (but not always) correspond to the occluding edges of objects in the scene. Marr didn't have access to good information about the general nature of everyday images and Marr and Hildreth (1979) simply had to suppose that the image luminance pattern across the image of an occluding edge would have a locally maximal intensity gradient across several adjacent spatial scales. There is now considerable knowledge about the statistical nature of natural images (for a review, see Simoncelli and Olshausen, 2001 and Geisler, 2008). Since Marr there has accumulated detailed knowledge about the properties and reliability of various types of feature-finding mechanism (for recent work see, Georgeson, May, Freeman and Hesse, 2007; see also Morrone and Burr, 1988; Watt and Morgan, 1985; Watt, 1988).

However, the deep issue concerns how edge information might be represented. A common notion is that of an edge map: an image where the values are some measure of the presence of an edge. An edge map can hold all the information about edges in an image, but cannot tell you directly very much about any of it: only how-much edge at such-and-such a place. Marr's notion of the full primal sketch was richer in that he supposed that there would be primitive tokens standing for localized pieces of edge. Each token would be described by various parameters relating to the piece of edge that the token referred to. This area – what the symbolic description should be – remains poorly explored (see Watt, 1988 and Watt, 1991 for some post-Marr discussion).

We use the framework developed by Watt (1991) to take the notion of a primal sketch further: an image description that can be computed from any digital image. The purpose of this paper is to show one way in which a symbolic form of image description might be built and how that might open up new explorations of vision. We expect that the model we describe is wrong in many different details. However, we persist because we think the architecture of a model which generates a symbolic description of an image is useful. It is not known with any certainty how such a model might be implemented in neurones inside the brain, but recent attention to long-range interactions between relatively remote receptive fields may hold the key (eg. Li and Gilbert, 2002; Polat and Sagi, 1993;Yen and Finkel, 1998).

Having explained the model in some detail, we will turn to a simple but well studied example domain, visual perception of printed text, to show how the notion of an image description can complement an existing set of rigorous and comprehensive psychophysical studies (for a recent review, see Legge, 2007). This domain is a long way removed from scene-perception, but has several useful features. The business of grouping – what goes with what and most importantly what does not go with what – is central and unambiguous to the effective description of a page of text. Pages of text are designed for a specific purpose, and with several centuries of design experience one might suppose them to be well designed. The existence of a specific purpose – word identification and word sequence representation – makes it possible to establish whether a model of vision is succeeding in doing something useful.

Making image descriptions

The model we describe has 3 steps (details are provided in the appendix). The first of these is familiar to all vision scientists: filtering. This process takes an optical image and produces a range of different copies, each filtered to enhance some particular feature. The second step is more novel: to produce descriptions of the significant structures in each of the filtered images. This second step is not as familiar but is critical because it represents a shift from the analogue domain of images into a quasi-linguistic domain of image primitives. The third step is to group together the image primitives, where appropriate, to produce descriptions of larger scale entities in the image.

Step 1: filtering

The early stages of vision appear to employ rather simple linear filtering processes (e.g. Campbell and Robson, 1968; Wilson and Bergen, 1979). The input to such processes is an image and the output is an image representation, referred to as a filtered image. The filtering process is modelled by convolution between an image and a filter function: if the filter function is thought of as a receptive field, and the output of that receptive field as a firing rate, then the filtered image shows the output of a sheet of identical receptive fields densely covering the whole image. The value at any one point in a filtered image is the response amplitude (firing rate) that would be found for a receptive field centred on that point. Technically speaking, the filtered image is an image function because of its formal structure: to read it, you specify a point in image space and then you can find out the response value.

Visual filters are selective for elongated structures in images and for present purposes vary in three different parameters: they have an orientation preference; they have a spatial scale, or a width preference; and they have 4 different spatial phases, or a preference for a particular type of luminance pattern – light bar, light/dark edge, dark bar, and a dark/light edge. Example filters are shown in Figure 2. The 4 phases are shown for a vertically oriented filter of one spatial scale.

Figure 2 about here

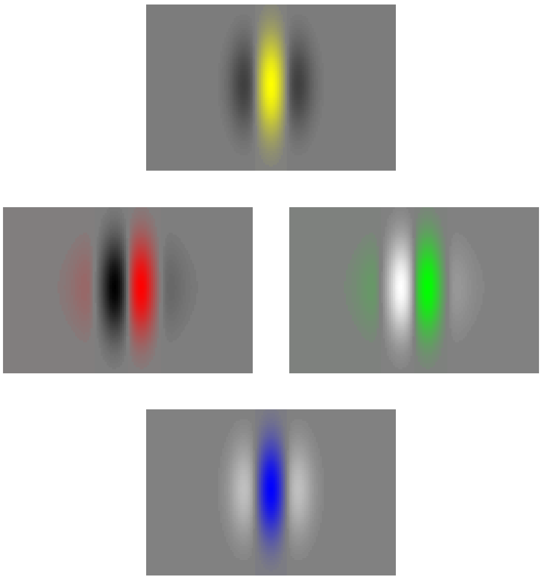


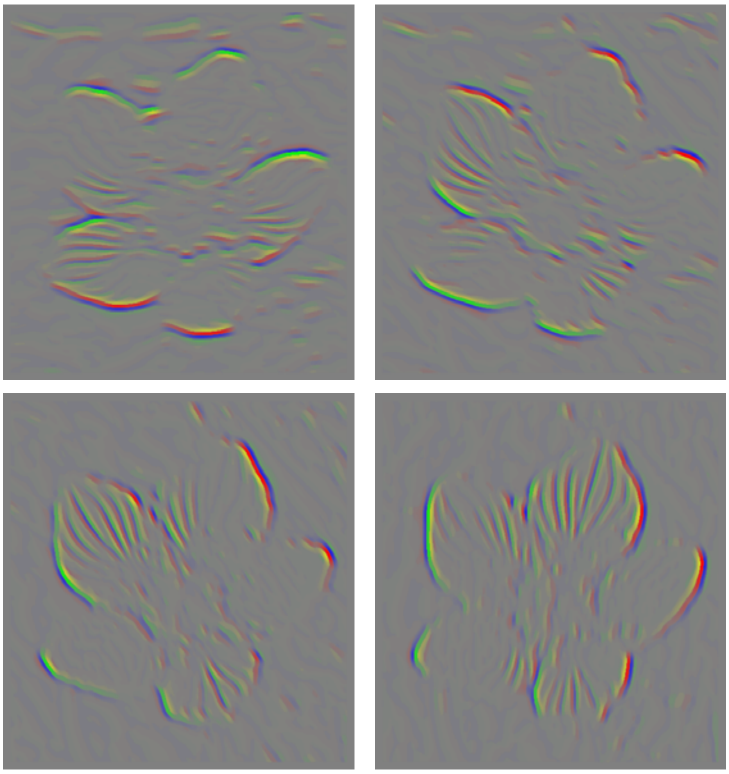
Figure 2: *This figure shows sample filters. The four phases for one orientation and spatial scale are shown. Note that the filters form pairs that are just the negative of each other. The colour coding of the different phases relates to subsequent figures.*

In essence, filtering is equivalent to cross-correlation: a filter will respond strongly to parts of an image that have structure which is similar to the filter. Figure 3 shows the spatial pattern of response of filters in 4 orientations to the image. The responses of filters with different phases are shown in different colours. These spatial patterns are called filtered images. Any image can be taken and filtered: the process is entirely blind to the contents of the image.

Several observations can be made about the filtered images in Figure 3. Notice, for example, that there are very strong responses to the edges of the petals, and weaker ones to the markings on the petal. The apex of the top petal in the image causes a strong response in the horizontal filter (top left). Moving clockwise round the petal edge, you can see that the place where there is a strong response moves progressively round through the other orientations to the vertical filters. Different parts of the edge of the petal cause responses in filters of different orientations.

Although the filters are fixed structures, the pattern of their response is a direct consequence of the image. Moreover, provided a full set of orientations and scales is used, then there is nothing lost from the image.

Figure 3 about here



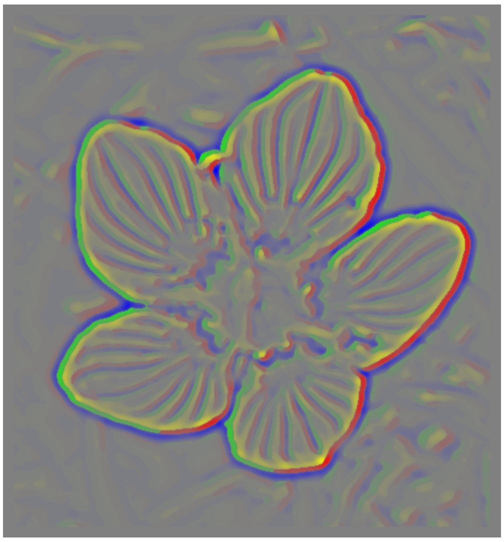


Figure 3: *Filtered images are shown, for the top petal in* Figure 1*. For each, grey represents no response, and the brighter the colour, the stronger the response. Different colours correspond to different filter phases. The different graphs show the responses of filters with different orientations. The lowest panel shows the sum of all the orientations, illustrating how complete the filtered images can be.*

Step 2: finding primitives

Filtered images contain characteristic spatial forms of response pattern. The average spatial form of filter response (after rotation to align outputs from different orientation filters) is shown in Figure 4 with height standing for filter response amplitude. The predominant form of filter response is an elongated ridge of high response with a sharp fall off in response on either side. The coloring of the surface shows the variability in filter response: red areas have the highest variability. This suggests that the main differences between individual responses lie in their length.

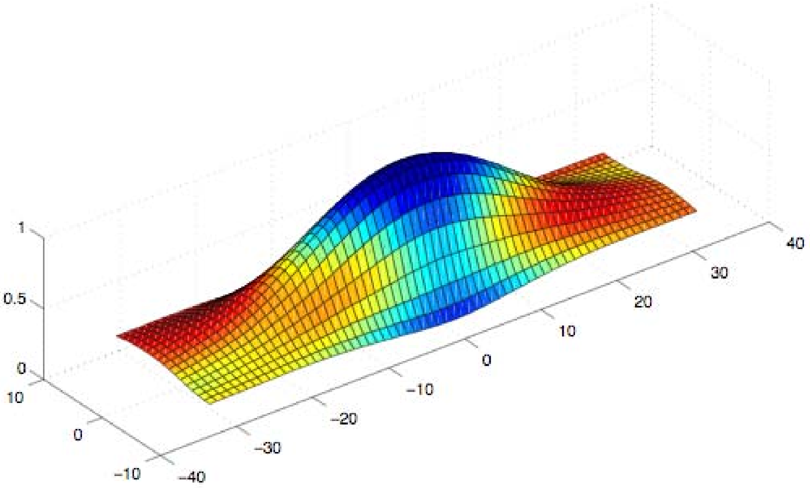


Figure 4 about here

The main area covered by a ridge will be referred to as an image region. Each region is an area of the image where a filter of a single phase produces a spatially uninterrupted region of high response. The response of a filter across an image is a set of such regions, varying mainly in four parameters: position, orientation, length and amplitude.

The critical part of this is that each region is determined by the image contents: its form is a strong consequence of the structure of the image at the place where it occurs. These regions could be the elementary features that are used for subsequent image understanding, and so we will treat them as primitives (see Watt, 1991 for more details).

Each region is described by a set of parameters which encapsulate all that is important and useful about that region:

i) position
(centroid in two dimensions),

ii) the orientation of its longest axis
(obtained by finding the best fitting straight line through the ridge),

iii) the length of the region along that axis
(spatial standard deviation of filter response along ridge)

iv) and the overall response mass
(sum of filter response within region).

These parameters are all robust in that they are not much affected by small details and variations in the filter responses.

Figure 5 shows the regions that are found from the filtered images in Figure 3 (for simplicity, just one phase is shown). Each region is represented by a rectangular block and the dimensions of the block correspond to the measured properties of the region – orientation, length and width.

Figure 5 about here

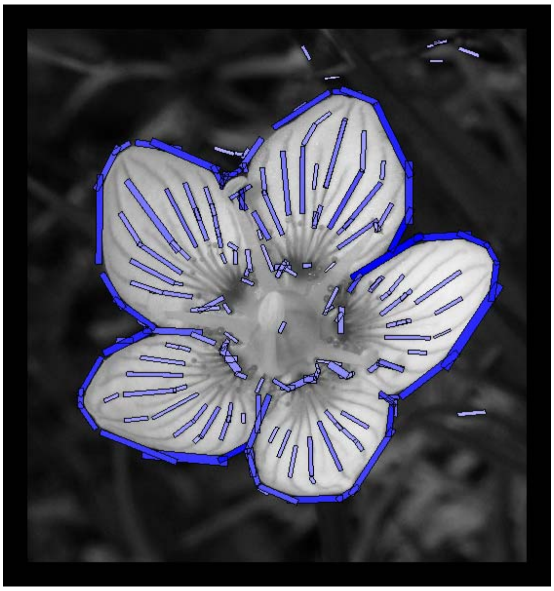


Figure 5: *A section of the example image is used for this figure. The panel shows a graphical representation of the regions as described.*

To render the process of generating an image description practical, we need a rule to determine the actual boundary of each region. In a simple operational sense, we can say that every region ends where the response reaches zero, and then use some slight further processing to set the filter responses to zero where we don't want them. Setting a filter response to zero is, itself, a simple threshold operation: a filter output has to exceed some threshold value before it can be used further. In simple terms, this could be an image wide threshold: the same threshold is applied everywhere in the image. There are many other alternatives, such as local thresholds related to the contrast of the image at that point – local gain control mechanisms are effectively this (for example, see Heeger, 1992).

Step 3: grouping primitives

The final step is concerned with identifying which regions might belong together. Regions tend to lie close to each other in space where they belong together. Proximity is known to be a useful and powerful cue for grouping. Overlap – ie spatial co-incidence – is the most extreme form of proximity. It is also the only form that does not require a spatial metric or a critical distance. In this sense, it offers the simplest form of grouping.

Any continuous line will give rise to a series of region primitives from different orientation filters that overlap in space at their ends and so can be grouped along the contour. This can be seen in the responses to the edge of the petal in. The resultant region descriptions can be grouped together. The regions in Figure 6 are the subset of those from Figure 5 which are mutually grouped by virtue of filter response overlap. As can be seen, this simple grouping process, using only spatial overlap between filters of different orientations (but similar in other properties), does a reasonable job of finding the continuous edge of the object. This is a fundamental step in visual processing of scenes. Marr (1976) referred to the outcome as curvilinear grouping, although he was vague on how it could be achieved.

Figure 6 about here

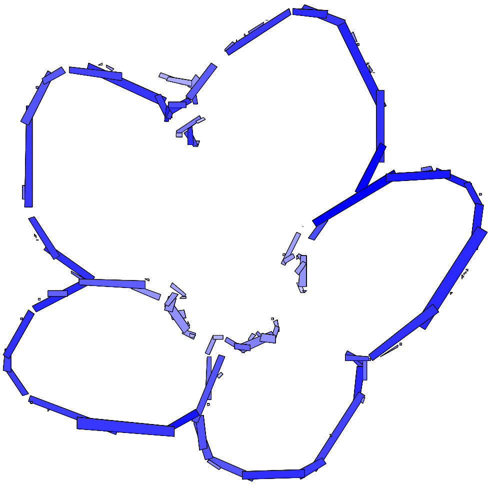
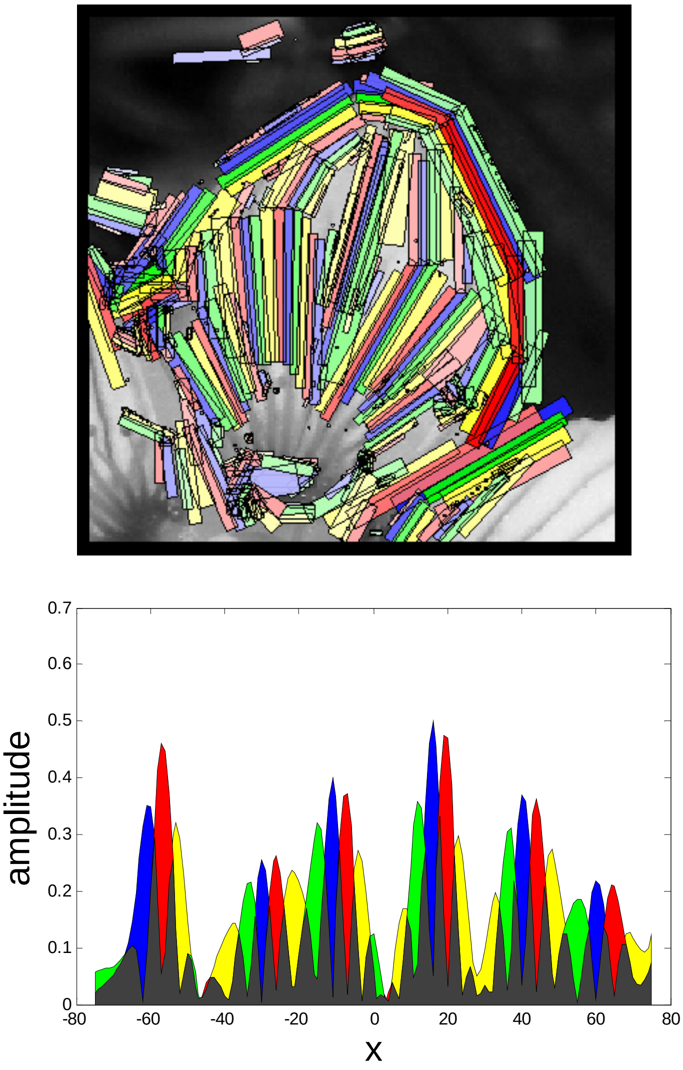


Figure 6: Regions from different orientation filters that are joined because the underlying filter responses overlap. This form of overlap is a simple way to do curvilinear grouping of contour segments into complete curves.

Most objects are circumscribed by their occluding contour: their seen edge from a particular viewpoint. However, inside the occluding contour, there may be considerable further visual information and further image regions. This is the case for the petals of the flower. The top panel of Figure 7 shows one of the petals (the upper left one), with all of the associated regions. The stripes in the petal give rise to a series of parallel regions, overlapping each other across the petal. The second panel shows a cross section through the filter responses across the petal, indicating in gray where the filter responses overlap. This form of overlap can also be used for grouping – and corresponds to what Marr (1976) called theta aggregation.

Figure 7 about here



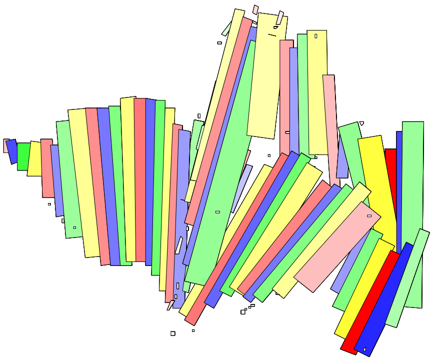


Figure 7: Grouping by overlap is illustrated. Where filter responses overlap (marked dark grey in the middle panel of this figure, the regions can be grouped together. The lower panel shows the set of regions that become grouped together in this way.

There is a good reason for doing grouping in this way, which is related to the notion of proximity. The intensity values in images have a correlational structure which varies with distance: intensities in neighbouring pixels tend to be highly correlated and intensities in distant pixels tend to have a very low correlation (Baddeley, 1997). It has long been recognized that there is an important role for proximity in grouping (for recent work, see Claessens and Wagemans (2005) and Kubovy, Holcombe and Wagemans (1998). The scheme we use here can be thought of as an extreme version of grouping by proximity: overlap is proximity over a distance of zero.

Using image descriptions in an example domain: printed text

We have described a simple process that leads to a description of an image in terms of elongated structures with measured mass, length, orientation, position. This process can be applied to any image and although it has a number of design decisions, the only free parameter that remains is the threshold level used to delimit regions. We will now explore a specific domain with the image description concept: the printed page. Although this seems like a long way from the normal types of image that vision deals with and therefore not an obvious example, the printed page has two significant benefits: i) we can be definite what a successful outcome for the performance of a model of visual perception should be; and ii) there is a body of research which has established experimentally what the limits on successful performance are for human observers.

Vision should deliver a description of a printed page that has one group for each word; those groups should each have information that allows the direct identification of the word. The identification step is a powerful one: the proposal is that there should be a dictionary of image descriptions, so that given the description, the identity of the word can be looked up. This provides the means to quantify the performance of a visual model, without having to inspect its output. That quantified behaviour can then be compared directly with data from real observers. In principle, one could do exactly the same thing for images of flowers (and it should be done).

Typography, at its crudest, is a design process that selects where on a page to place fixed forms: letters. The letters are placed with some obvious and easily understood constraints. They don't overlap, for example, and in fact they don't usually touch each other (the exceptions, which are less obvious in computer typeset material, are for certain combinations of letters, such as ‘tt’ and ‘fi’). Presumably this constraint relates to the mechanics of typesetting – letters starting out as forms in enclosing rectangles – and also relates to the visual process – which is particularly adept at isolating mutually inter-connected regions of ink.

Centuries of typography have presumably refined practice so that current practice will be close to the optimum: maximum legibility and minimum cost. The optimum will depend on various factors, but a very significant one is the cost of paper and so we can expect that the amount of white space adopted by typographers will tend to be as small as convenient. However, the constraint of legibility works in the opposite direction. It is clear that too little white space makes reading a piece of text difficult to read (see for example, Hartley and Burnhill, 1977).

The essence of reading a paragraph of text lies in selective grouping: the letters of a word need to be grouped together, but not with letters of other words. It is trivial to see that the use of white space is the cue to this.

There is a considerable literature on the psychology of reading, and in two areas there are results that are of direct significance for vision. There has been considerable research into the effects of letter and word spacing on efficient reading. Studies of the effect of word spacing have tended to study the consequences of reduced word spacing for eye-movement patterns whilst a reader follows a page of text (Morris et al, 1991; Epelboim et al, 1994; Rayner et al, 1998). Reduced word spacing has consequences for where the eye will be moved to next.

A word provides several sources of information about its identity: the sequence of letters; the overall shape defined by the sequence of letter shape variations; even the length of the word alone is informative. There is a large literature on this that has typically involved disrupting one or other cue to identity. An early claim was that word length and shape (defined as the locations of ascenders and descenders in a word) did not provide much actual information because they are highly ambiguous (Groff, 1975). This was corrected by Haber and Haber (1981) who showed that taking word syntax into account could reduce massively the ambiguity. Subsequent studies have either favoured the idea that word shape is useful (eg Haber at al, 1983) or have reached a contrary view (eg Paap et al, 1984). Underwood and Bargh (1982) argued that the consequences of altering word shape were ameliorated because other routes to word identification were available and switched to. A detailed study in which words were created variously in mixed case, with mixed letter size and mixed letter spacing helped to clarify the effects (Mayall et al, 1997). Their main conclusion is that the word shape effects tend to arise when they lead to either inappropriate grouping of letters or when they disrupt trans-letter features.

We now explore the visual consequences, at the level of image descriptions, of various typographical practices. The results will be related to this literature. In discussion after the exploration, we will consider some further recent studies from a perspective of visual perception of words.

Images of printed text

To help in exploring this domain, we have obtained 150 images of printed text. The images are photographs of pages of books, fiction and non-fiction, from a broad range of publication dates (1950 – present), and paperback and hardback. Each photograph was processed so that the horizontal width of the character ‘n’ was a standard size. In all that follows, this dimension (referred to as n-size) is the unit of distance.

Figure 8 shows a sample printed text image. It has been marked up into the rectangles enclosing characters (in green), the spaces within a word (blue) and the spaces between words (yellow). Although it is an unremarkable example of typography, it has several remarkable features, treating it simply as an image from a visual point of view.

Figure 8 about here

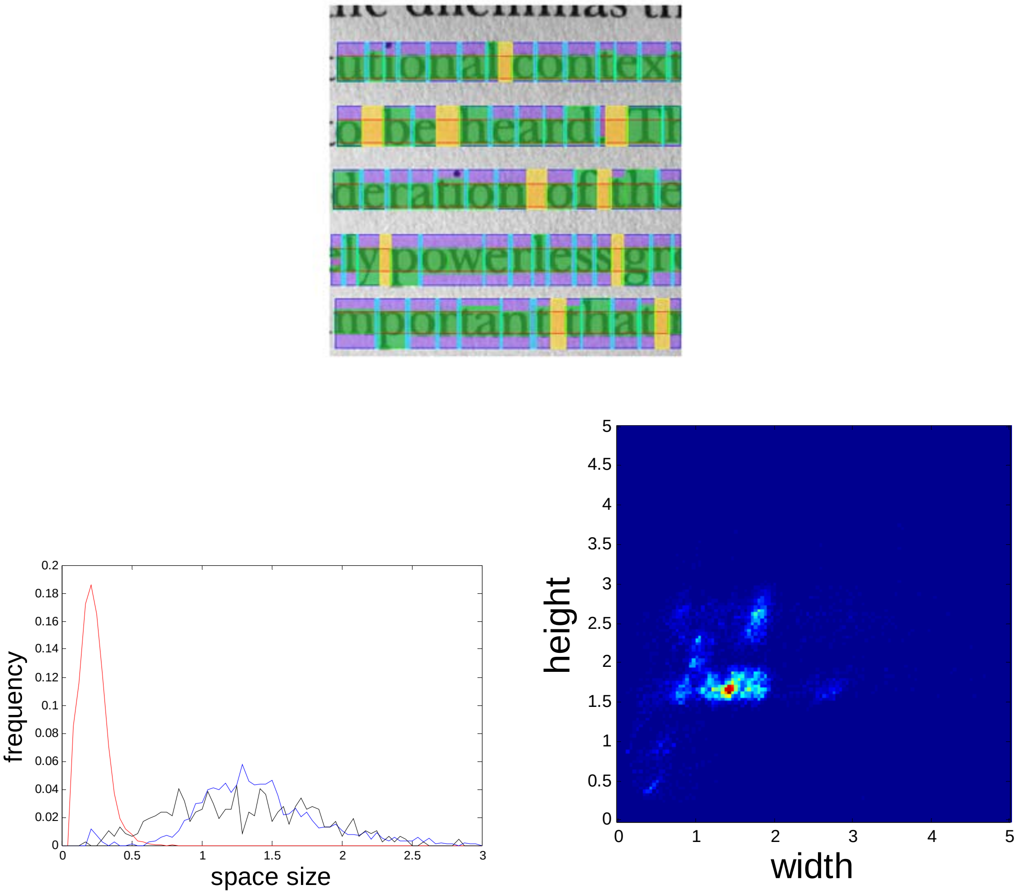


Figure 8: Some spatial properties of printed text. An example image of a page is shown at the top of the figure. Beneath this are drawn graphs showing the distribution of white space parameters(red – character spaces; blue – word spaces; black – line spaces). Note that the inter-character spaces are much smaller than the other two distributions. Alongside is a 2D histogram of character heights vs widths. Note that this shows a small number of distinct clusters.

For example, inspection of the marked-up text image at the top of Figure 8, shows that the word spaces (yellow) are much wider than the character spaces (blue). In fact, the ratio of the width of spaces within words to the width of spaces between words (or between lines) is about 7 for this example. This is enormous: a ratio of 1.1 (10% difference) would be discriminable with some effort, and a ratio of 2 would be discriminable without any effort at all. Similar considerations apply to the space left between lines. On the face of it, the use of white space is extravagant. Figure 8 (bottom left) shows the distributions of character spaces, word spaces and line spaces from our set of 150 images of text. This property of relatively large spaces between words is quite general.

Examination of the form of printed characters reveals more interesting features. One would suppose that printed characters were to be as discriminable from each other as possible. Figure 8 (bottom right) shows the joint distribution of character height and width across the sample of 150 pages. There are clear clusters: heights of around 1.5 and 2.5, with a few between (heights less than 1 are due to punctuation marks); and widths of 1, 1.5 and 2 (widths greater than this are due to characters joined together). There are perhaps just 4 or 5 different clusters in this diagram, which does not suggest high discriminability.

Figure 9 (top) shows the distribution of ink across all the words (top 3 panels) and characters (bottom panel) of our sample of text images. For the words, the three panels show the left edge of the words (aligned at their left edge), the body of words (aligned at their centre) and the right edge of words (aligned at their right edge). Each character is aligned at its left edge and the baseline of the line of text. In terms of generating highest discriminability, one would expect these distributions to be uniform. Where they are very high or very low, they show a feature common to all stimuli which therefore does not provide differential information. In both cases, words and characters, the use of ink does not suggest the best strategy for creating discriminable stimuli.

The actual patterns are also interesting. For words, there is a striking periodic pattern of vertical lines. This periodicity has been observed before (Wilkins, Smith, Willison and Beare, 2008), but has not been explained. Note also that this pattern is superimposed on, and obliterates any indication of the boundaries between letters. For characters, the area with highest probability of ink is a vertical line running down the left of the area. Nearly all characters have some ink in this area. There is a second vertical line running down the right of the area, where 50-60% of characters have some ink. In between is an area with much less ink. On either edge – left and right – there are also areas above and beneath the main body where some characters have ink.

So this brief and incomplete exploration of printed text as images has revealed several features that are worth explaining. The use of white space is curious, not least in that the amount used seems to be excessive, given the cost of paper. The use of ink is similarly curious. Ink appears to be applied in very regular structures that are not optimum for discriminability.

Figure 9 about here

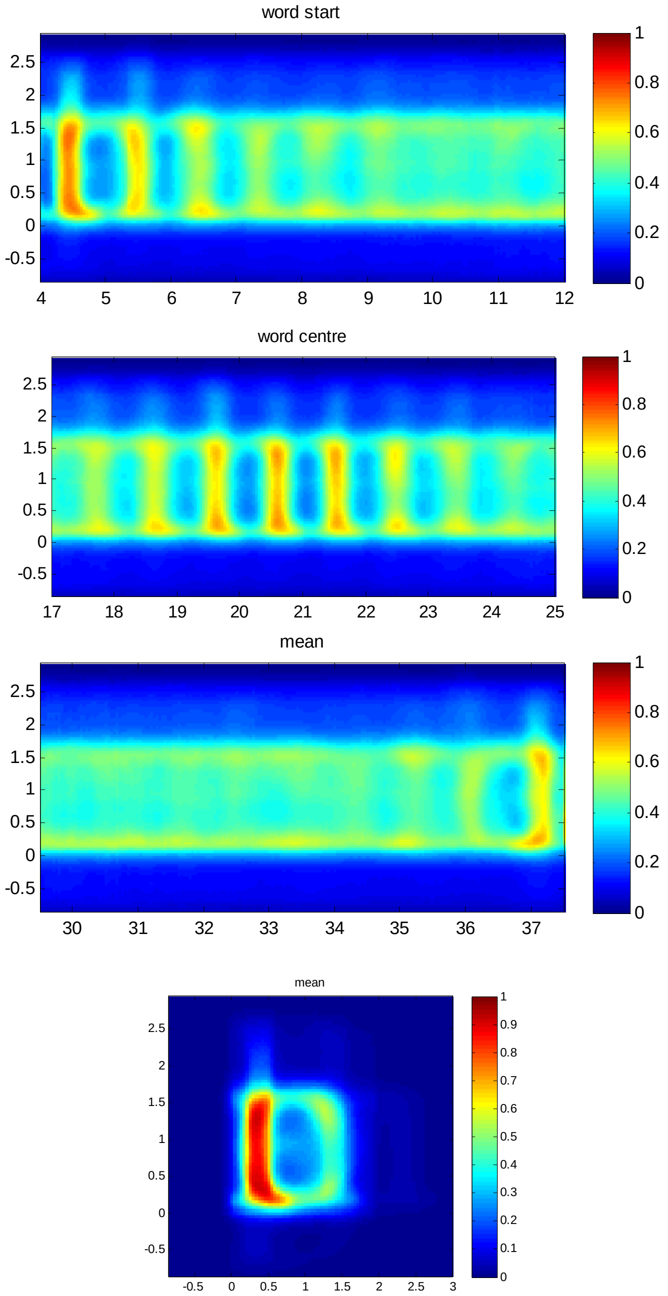


Figure 9: the top 3 panels show the mean image for words. All the words in the sample of 150 images were isolated and added after being aligned at (i) their left edge,(ii) their centre and (iii) their right edge. The pattern revealed shows that a major structure of the printed word is a periodic vertical pattern. The bottom panel shows the same analysis for individual characters.

Image descriptions of printed text

We now turn to apply the image description model to one of these images of printed text. Figure 10 shows a small section of such an image, the filter responses (just for one orientation) and the image description that is produced. Inspection suggests these properties:

1. The regions produced by the page of text tend to be longer, and more uniform in length, than those in the scene image.
2. The grouping of regions for the text is substantially higher and more simply organized than for the scene.

Figure 10 (bottom left) shows the distributions of region length for natural images (black) and text images (red), confirming the first observation. The text images have a very marked peak of lengths at a value of 2. Next, if we take a single region, and then find the positions of all the regions that are grouped with it, we can mark out a zone of space that shows something of the spatial organization of the group. Figure 10 (bottom right) shows the average zone spatial grouped regions: the lines circumscribe the positions of grouped regions. The zones for text images spread laterally much further and are more constrained vertically. This shows that the grouped regions produced by text form along the length of words.

Figure 10 about here

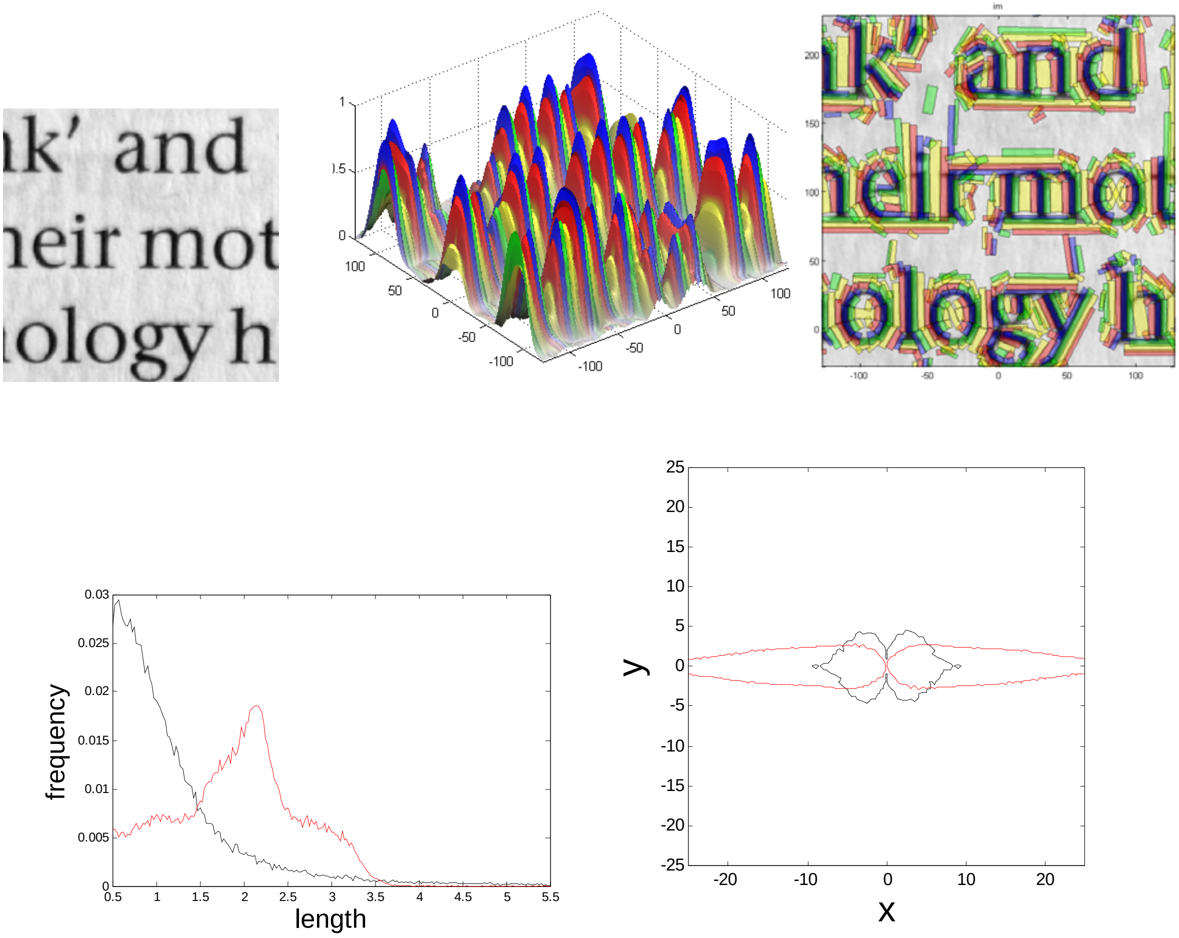


Figure 10: The top shows an example image of text, with a filter response pattern (for just vertical filters) and the image description that is produced (for all orientations). The lower left panel shows the distributions of region length for the natural images of scenes

(black curve) and text images (red curve). The text images have a distinctive distribution of lengths that is longer and more uniform. The lower right panel compares the same two types of image for the relative positions of grouped regions. The lines circumscribe the places, relative to a given region, within which the other regions in the group will lie, on average. The text images show much more constrained structure.

Quantitative analysis of images of text

We now turn to look in some detail at the specifics of the image description produced by pages. For this part of the analysis, we use synthetic images of words generated by placing characters in an image with Times Roman font at 48 point font size. In this way, we can control exactly the spatial relations between characters and words. The words were the 10000 most frequently used words in written English (Leech, Rayson and Wilson, 2001). This set accounts for 93% of all written words.

Analysis 1: filter parameters for grouping, segmentation and recognition

Our analysis starts with isolated words and the issue of grouping filter responses into a single entity. Figure 11 shows an image and the image description for a single word. The dark lines joining the regions indicate that they are joined together according to the overlap rule. In this case, with this particular value for spatial scale and threshold, the regions are all grouped together so that the whole word is represented as a single group.

Figure 11 about here

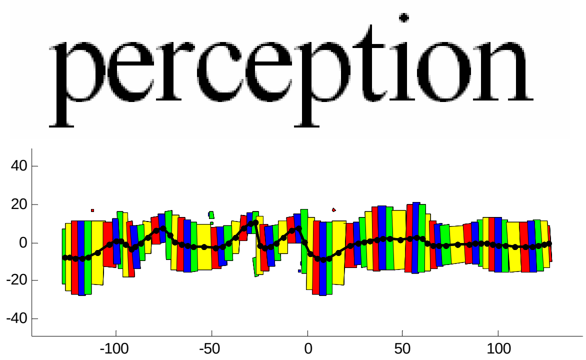


Figure 11: This figure shows an example image of a word. Beneath the word is the image description. The black dots indicate regions that are grouped together because the filter responses overlap in space.

For the first analysis, the single word stimuli were processed as the example above, except that spatial scale and threshold level were varied. For each combination of scale and threshold, the success of the grouping was established. Correct grouping is defined as grouping of all the regions produced by the word image that lie between the leftmost and rightmost extremities of the word in the image. Figure 12 (top left) shows the proportion of words that are correctly grouped, as a function of spatial scale and threshold level. As can be seen, there is a substantial region of this space where words are correctly grouped.

The analysis now considers the issue of correctly segmenting separate words. For this analysis, an image with 6 words was created: 3 each in two lines of text. Word and line spacings were set to the mean values for a sample of 150 pages of text. The word spacing was set to 1.5 times the n-size of the text, and the line-spacing was set to 4.5 times the n-size. Word spacing is measured as the gap between the rightmost extremity of one word and the leftmost extremity of the next word. Line spacing, on the other hand is measured as the distance from the base of letters n (or equivalent) in one line to the base of letters n in the next line. The distance from the base of a descender (eg. letter y) to the top of an ascender (eg letter h) on the same line is 3.3 times the n-size.

Figure 12 (top right) shows the proportion of images where the visual process resulted in a separate grouped description for each of the 6 words. If any group resulted from more than one word, then this was counted a failure. As can be seen, there is still a substantial region of the parameter space where success is guaranteed, although it is different from the area for grouping. This is not unexpected – larger scales and low thresholds will tend to cause the responses to letters and to words to merge (causing correct grouping and incorrect segmentation respectively).

Figure 12 about here

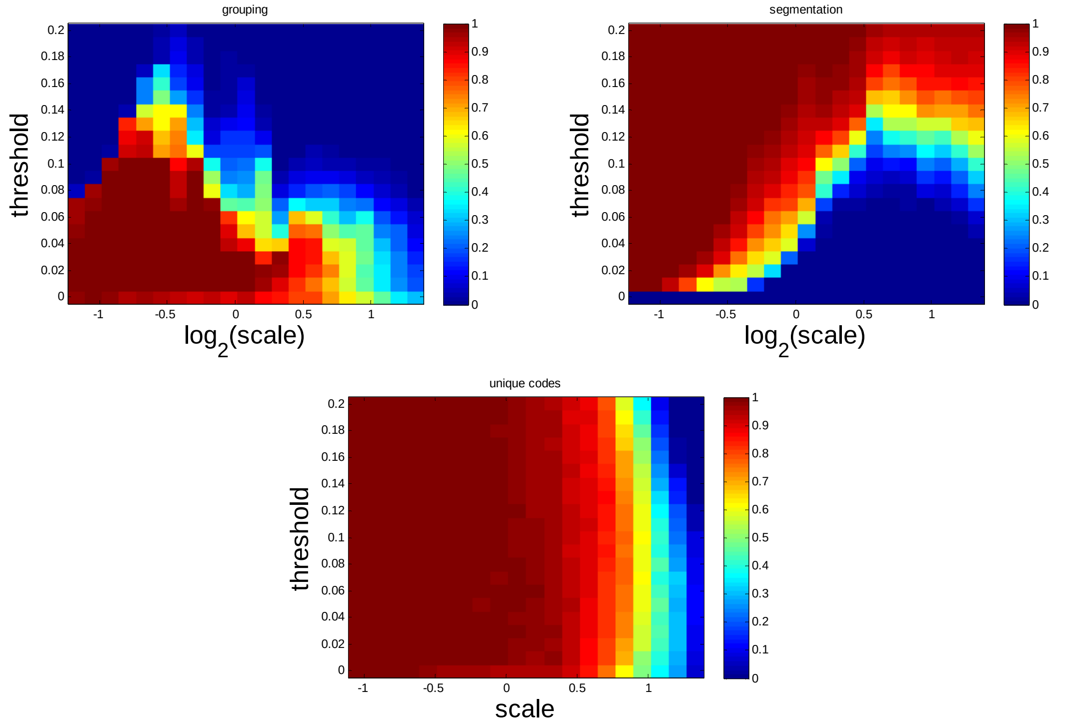


Figure 12: These 3 panels show the effects of filter scale and threshold value on 3 measures of performance: (i) grouping of letters into words, (ii) segmentation of words from surrounding words, and (iii) production of unique codes for each word in a 10000 word database. Spatial scale is given in units of n-size to log(2), so a value of 0 on the x-axis corresponds to 1 n-size.

Our final analysis in this section is more ambitious. Inspection of Figure 11 indicates that the grouped description is actually a simple linear sequence of regions. Does this sequence of regions contain enough information to specify the word? The regions vary in all of their parameters, but we will consider only 2: vertical lengths and vertical positions.

The simplified computational question considered was this. Each word from the dataset is processed to produce a sequence of regions. For each region, the 2 dimensions are each quantized to just 3 categories, so that each region is represented as one of just 9 different possibilities. How many words produce unique codes? Category boundaries on each dimension were explored using a simplex search to find the set of boundaries that yielded the greatest number of unique codes. Figure 12 (bottom) shows the results as a proportion of words with unique codes (ie successes) as a function of spatial scale and threshold. There is a substantial region of the parameter space, including that range required for grouping and segmentation, where the rate of success is very close to 1. Across this region of parameter space, the category boundaries that the simplex found were within 1% of an equal split of the distribution of lengths or vertical locations. The category boundaries for length correspond in Figure 4 above to the points along the ridge at x values of 12 and 24, nicely bracketing the area of highest variance.

This observation, based on specific information derived from actual printed words rather than abstract conceptualizations of word shape, illuminates the question of whether it is possible that words could be recognized without recognizing letters (such as by word shape). Groff (1975) claimed that word shape did not distinguish enough words, especially short high frequency words. Haber and Haber (1982) claimed the opposite. The present result suggests that words can be identified reliably without recourse to letters, in principle if not in practice.

We must note immediately, that it would be a leap to go from observing that 99% of the 10,000 most common words yield unique codes, to supposing that this could be a mechanism for word recognition. These codes are inscrutable: you cannot infer the word from the code in the same way that a sequence of identified letters usually leads to the word. So the unique code for a new word will not be known. So, the most that can be argued is that these unique codes are of some assistance in reading under the right circumstances.

The more general and interesting point is to note how a powerful code can be made from these simple regions categorized into just 3 clusters in each of just 2 dimensions. As a demonstration that image descriptions could be useful, we believe this is relatively convincing.

When we put these 3 analyses together, we note that there is an area of the scale and threshold space towards the bottom left of the plots in Figure 12, where successful grouping of letters into words, successful segmentation of different words and successful recognition can all occur.

Analysis 2: white space parameters

The preceding analyses have shown that the basic requirements of intra-word grouping and simultaneous inter-word segmentation are met by our model of image description for a range of spatial scales and thresholds, when the text is typeset according to normal white space practice. This suggests that the fundamental purpose of white space is to ensure an appropriate pattern of grouping in the visual description. To assess this, we have generated synthetic paragraphs, each with 6 different words arranged into two lines of three words, and measured successful grouping and segmentation as a function of word spacing and line spacing.

Inspection of Figure 12 shows that for the given word and line spacing employed in that data, there is a limited band of spatial scales that will lead to correct grouping and segmentation. For a threshold value of 0.06, for example, the smallest spatial scale that reliably leads to 90% successful performance is about 0.5 n-units (set by grouping limits) and the largest is about 1 n-unit (set by segmentation limits). So the range of spatial scales that generates good performance for these word and line space values is about 1 octave. This measure, can be interpreted as a measure of how likely the visual system is to have exactly the right filter to do the processing. A bandwidth of an octave is a comfortable outcome, as the visual system is almost certain to have at least one filter with a spatial scale somewhere inside that range.

Increasing word or line spacing does not change the probability of correct grouping performance, and hence does not change the smallest spatial scale that can support successful performance. However increasing word or line spacing increases the probability of correct segmentation, and hence increases the largest spatial scale that can achieve good performance. So, for any combination of word and line spacing, we can calculate the range of spatial scales that can support good performance.

Figure 13 (left panel) shows the range of spatial scales, calculated as a bandwidth in octaves, that can support performance at 90% correctly grouped and segmented words, as a function of both word spacing and line spacing. As can be seen, the effects of line spacing and word spacing are almost independent of each other. A word spacing of 1.5 n-units, and a line spacing of 4.5 n-units is required to provide optimum bandwidth of just about 1 octave.

Figure 13 about here

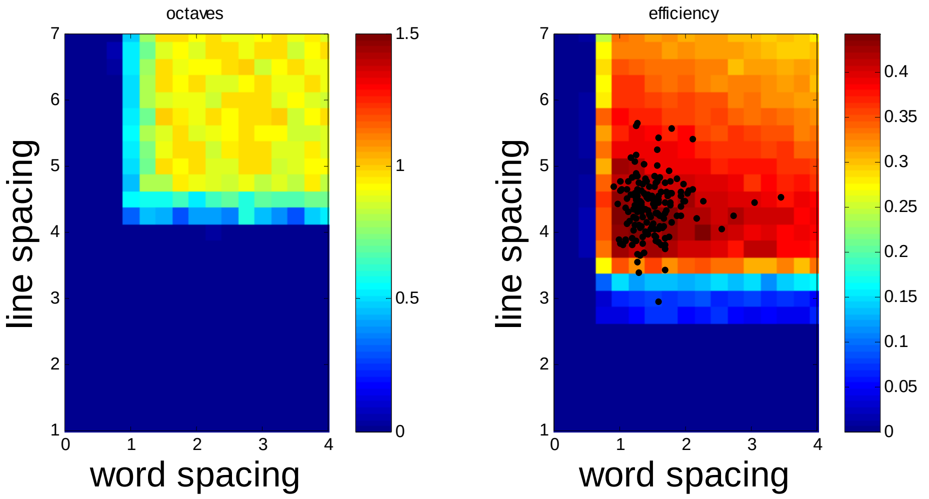


Figure 13: This figure shows an analysis of the effects of white space parameters on visual performance. On the left, the measure of performance is the range of filter scales (in octaves) that will successfully group letters and segment words. As can be seen, that range is 1 octave for word spacing greater than 1 n-size and line spacings greater than 4.5 n-size. The panel on the right shows an efficiency analysis: efficiency is defined as the number of correctly grouped and segmented words divided by the maximum possible number of words per unit area. This panel also shows the actual values for white space employed in our sample of text images. As can be seen, efficiency is best at

around 45%, and most examples of printed text use parameters that will yield efficiency close to this.

Figure 13 (right panel) shows a simple efficiency analysis. We define efficiency here as the number of correctly grouped and segmented words per unit area, divided by the maximum possible number of densely packed words in the same unit area. The highest efficiency is of the order of 45%, which occurs for small but not negligible word and line spaces. The consequence of making either line or word spacing smaller than the optimal (for the model) is to lose nearly all efficiency because words merge together. The consequence of making either white space parameter larger is just to waste some paper with a smaller drop in efficiency.

It is of interest to compare this result with the analysis of real paragraphs from books. These parameters were measured from the sample set of images of pages of books, described above. The joint distribution of word and line spaces from that sample is shown as black dots on the efficiency figure. It is quite tightly constrained. The fit between the efficiency implications of the visual model and typographical practice is very close indeed.

It is also of interest to compare this result with the existing empirical literature. Adequate word spacing has been established as important for reading. The present results simply add flesh to that finding by showing how grouping and more importantly segmentation break down when there is not adequate word spacing. The finding of Mayall et al (1997) that some changes in word shape result in inappropriate letter grouping and loss of trans-letter patterns is also worth exploring further in the light of the behaviour of this model.

Closing observations about pages of text

There is a great deal more could be said about typography as it relates to vision. However, for present purposes, typography has been a vehicle for illustrating a notion of image description, not an end in its own right. That we have a couple of novel explanations for features of typography is a welcome bonus.

This analysis has demonstrated two fundamental properties of printed text, as it relates to image descriptions.

1) The grouping and segmentation of a paragraph can be accomplished readily by using filter response overlap provided white space parameters are not less than a critical value. This explains why the amount of white space required is so high, resulting in an efficiency of under 50%.

Our starting inspection of white space in printed text in Figure 8 noted that the use of white space in printed text was highly wasteful. Whilst this results in a low efficiency for the use of paper, it does result in a readily readable layout which uses properties of this model of the visual system to achieve letter to letter grouping and word from word segmentation. The explanation of the large amounts of white space in a page relates to the nature of the filters that respond to pages of text.

2) Words can be recognized effectively by sub-letter features. These features seem to bridge the gap between word letter sequence and word shape cues.

Our starting inspection of words in Figure 9 identified regular vertical periodic structure. According to our analysis, this regular structure is important for grouping a word. Our analysis of unique codes for words has utilised this structure, treating it as a 1D bar-code in which the bars vary in vertical position and vertical length (but not in horizontal separation). The use of 1D bar-code descriptions has powerful computational benefits, but also costs. The benefits are that they are very robust in an unpredictable environment, they can be very easily adjusted to suit a new typeface, and they are computationally simple to compute. The costs are that a 1D sequence has inherent limitations such as it is not robust to reflection (making a mirror image), inversion (turning upside down) or a combination of the two. It requires letters to be highly constrained in how they are formed and how they are placed. They have to be formed out of principally vertical strokes of a small number of uniform sizes. They have to be placed so that they are aligned vertically, parallel in orientation.

Closing observations about image descriptions

We finish by emphasizing some aspects of our image description model. We start this by looking at what the notion adds to the existing literature on visual aspects of reading. Then we conclude with a more general point about visual perception.

Image descriptions of text

The starting point for creating an image description is a filter. A number of empirical psychophysical studies have identified important characteristics of the filter or filters that underlie the recognition of letters and words. Legge, Pelli, Rubin and Schleske (1985) showed that the critical band of spatial frequencies for reading is around 2 cycles per character. That corresponds very close to the spatial scale that we have found to be optimum for grouping, segmentation and recognition. Our best filter has a scale of 0.6 n-size (see Figure 12), which when expressed in cycles per character becomes 2.1 cycles per character (character size is approximately 1.25 times n-size).

Solomon and Pelli (1994) have shown that letter identification is accomplished by the use of a single visual filter, also finding a centre frequency of around 2-4 cycles per character. Our data in Figure 13 show that the bandwidth of the range of filters that can do the task is quite narrow (around 1 octave), which also suggests that a single filter would be involved. Majaj, Pelli, Kurshan and Palomares (2002) extended the analysis of filters and letter identification, confirming the Solomon and Pelli finding for a wider range of typefaces. They also found that when the text was made larger, there was a tendency for the filter used to be at a higher frequency, when expressed as cycles per character, implying that finer detail is used. Chung, Legge and Tjan (2002) found similar results. Our data in Figure 12 show that there is little if any theoretical cost in using a smaller scale (ie higher frequency) filter, at least over the range we tested.

Chung and Tjan (2007) have qualified this by showing that when one adds further letters into the display (but still only requiring the identification of a single letter), there is a tendency to uniformly use finer scale filters. In this case of letter identification in the presence of other letters, there is a potential cost associated with grouping. Although grouping all the image primitives for a word together (or a single isolated letter) is a good principle for word recognition, grouping could make the identification of individual letters more difficult. Using finer scales will tend to result in less grouping.

There has been little attention paid to stages in the process that would correspond to our image primitives and the general notion of image description. However, Pelli, Burns, Farell and Moore-Page (2006) studied psychophysically some of the properties of the visual processes that achieve the detection and identification of letters in the presence of visual noise. They found that the successful identification of a letter is based on the detection and subsequent combination of a small number (around 7) of component features, not on the detection and identification of the letter as a whole form. Subsequently, Pelli and Tillman (2007) have shown that the recognition of a word in turn depends in large measure on letter identification. In both cases, the alternative hypothesis was that the identification was achieved by the application of a whole object (letters or words respectively) template. A whole object template would provide better performance for detection and identification in high visual noise, and so the failure to show whole object templates is significant.

These authors did not make the step of identifying the features involved. Their notion of feature based recognition and our image descriptions share a number of common properties. The manner in which they use the term feature and we use the term image primitive differs fundamentally, however. Pelli et al (2006) use the term feature to refer to a piece of the image, whereas we use the term image primitive to refer to an internal description of a variable piece of image. It may well be that our image primitives are indeed descriptions of the image features that Pelli et al refer to.

The final stage in our model is the grouping stage. That is the stage which, in our formulation, determines the consequences of white space. There have been a few studies of the role of white space in word recognition and reading, and these are all consistent with the typographical convention of using quite large amounts of white space. Rayner, Fischer, and Pollatsek (1998) found that the absence of word spaces has strong interfering effects on word identification and eye movement control. Chung (2004) has found a drop in reading speed for smaller than normal line spacing in text. Our analysis of white space is in line with these findings.

Image descriptions

The main purpose of this paper is not to explore or explain typography, and certainly not to stray into the minefield of research into word recognition and reading. Instead, the paper simply hijacks that area of research to illustrate a more general point about research into vision concerning the importance and utility of image descriptions in low-level vision.

It is necessary to distinguish between making information about the stimulus available, and making information about the stimulus explicit. In crude terms, it is one thing to show that a series of neural processing operations results in a neurone responding differently to two types of stimulus (that the information is available) – but it is quite another to show how the rest of the brain would know which neurone was carrying that information (making it explicit).

Image descriptions are not mental pictures because they make image structure explicit. It is perhaps a commonplace, but one that should be re-iterated frequently, that the outcome of visual perception is not a mental picture. This seems completely uncontroversial because a mental picture would require a perceiver, opening up an infinite regression. There certainly are picture-in-the-head stages in vision and in fact much of the detailed neurophysiological and psychophysical knowledge of vision relates to stages that are essentially of this form. Pictures-in-the-head, images, make information about stimuli available but not explicit. Image descriptions are required to do that.

Despite the fact that we have used image descriptions to achieve word recognition, they are not image recognition. This is important because the outcome of visual perception is not just a set of recognitions – assignations of stimuli into pre-defined categories. This also seems completely uncontroversial: I can see completely novel squiggles on a page without recognizing them. A description stage can be flexible to deal sensibly with novel stimuli (see Watt and Phillips, 2000 for a more detailed account of this).

Our model is able to do things like word grouping and segmentation, and perhaps even word recognition, because it moves away from the data format of images, and employs a language-like descriptive stage. The limitations on how successfully it can do it are certainly limitations in the image, but the processes involved are not image-based processes.

In its essence, the image description we have explored here is a language to describe useful patterns in images. At the same time, it can be seen that the image descriptions that we present here have the additional feature of equally being a language in which a wide range of visual tasks can be described. In that sense, image descriptions become a common interface between optical information, which is image based, and task demands which are typically propositional. In theory, this common stage could be at a range of different levels.

It is possible although computationally complex, to specify most visual tasks in terms of image and pixel operations. At this level, there is only one type of descriptor, the pixel, and it has only one property, a scalar value. Any image can be described as a list of pixels, although the length of the sentence would be enormous. Similarly, any visual task can be specified as a set of operations on a list of pixels: once again the sentence length would be enormous. Equally, it is possible to create a descriptive language that is very high level, has multiple different types of image primitive, each with its own set of properties, and multiple different types of relation between them. Such a descriptive language would have very short sentences, but would not naturally deal with novel images. The level we have chosen, which corresponds broadly to local elongated structures, seems like a compromise between ability to deal with novelty and sentence length.

Closing observation about the model

We conclude with one observation about the enterprise described here. We have described a model that is an actual software system which takes an actual gray-scale image (photograph) as its input and produces a symbolic description as its output. That output can be used to make decisions about the contents of the image: we have shown how it can be used to segment and then identify words. The model is unusual in this respect.

There is not yet enough hard evidence to fully justify many features of the model, but summarizing existing evidence is not its primary function. The model exists to explore ways in which studies of visual tasks can be modelled in concrete, computable terms from a gray-scale image to a modelled response in a visual task. The range of applications of such a type of model is large. A few examples would include: visual contrast detection; contour shape analysis; face recognition and related tasks; and scene salience computation.

References

Baddeley R.J. (1997) The correlational structure of natural images and the calibration of spatial representations. Cognitive Science, 21:351--372

Campbell F.W and Robson J.G (1968) Application of fourier analysis to the visibility of gratings. J Physiol 197, 551-566

Chung STL (2004). Reading speed benefits from increased vertical word spacing in normal peripheral vision. Optometry and Vision Science 81: 525-535

Chung S.T.L., Legge G.E. and Tjan B.S. (2002) Spatial-frequency characteristics of letter identification in central and peripheral vision. Vision Research 42, 2137-2152

Chung S.T.L. and Tjan B.S. (2007) Shift on spatial scale in identifying crowded letters. Vision Research, 47, 437-451

Claessens, P. M., & Wagemans, J. (2005). Perceptual grouping in Gabor lattices: proximity and alignment. Perception & Psychophysics, 67, 1446-1459

Epelboim J., Booth J.R. and Steinman R.M. (1994) Reading unspaced text: implications for theories of reading eye movements, Vision Research 34, pp. 1735–1766.

Geisler, W.S. (2008) Visual perception and the statistical properties of natural scenes. Annual Review of Psychology, 59, 10.1-10.26.

Georgeson, May, Freeman and Hesse, (2007) From filters to features: scale-space analysis of edge and blur coding in human vision. Journal of Vision, 7(13), 1-21

Groff P. (1975) Research in Brief: Shapes as Cues to Word Recognition. Visible Language, 9, 1, 67-71

Haber R.N. and Haber L.R. (1981) The shape of a word can specify its meaning. Reading Research Quarterly, 16, 334-345.

Haber L.R, Haber R.N, and Furlin K.R, (1983) Word length and word shape as sources of information in reading. Reading Research Quarterly, 18, 165-189

Hartley J and Burnhill P (1977) Fifty guide-lines for improving instructional text. Innovations in Education and Teaching International, 14, 65 - 73

Heeger, D.J., 1992. Normalization of cell responses in cat striate cortex. Visual Neuroscience 9, pp. 181–197.

Hubel, D.H. & Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology, 195, 215-243.

Hubel, D.H. & Wiesel, T.N. (1977). Functional architecture of the macaque monkey visual cortex. Ferrier Lecture, Proceedings of the Royal Society of London B, 198, 1-59.

Kubovy, M., Holcombe, A. O., & Wagemans, J. (1998). On the lawfulness of grouping by proximity. Cognitive Psychology, 35, 71-98

Leech G., Rayson A. and Wilson A. (2001) *Word Frequencies in Written and Spoken English: based on the British National Corpus.* Longman, London. http://www.comp.lancs.ac.uk/ucrel/bncfreq/

Legge, G.E. (2007). *Psychophysics of Reading in Normal and Low Vision* . Mahwah , NJ & London : Lawrence Erlbaum Associates.

Legge, G.E., Pelli, D.G., Rubin, G.S., & Schleske, M.M. (1985). Psychophysics of reading. I. Normal vision. Vision Research, 25, 239-252

Li W, & Gilbert CD. (2002) Global contour saliency and local colinear interactions. Journal of Neurophysiology 88, 2846-56

Majaj N.J., Pelli D.G., Kurshan P. and Palomares M., (2002) The role of spatial-frequency channels in letter identification. Vision Research 42, 1165–1184.

Marr, D. (1976). Early processing of visual information. In Philosophical Transactions of the Royal Society of London B, 275, 483-524.

Marr, D. (1982). *Vision : A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco : W.H. Freeman and Co.

Marr, D. & Hildreth, E. (1980). Theory of edge detection. In Proceedings of the Royal Society of London, volume 197 of B (pp. 441-475).

Mayall K, Humphreys, G W, Olson, A. (1997) Disruption to word or letter processing? The origins of case-mixing effects. Journal of Experimental Psychology: Learning, Memory, and Cognition. 23, 1275-1286

Morris R.K., Rayner K. and Pollatsek A. (1990) Eye guidance in reading: the role of parafoveal letter and space information, Journal of Experimental Psychology: Human Perception and Performance 16, pp. 268–281.

Morrone, M.C. & Burr, D.C. (1988) Feature detection in human vision: a phase-dependent energy model. Proc. Roy. Soc. B 235 221-245.

Oruç, I., Landy, M. S., & Pelli, D. G. (2006) Noise masking reveals channels for second-order letters. Vision Research, 46, 1493–1506

Paap KR, Newsome SL, Noel RW. (1984) Word shape's in poor shape for the race to the lexicon. J Exp Psychol: Hum Percept Perform. 10, 13-28.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006) Detecting features and identifying letters. Vision Research, 46(28), 4646-4674

Pelli, D. G., & Tillman, K. A. (2007) Parts, wholes, and context in reading: A triple dissociation. PLoS ONE 2(8): e680. http://www.plosone.org/doi/pone.0000680

Polat, U. & Sagi, D. (1993). Lateral interactions between spatial channels: Suppression and facilitation revealed by lateral masking experiments. Vision Research, 33, 993-999.

Rayner, K., Fischer, M. H., & Pollatsek, A. (1998). Unspaced text interferes with both word identification and eye movement control. *Vision Research*, 38, 1129-1144.

Simoncelli E.P, Olshausen B.A. (2001) Natural image statistics and neural representation. Annual Review of Neuroscience 24: 1193-1216

Solomon & Pelli (1994) The visual filter mediating letter identification. Nature 369, 395-397

Terry P (1976) The effects of letter degradation and letter spacing on word recognition. Journal of Verbal Learning and Verbal Behavior, 15, 577-585

Underwood G, and Bargh K, (1982) Word shape, orthographic regularity, and contextual interactions in a reading task. Cognition, 12, 197-209.

Watt, R. J. (1988). *Visual Processing : Computational, Psychological and Cognitive Research.* London : Lawrence Erlbaum Associates Ltd.

Watt RJ 1991 *Understanding vision*. Academic Press, London

Watt RJ, Morgan MJ. (1985) A theory of the primitive spatial code in human vision. Vision Research 25, 1661–1674

Watt, R. J. & Phillips, W. A. (2000) The function of dynamic grouping in vision. Trends in Cognitive Sciences 4, 447-454

Wilkins A.J., Smith J., Willison C.K. and Beare T. (2008) Stripes within words affect reading. Perception, in press.

Wilson H.R. and Bergen J.R. (1979), A four mechanism model for threshold spatial vision, Vision Research 19, 19–32.

Yen, S. & Finkel, L.H. (1998). Extraction of perceptually salient contours by striate cortical networks. Vision Research, 38, 719-741.

Appendix: computational details of model

Images are typically made to a size of 512 pixels square. Each pixel holds a gray-level value with up to 64-bit pecision, as needed. All computations were performed in Matlab©. Steps (1) and (2) are standard image processing operations. Step (3) is non-standard.

Step 1: filtering

The image is convolved with a set of filters, to produce a set of filtered images. The filters vary in orientation in steps of 30° round the clock. The basic form of a horizontal filter is the product of the second derivative of a Gaussian vertically and a Gaussian (of twice the linear dimension) along the horizontal axis. This provides two of the required phases. The other two phases are obtained from the Hilbert Transform of the filter. In the fourier domain the two filters are easily made as real and imaginary components by setting the negative vertical frequencies to zero.

Convolution between the image and the filter is achieved by multiplication in the fourier domain.

Step 2: thresholding

A simple threshold is applied to all the filtered images. All points where the absolute value of the filtered image is less than some critical value are set to zero.

Step 3: image description

Each zero-bounded region of pixels is extracted from each filtered image to make a collection of items. Each item in a collection has 3 components: an x-location (x), a y-location (y) and a filter response value (v).

The mass of the region is calculated as:

```
mass=sum(v)
```

The position of the region is given by:

```
posx= sum(x*v)/sum(v);
posy= sum(y*v)/sum(v);
```

The orientation of the region is given by:

```
orient=atan ((xycov)*2/(xvar-yvar))/2;
  where:
                xycov=meanxy-posx*posy;
                xvar=meanxx- posx * posx;
                yvar=meanyy- posy * posy;
                meanxx=sum(x*x*v)/sum(v);
                meanyy=sum(y*y*v)/sum(v);
                meanxy=sum(x*y*v)/sum(v);
```

The length of the region is given by:

```
length=sqrt(xvar*cos(orient)^2 + 2*xycovar*cos(orient)*sin(orient) + yvar*sin(orient).^2);
```