Supplementary Material

Quantifying peripheral and foveal perceived differences in natural image patches to predict visual search performance

A. E. Hughes¹,³, R. V. Southwell¹,³, I. D. Gilchrist² & D. J. Tolhurst¹

¹ Department of Physiology, Development & Neuroscience
2 University of Cambridge
3 Downing Street
4 Cambridge, UK, CB2 3EG

² School of Experimental Psychology
3 University of Bristol
4 12A Priory Road
5 Bristol, UK, BS8 1TU

³ present address: University College, London, UK, WC1E 6BT
1. The preliminary ratings experiment

In order to determine the kinds of stimulus manipulation that might make appropriate distractor patches in the key search experiment, a preliminary difference ratings experiment was performed. Five image patches were chosen as the targets (Supplementary Fig.1), and then the colour or shape of these was transformed in a number of ways to make a number of comparison stimuli for each target (see e.g. Supplementary Fig.1).

Supplementary Figure 1. A, bitmap images of five objects were used to construct variant images in a pilot ratings experiment, to try to determine which changes would be differentially visible foveally and peripherally. In the main search and subsequent rating experiment, images derived only from the “cat” and the “flower” parent images were used. B, some examples of variant image patches made from the “cat” parent image. The patches had a circular outline gradually blending into the grey background. “Colour” changes: c1 a change in hue; c2 a change in chroma/saturation; c3 a change in overall brightness. “Shape” changes: s1 is blurred; s2 has the centre of the image rotated and then blended into the image rim; s3 the whole patch is rotated; s4 the central part of the image was broken up into 9 squares which were then shuffled and blended into each other.
In a trial, the target was compared to one of the variants. Either the target or the variant (chosen at random) was presented for 833ms (100 frames) and then there was a short interval of 83ms when the screen was uniform grey apart from a fixation spot. The other stimulus of the pair was then presented for 833ms. Following a second 83ms blank interval, the first presented stimulus of the pair was shown again for 833ms. This 3-interval protocol is discussed in To, Lovell, Troscianko, & Tolhurst (2010).

Nine observers viewed 286 image pairs foveally and, in separate blocks, viewed the same 286 pairs at an eccentricity of 7.3 deg in the upper left visual field (Supplementary Fig.2). The order of stimulus presentation was differently randomised for the foveal and peripheral stimuli, and differently for each observer. Three foveal blocks of about 95 stimuli were alternated with 3 peripheral blocks, with some observers starting with a foveal block and others with a peripheral block.

Supplementary Figure 2. The fixation spot was in the centre of the CRT display. A, for peripheral viewing of the image pairs, the stimulus patches were presented 7.3 deg in the upper left field, while the observer maintained fixation. B, For foveal viewing, the fixation spot disappeared during the actual stimulus presentation (833 ms) and the stimulus patch was centred where the fixation spot had been.

In both foveal and peripheral experiments, the standard stimulus pair with a magnitude difference deemed to be “20” (Supplementary Fig.3) was viewed foveally and was presented as a reminder every 10 trials. The methodology is described in detail in To, Gilchrist, Troscianko, & Tolhurst, (2011). Observers were instructed that, if they could perceive no differences between an image pair they should give a rating of zero. However, if they could perceive some difference, they must decide quickly whether, e.g., the difference was bigger or smaller than the colour saturation difference in the standard. And, finally, they must assign a numerical rating in proportion to how many times bigger or smaller they perceived the stimulus pair to be, compared to the standard. Observers were made aware that some of the image pairs would have no difference; this was to encourage them to give a rating of zero if they perceived no difference and not to suppose that there must have been a difference which they had missed.
The foveal ratings of each observer were standardised by dividing by that observer’s median foveal rating, and then multiplying by the grand median of all foveal ratings across all observers. Then, for each foveal stimulus pair, the standardised ratings of the 9 observers were averaged. The peripheral ratings were similarly standardised to the grand median of peripheral ratings and averaged.

**Supplementary Figure 3.** The standard image pair against which the observers were asked to scale their ratings of the differences between the patches. While the standard pair differs in the degree of colour saturation, the test pairs could differ in other aspects of colour, or in shape (e.g. main text Fig.1). The standard patches were the same size as the test patches, and they were viewed foveally, even when the test stimuli were presented peripherally.

Supplementary Fig.4 plots the averages of the 9 observers’ standardised ratings for the peripheral viewing of each stimulus pair against the average of their standardised ratings for foveal viewing of the same pairs. Part A shows that the perceived magnitude of colour changes in stimuli is little affected by peripheral viewing, while shape changes are perceived with less magnitude in periphery than in fovea, confirming To et al (2011). The coloured rings in Supplementary Fig.4B show the parts of stimulus space that contained perceived stimulus differences appropriate to the hypotheses underlying our search task. “High” stimuli would contain distractors that were easy to distinguish from the target both foveally and peripherally. “Low” stimuli would contain distractors that were difficult to distinguish both foveally and peripherally. The key distractor stimuli in our search experiment are the “metamer” stimuli (after Freeman & Simoncelli, 2011); they would be distractors easy to distinguish foveally, but difficult to distinguish peripherally. Rosenholtz, Huang, Raj, Balas, & Ilie (2012) might call them “mongrels”. According to the Duncan & Humphreys (1989) explanation of search performance, search time will depend on how easy it is for the observer to distinguish the target from the distractors. If the distractors come from the metamer part of stimulus space, will search be fast (because the distractors are easy to distinguish foveally) or slow (because they are hard to distinguish peripherally)?
Supplementary Figure 4. Results of the preliminary rating experiment. The peripheral rating for each stimulus pair is plotted against the foveal rating for the same pair. The values are the averages of the ratings from 9 observers. A, The red symbols and regression are for stimulus pairs where the image change was in some aspect of “colour” (luminance, hue or saturation/chroma), while the blue symbols and regression are for image pairs where the transformation was in some aspect of shape. Black symbols show ratings for image pairs where there was, in fact, no change. B, The same results are plotted, but the coloured circles approximately show the ranges of comparative foveal and peripheral ratings that would be suitable for the choices of distractor stimuli in the search experiment. “High”, “low” and “metamer” stimuli are defined in the Supplementary text.

2. Construction of the search stimuli

The preliminary ratings experiment was performed with image patch pairs constructed from 5 different originals (main text Fig.1). However, only the rating results for the “cat” and the “flower” families showed a good split into the 3 categories (Supplementary Fig.4B) that we needed for the search experiment. For each of the two target images (the undistorted “cat” and the undistorted “flower” patches), 7 distractor patches derived from the same cat or flower original were required from each of the 3 categories on the preliminary ratings graph (i.e. the easy/easy, hard/hard and metamer areas). However, the preliminary experiment had not included enough image transformations to give 7 distractors in some particular categories, and so some new transformed images had to be generated from the original cat and flower images, based on the known ratings for the image transformations that we had actually created and tested. For example, a hue change of +10 degrees was assumed to look equally different from the parent as a hue change of -10 degrees. We would later confirm that the new transformations did largely produce the foveal and peripheral ratings that we expected in the ratings experiment described in detail in the main text (main text Fig.6).

As in the rating experiment, the target patch and the distractors were set in circular patches of 68-pixel diameter with a gradual fade into the background, which was again 800 x 600 pixels of uniform

Hughes et al, Supplementary Material, page 5
grey. Image patches were placed only within the central 600 x 600 pixel square area of the 800 x 600 rectangular background. The 600-pixel square was split into four equal quadrants, delineated with thin visible grey lines (main text Fig.3; Supplementary Fig.6), so that the observer could respond in the search task by identifying the target’s location to one of the four quadrants.

**Placement of the target patch and distractor patches**

Image patches were placed quasi-randomly on the background, to ensure an even distribution of patches between the four quadrants. If placement had been done truly randomly, search arrays with empty or sparse quadrants might have been generated, especially for the smaller set sizes. Such arrays would be unintentionally easier, as the observer would be able to discount these empty quadrants as target locations from the outset.

For the purposes of search array construction, each of the 4 visible quadrants was considered to be further split into four sub-quadrants, giving 16 sub-quadrants (of 150 x 150 pixels) in total (Supplementary Fig.5). Each image patch (target or distractor) was placed into a unique 150-pixel sub-quadrant, with no more than one patch per sub-quadrant. The position of the 68-pixel diameter patch within the 150-pixel sub-quadrant was randomised, so that the final search array did not look as if it had been generated on a strict grid.

**First, the target was randomly assigned** to one of the 4 major visible quadrants, and then it was assigned at random to one of the 4 subquadrants (Supplementary Fig.5A).

**Then, the first distractor was placed.** One of the 4 major quadrants was chosen at random. If this already contained the target, the distractors were assigned sequentially to the next quadrants in cyclical numerical sequence. E.g. if the target and first selected distractor quadrant were both 2, the first 3 distractors were placed in quadrants 3, 4, and 1 to place one patch in each quadrant. If, however, the chosen quadrant did not contain the target (Supplementary Fig.5B), starting at that quadrant, 3 distractors were placed within the next three quadrants in numerical order, but skipping the one with the target. E.g. if the target was in quadrant 4 and first selected distractor quadrant was 2, the distractors were placed in quadrants 2, 3, and 1. All major quadrants now had one patch (Supplementary Fig.5C).
Supplementary Figure 5. Construction of search stimuli by quasi-randomly assigning patch locations within a grid. The display had one vertical and one horizontal line that were visible. These divided the display area into 4 major quadrants. For stimulus construction, each major quadrant was considered to be made of 4 subquadrants. The extra gridlines and numbers are for illustration, and were not visible to the observers. A, in this example, the target was allocated at random to subquadrant 3 major quadrant 4 (red numbers). Within the subquadrant, the centre of the patch was randomly allocated such that its borders remained within the subquadrant. B, a quadrant was chosen at random for the first distractor – quadrant 1. Since, in this case, the quadrant was not occupied by the target, a distractor patch was assigned at random to one of its subquadrants (#4). C, The next 2 distractors were assigned to the quadrants in numerical order (#2 and #3), starting at the one after the first-placed distractor and skipping the quadrant with the target (no need to skip in this example). Each quadrant now had one patch. D, returning to the quadrant initially chosen for the first distractor (#1), the fifth patch was added to the array. Subsequent patches were placed in numerical order, cycling through the numerical code for the quadrants (2, 3, 4, 1, 2), and the result is shown as Supplementary Fig.6A.

Finally, the remaining distractors were placed sequentially in the 4 quadrants, starting with the quadrant that had originally been chosen to take the first distractor (this might have been the one containing the target; Supplementary Fig.6B). This ensured as even a distribution of image patches between the four quadrants as possible, and also that the quadrant containing the target sometimes contained more patches than some other quadrants and sometimes less (Supplementary Fig.6). Within each quadrant, the next image patch was placed into a sub-quadrant chosen at random from the remaining unoccupied sub-quadrants. The end result was an even distribution of image patches, with no quadrants being left empty or with substantially more patches than others.

Hughes et al, Supplementary Material, page 7
Families of search stimuli

For the 2 target stimuli ("cat" and "flower") there were search arrays with homogeneous and with heterogeneous distractors, at each of the 3 classes of target/distractor difference (easy/easy, hard/hard and metamer). For each of these 12 categories, we had a pool of 7 different distractor patches to choose from. Search arrays were made with 3 sizes (4, 9 and 14 distractor patches), giving 36 conditions altogether and, for each of those conditions, 10 different instances of search array were generated. For each condition of array, the 10 instances were generated with the target in a range of different quadrants, subquadrants and positions.

To make search arrays with homogeneous distractors, 2 of the 7 distractor patches were chosen to make the search stimuli. For a given set of 10 search arrays, 5 arrays were made with one of these distractors and 5 with the other. With a set size of 5 (4 distractors), 3 distractors were simply assigned to the empty major quadrants after the target had been randomly placed. The fourth (final) distractor was essentially assigned randomly to one of the 4 major quadrants. With a set size of 10 (9 distractors), 2 of the major quadrants ended with two patches and 2 quadrants had 3. The target might have been in a quadrant with 2 or 3 patches. With a set size of 15 (14 distractors), the procedure essentially meant that one quadrant was chosen at random to house only 3 patches, whilst the others might have housed 4. The target could have been in a group of 3 or 4 patches.

To make search arrays with heterogeneous distractors, the same basic sequence was followed, except that, for any one search array, distractor patches had to be selected from a population of 7 rather than just 1. With a set size of 5, 4 of the 7 distractor patches were chosen differently at random (without replacement) for each instance of the search array, and the order of their selection for placement in the major quadrants was randomised. With a set size of 10, all 7 heterogenous distractors were taken once but 2 of them were chosen to be duplicated, differently at random for each search array. The order of the 9 distractors was randomised before assignment to the major quadrants. Finally, for the set size of 15, each of the 7 distractors was used twice, the order of the 14 being randomised before placement in the major quadrants.
Supplementary Figure 6. Two examples of search arrays from one set of 10. A is the completed sequence outlined in Supplementary Fig.5. The patches are all made from the original “cat”, which is the target in bottom-right in A and in top right in B. The array family has heterogeneous distractors, size 10, easy/easy. Note that: (i) the target patch is in a different major quadrant in the two instances; (ii) the target is in the same quadrant as 2 distractors in B, but with only 1 in A; (iii) the quadrants with 2 patches are side-by-side in one example, but diagonally located in the other; (iv) the duplicated distractors are different in the two instances. In example B, the target was first assigned to quadrant 2. The first distractor was also randomly assigned to quadrant #2 but, because that was already occupied by the target, the first distractor was actually placed in quadrant #3 and subsequent distractors were place in quadrants 4, 1, 2, 3, 4, 1, 2 and 3.
3. Additional details of Results

3A. Errors during search

We recorded search time (time between array onset and the observer’s button push) and whether the observer had correctly identified the screen quadrant containing the target. In the 1800 total trials, there were 77 errors (4.3%) where the observers had chosen the wrong quadrant. Most (69) of the errors were evoked by arrays in the “low” discriminability condition, especially for the heterogeneous arrays. The feature of the “low” condition distinguishing it from the other two conditions is in having low TD and DD discriminability foveally.

<table>
<thead>
<tr>
<th>Search array condition</th>
<th>Cat family</th>
<th>Flower family</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size = 5</td>
<td>Size = 10</td>
</tr>
<tr>
<td>homogeneous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>1 (1,0)</td>
<td>2 (1,0)</td>
</tr>
<tr>
<td>metameter</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>high</td>
<td>0</td>
<td>2 (0,1)</td>
</tr>
<tr>
<td>heterogeneous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>3 (2,1)</td>
<td>12 (6,4)</td>
</tr>
<tr>
<td>metameter</td>
<td>0</td>
<td>2 (1,0)</td>
</tr>
<tr>
<td>high</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Supplementary Table 1.** The number of errors obtained for each search array condition. An error occurred when an observer incorrectly chose one of the 3 quadrants where the target was not actually present. The error trials were discarded from our analysis of search times. For each of the 36 conditions, there were 10 different search arrays and 5 observers participated in the experiment. Thus, the numbers of errors listed in each cell in the Table is out of 50 trials. All but 8 of the 77 errors were made for “low” discriminability search stimuli. Most of the 77 errors were made by 2 of the 5 observers (30 and 21 respectively); their individual error scores are shown in brackets.

3B. Ratings for stimuli that differ in brightness

In our previous studies (To, Gilchrist, Troscianko, Kho, & Tolhurst, 2009; To et al., 2011), we typically found the perceived differences of most image pairs to be reduced in peripheral vision compared to foveal (see also the preliminary ratings experiment, detailed in the main paper). However, Fig.6 of the main text (especially part B) shows that many image pairs evoked higher perceived difference ratings peripherally (the line of equality is shown). One reason for this difference results from a difference in the specific image changes that we used to make the present search arrays. Supplementary Fig. 7 replots the 12 degree data of Fig.6B from the main paper to show the kinds of image change employed. The grey symbols show the results for image pairs which differed in some spatial transformation without any colour changes. These stimuli did evoked substantially lower magnitude ratings peripherally, as we have shown previously. The brown symbols are for image pairs where the changes were in hue and/or chroma (saturation); they lie close to the line of

Hughes et al, Supplementary Material, page 10
equality, not dissimilar from our previous findings. The remaining image pairs (orange symbols) were ones where the change included a change in overall brightness (“L” in L*c*h space); we have not previously employed such stimuli.

**Supplementary Figure 7.** The ratings for 12 degrees eccentricity are plotted against foveal rating; the ratings are averaged across the observers. The orange symbols are for image pairs where the difference included a change in overall brightness. The brown symbols are for pairs where the difference included any other colour change that did not include brightness (i.e. hue and/or chroma). The grey symbols are for image pairs, where the changes were in shape only and not in any aspect of colour. Black symbols are those stimuli where there was in fact no change.

Peripheral viewing has increased the rating given to the brightness stimuli; most orange data points lie above or close to the line of equality. A General Linear Model showed that the regression of peripheral rating upon foveal rating was significantly improved if the brightness-change results were allowed a different intercept and slope from the rest of the colour-change stimuli which in turn were allowed a different intercept and slope from all the shape-change stimuli; an ANOVA compared nested models with and without the 3-way brightness/colour/shape stimulus category as a fixed factor governing separate slopes and intercepts ($F(1,126) = 84.53, P \approx 0$). Using the regression through the 23 brightness change data (orange circles) as the reference, the slope of the regression through the other colour change stimuli ($n = 41$; brown circles) was slightly shallower (76.5% of reference slope, $t = 2.08, P = 0.04$); the line through the shape change data (grey) was very much shallower (39.4% of reference slope, $t = 5.63, P \approx 0$), in concordance with the results from To et al (2011). The rating data were “centred” so that the intercepts given by the GLM showed the expected peripheral rating for stimuli evoking a foveal rating of 25. This “centresep” for the brightness change data was 7.95 ($t = 6.64, P \approx 0$). The centresep for the other colour change stimuli was significantly less ($\Delta = -6.13, t = 4.08, P \approx 10^{-5}$) so that it was only just greater than zero (1.82) Paired t-tests confirmed that the brightness-change ratings were indeed significantly higher peripherally.
than foveally ($\Delta$rating = 7.79, $t = 6.03$, $n = 23$, $P \approx 0$) but that the other colour-change data were, on average, about the same foveally as peripherally ($\Delta$rating = 0.73, $t = 0.56$, $n = 41$, $P = 0.58$).

This finding is particularly surprising given that the standard assumption is that visual information is degraded in the periphery (Strasburger, Rentschler, & Jüttner, 2011) and thus stimuli could be expected to look more similar. This effect could reflect a type of internal scale adjustment, whereby subjects feel compelled to give a range of ratings for both foveally and peripherally presented stimuli. As metameric stimuli in the periphery end up with low difference values, subjects may therefore artificially “boost” the ratings of other types of stimuli to increase the range of values given for peripheral stimuli, although it is not clear why this should happen only with stimulus pairs that changed in brightness. Alternatively, there is evidence that stimuli presented in the periphery can in fact be perceived as brighter than those in the fovea when the observer is dark adapted (Marks, 1966; Osaka, 1975), although this finding has not always been replicated for photopic vision (Zihl, Lissy, & Pöppel, 1980). There is currently little understanding of why this might be the case (Strasburger et al., 2011). It may be that, if receptive fields and summation areas are larger in the periphery, then the peripheral image patches are more closely matched in size to peripheral receptive fields than the foveal ones, allowing detection of an overall patch average brightness rather than local edge contrasts. As our results suggest that this effect generalises to experiments using naturalistic stimuli, future experiments could further investigate the role of brightness in peripheral perception and the underlying mechanisms.

3C. Peripheral ratings for stimulus patches that did not actually differ

The black symbols in Supplementary Fig. 7 show the ratings given for those image pairs where there was, in fact, no difference; ideally, the observers should have given a rating of zero to these stimuli, and the observers were told that some stimuli would indeed have no difference. However, it can be seen that the averaged ratings of the 11 observers for these 23 image pairs are not zero, but more interestingly they tend to be greater peripherally than foveally, particularly at 12 degrees. Retrospectively, this can be seen in the graphs of (To et al., 2011). Supplementary Fig.8 investigates this in more detail. It shows histograms of the individual ratings of the observers to the 23 identity pairs (i.e. the ratings of the 11 observers are not averaged together). The identity pairs evoked fewer ratings of zero at 12 degrees eccentricity (Supplementary Fig.8B) than foveally (Supplementary Fig.8A), and the overall averaged rating is lower foveally.
Supplementary Figure 8. The ratings given to image pairs that, in fact, had no change and might have been expected to elicit a rating of zero. 11 observers each saw 23 such image pairs at each eccentricity, giving 253 ratings foveally (above) and 253 for 12 degrees peripherally (below). The bar charts show how many times (out of 253) different rating values were assigned.

The foveal and peripheral rating distributions for identity stimuli are significantly different in form. The mean rating foveally was 0.85 while the mean peripheral rating was 4.06. Supplementary Table 2 lists the output statistic for a number of tests for comparing 2 samples. All are highly significant, at $P \approx 0$.

One explanation for this phenomenon is that visual perception in the periphery may be more subject to the effects of internal noise; perhaps, this is the spatial uncertainty of specific features leading to the hypothesis that peripheral objects are coded as a set of summary features (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Levi, 2008; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Pelli & Tillman, 2008; Rosenholtz et al., 2012). Subjects do in fact believe that they perceive differences between the two stimuli, despite this not being physically true.

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>Present data from main text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired t-test</td>
<td>$t = 6.65$</td>
</tr>
<tr>
<td>Mann-Whitney U test</td>
<td>$z = 8.64$</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov 2 sample</td>
<td>$D = 0.34$</td>
</tr>
</tbody>
</table>

Supplementary Table 2. The summary statistics of 3 tests of the similarity between the distributions of ratings for non-changing stimuli, as rated both foveally and peripherally.
In the main text, Fig. 7 summarises our attempts to use 12 peripheral ratings of perceived difference to model search slope and search time. Here we reproduce the Figure (Supplementary Figure 9) but we use different symbols to highlight the results for the “low” (green), “high” (red) and “metamer” (blue) arrays. The different symbol shapes distinguish cat from flower and homogeneous from heterogeneous.

**Supplementary Figure 9.** This Figure is reproduced from the main text with different symbols to show the “high”, “low” and “metamer” classes of array. Low discriminability amongst constituent patches is green; high discriminability is red; metamer discriminability is blue. A, following Rosenholtz et al. (2012), search slope (n = 12) is plotted against the discriminability of the target from the distractors. TD discriminability is the averaged TD rating value for the array at 12 degrees eccentricity. Circles for “cat” homogeneous stimuli; squares for “cat” heterogeneous stimuli; triangles for “flower” homogeneous stimuli; inverted triangles for “flower” heterogeneous stimuli. The black line is the regression through all 12 data. B, the experimentally measured search slope (n = 12) is plotted against the slope predicted by a multilinear regression on TD and DD at 12 degrees, Eqn.2. Symbols as in A; the black line is the line of equality. C, the actual search time (n = 36) is plotted against the time predicted by the nonlinear fit to Eqn.4. Symbols as in A.
A. Relationship between actual search slope (msec/item) and target-distractor discriminability at 12 deg.

B. Comparison of actual search slope (msec/item) with the predicted search slope.

C. Correlation between measured search time (sec) and predicted search time (sec).
References


Hughes et al, Supplementary Material, page 16
