What image features guide lightness perception?

Minjung Kim
Fachgruppe Modellierung Kognitiver Prozesse, Technische Universität Berlin, Berlin, Germany
Department of Psychology and Centre for Vision Research, York University, Toronto, ON, Canada

Jason M. Gold
Department of Psychological and Brain Sciences, Indiana University Bloomington, Bloomington, IN, USA

Richard F. Murray
Department of Psychology and Centre for Vision Research, York University, Toronto, ON, Canada

Lightness constancy is the ability to perceive black and white surface colors under a wide range of lighting conditions. This fundamental visual ability is not well understood, and current theories differ greatly on what image features are important for lightness perception. Here we measured classification images for human observers and four models of lightness perception to determine which image regions influenced lightness judgments. The models were a high-pass-filter model, an oriented difference-of-Gaussians model, an anchoring model, and an atmospheric-link-function model. Human and model observers viewed three variants of the argyle illusion (Adelson, 1993) and judged which of two test patches appeared lighter. Classification images showed that human lightness judgments were based on local, anisotropic stimulus regions that were bounded by regions of uniform lighting. The atmospheric-link-function and anchoring models predicted the lightness illusion perceived by human observers, but the high-pass-filter and oriented-difference-of-Gaussians models did not. Furthermore, all four models produced classification images that were qualitatively different from those of human observers, meaning that the model lightness judgments were guided by different image regions than human lightness judgments. These experiments provide a new test of models of lightness perception, and show that human observers' lightness computations can be highly local, as in low-level models, and nevertheless depend strongly on lighting boundaries, as suggested by midlevel models.

Introduction

Lightness constancy is the remarkable ability of the human visual system to maintain a stable percept of surface reflectance across a wide range of lighting conditions. Reflectance is the proportion of incident light reflected by a surface, as measured in photometric units, and lightness is perceived reflectance. The interactions between lighting, material properties, and 3-D shape during image formation mean that recovering surface reflectance from image luminance is an underdetermined problem: Under different lighting conditions, surface patches with the same reflectance can yield different luminances in the retinal image, and surface patches with different reflectances can yield the same luminance. How the human visual system achieves lightness constancy remains poorly understood, and research on this problem is a fundamental topic in vision science.

Different theories of lightness perception identify different image features as playing important roles in estimating lightness. Anchoring theory (Gilchrist et al., 1999) states that the image patch with the highest luminance is a crucial reference point, and that other image regions are assigned lightness values relative to this region. Center–surround models (Heinemann & Chase, 1995; Shapiro & Lu, 2011) emphasize the role of the immediate surround of the region whose lightness is being judged. The oriented difference-of-Gaussians (ODOG; Blakeslee & McCourt, 1999) model, and its extensions LODOG and FLODOG (Robinson, Hammon, & de Sa, 2007), rely on oriented receptive fields at multiple scales. Adelson (1993) emphasizes the importance of perceptual segmentation, highlighting X-junctions as a possible cue to lighting boundaries (Beck, Prazdny, & Ivry, 1984).

Most experiments on lightness perception have examined human observers' lightness matches in scenes that were carefully designed so that different models predicted different lightness percepts. Here we take the

Citation: Kim, M., Gold, J. M., & Murray, R. F. (2018). What image features guide lightness perception? *Journal of Vision*, *18*(13):1, 1–20, https://doi.org/10.1167/18.13.1.

https://doi.org/10.1167/18.13.1 Received June 26, 2018; published December 4, 2018 ISSN 1534-7362

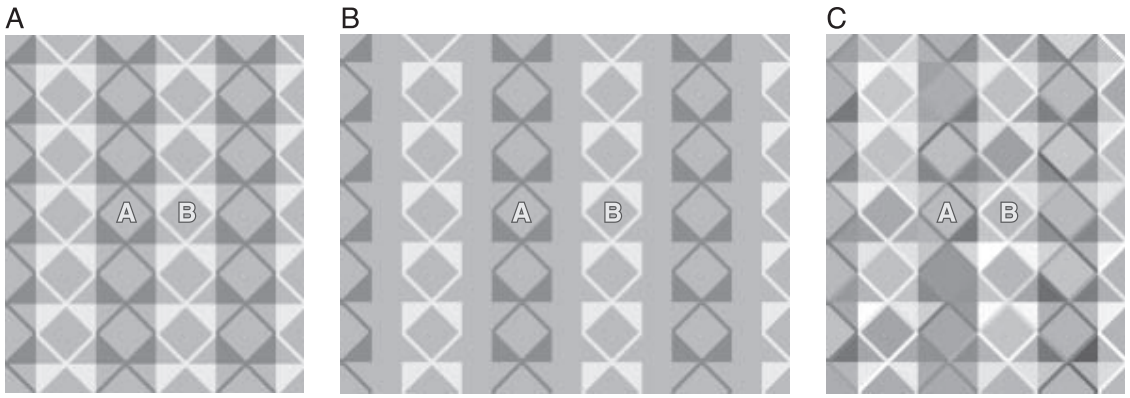

Figure 1. Stimuli in Experiment 1. (A) Standard argyle. (B) Broken argyle. (C) Noisy argyle. These stimuli are modeled after those of Adelson (1993). We measured points of subjective equality for all three stimulus types, and classification images for the noisy argyle. The letters A and B were not shown in the stimuli used in the experiment.

novel approach of measuring classification images to evaluate models of lightness perception (Ahumada, 2002; Murray, 2011; Volterra, 1930). Classification images measure the influence that local stimulus regions have on an observer's responses in a perceptual task, and so they provide information about what image features guide perceptual judgments. They are a psychophysical version of methods that are called *reverse correlation* or *spike-triggered averaging* in the neurophysiological literature (Ringach & Shapley, 2004). While most often used to study spatial vision, classification images provide a flexible experimental tool for identifying important features in a variety of domains, such as the perception of illusory contours (Gold, Murray, Bennett, & Sekuler, 2000), facial expressions (Kontsevich & Tyler, 2004), and translucency (Nagai et al., 2013). Different theories of lightness perception identify different image features as being crucial to computing lightness percepts, so classification images should provide a powerful way of testing these theories. In the domain of brightness perception, classification images have been used to examine simultaneous contrast effects (Shimozaki, Eckstein, & Abbey, 2005). Here we use classification images to examine more complex stimuli where lightness percepts may depend on scene structures such as lighting boundaries.

We use the argyle illusion (Adelson, 1993; Figure 1A) as a test case for evaluating models of lightness perception. In this illusion, one region (Figure 1A, diamond A) appears lighter than another (diamond B) even though they are of the same physical luminance. We chose the argyle illusion as our "fruit fly" because it is one of the strongest known lightness illusions, and one which has consistently resisted explanations by low-level models (e.g., Blakeslee & McCourt, 2012). Therefore, it poses a difficult and interesting problem for modeling visual perception, and understanding it may reveal general principles of lightness constancy.

Adelson (1993) explains the argyle illusion in terms of zones of uniform lighting—or, in the terminology of Gilchrist et al. (1999), *lighting frameworks*. In Figure 1A, diamond A appears to belong to a dimmer lighting framework than B, yet it has the same luminance; Adelson suggests that, from this, the visual system infers that A has a higher reflectance than B. Adelson further suggests that lighting frameworks in the argyle illusion are determined by nonreversing X-junctions at the boundaries between light and dark columns, which create a percept of dark, vertical shadows or semitransparent filters (Beck et al., 1984). Indeed, if the X-junctions are destroyed by splitting apart the columns (Figure 1B, the *broken argyle*), observers report a much smaller lightness difference between diamonds A and B.

In the present study, we investigate how human observers perceive and process the argyle illusion, and we compare human behavior to four computational models. We compare the strength of the argyle illusion for human and model observers, and we also compare the critical image features for human and model observers, using classification images to determine what image features influence observers' lightness judgments most strongly. Our results show that human observers' lightness judgments depend strongly on the luminances in the immediate neighborhood of the test patches being judged, in a way that tracks the boundaries of local lighting frameworks. Interestingly, none of the models that we tested are able to both replicate the argyle illusion and produce classification images that are even qualitatively similar to those from human observers. These findings show that making progress with computational models of lightness perception will require a better understanding of how lighting frameworks are established and of how the luminance of elements within lighting frameworks contributes to lightness percepts.

Experiment 1

In Experiment 1, we investigated what image features contribute to the argyle illusion for human observers. We measured points of subjective equality (PSEs) for standard, broken, and noisy argyle stimuli (Figure 1), and we measured classification images for the noisy argyle stimulus. We used PSEs to gauge the strength of the illusion and to screen observers for the much longer classification-image experiment. We measured classification images to determine what parts of the image contributed most strongly to the illusion.

Methods

Observers

We recruited 11 observers from the York University Centre for Vision Research. All were unaware of the purpose of our experiment except observer RM, who is one of the authors. In all experiments reported in this article, observers reported normal or corrected-to-normal monocular visual acuity in both eyes and gave written informed consent before participating. All procedures were approved by the Office of Research Ethics at York University.

Pilot studies showed that the lightness illusion was weaker in our stimuli than in the original argyle illusion (Adelson, 1993), likely because our stimuli were lower in contrast than Adelson's (see Stimuli). We chose two screening criteria, based on pilot data, to ensure that only observers who experienced a strong lightness illusion participated in the classification-image experiment. First, the observer's PSE had to be higher in the standard argyle condition than in the broken argyle condition. Second, the lightness illusion had to be at least half as strong in the noisy condition as in the standard condition; we operationalized this by requiring that the difference between the observer's PSEs in the noisy and broken conditions be at least half as large as the difference between their PSEs in the standard and broken conditions. Of the 11 observers, seven met the screening criteria, and four of those participated in the classification-image experiment.

Stimuli

The *standard argyle* (Figure 1A) consisted of lines, triangles, and diamonds arranged in a manner similar to the argyle figure from Adelson (1993). These image patches were set to light gray, dark gray, and middle gray (Weber contrasts = 0.625, −0.453, and 0), except for the two test patches labeled A and B in Figure 1A. These contrasts are 67% of Adelson's, and this contrast reduction was necessary to allow enough dynamic range to accommodate the contrast noise in the noisy argyle condition (see later). Background luminance was 176 cd/m^2. The test diamonds subtended 0.79° horizontally and vertically, and the whole stimulus subtended 3.97° horizontally and 4.76° vertically. The letters A and B in Figure 1 were not shown in the experiment.

The *broken argyle* was the same as the standard argyle, except that we introduced vertical gaps between the light and dark vertical strips (Figure 1B). The gaps were half as wide (0.39°) as the vertical strips (0.79°). The stimulus subtended 5.95° horizontally and 4.76° vertically.

The *noisy argyle* was created by adding an independent sample of Gaussian noise ($M = 0$, $SD = 0.18$) to the contrast of every patch (i.e., diamond, triangle, and line segment) in the standard argyle except the test diamonds (Figure 1C). The noise was patch-wise rather than pixel-wise, in that all pixels belonging to a single patch were modulated by the same additive contrast noise. We used patch-wise noise instead of pixel-wise noise in order to investigate how the geometric elements of the argyle stimulus contributed to the lightness illusion, and to reduce the dimensionality of the classification image. We used a different, independent sample of noise on every trial, in both the PSE and classification-image experiments.

All stimuli were shown on the LCD screen of a 21.5-in. late-2013 iMac positioned 0.57 m from the observer. The monitor had a resolution of 1,920 × 1,200 pixels, a pixel size of 0.247 mm, and a nominal refresh rate of 60 Hz. Our software inverted the monitor's gamma function to show the required luminances.

Procedure

PSE experiment: The PSE experiment had three conditions: standard argyle, broken argyle, and noisy argyle (Figure 1). Condition order was counterbalanced across observers. Each condition ran as one 7-minute block of 240 trials without a break. Each trial began with the stimulus centered on a gray background (176 cd/m^2) for 1,000 ms. After the stimulus disappeared, the observer pressed a key to indicate which of the two test diamonds appeared lighter. There was no feedback, and the next trial began 500 ms after the observer's response. In a randomly selected half of the trials, the stimulus was mirrored left to right so that test patch A was on the right and B was on the left. We adjusted the contrasts of the test diamonds using three interleaved staircases (one-up/three-down, one-up/one-down, and three-up/one-down; Wetherill & Levitt, 1965), with a step size of 0.04 Weber-contrast unit. The staircases modified the contrasts of both test diamonds at the same time—that is, incremented the contrast of one test diamond and decremented the other.

We obtained PSEs by making a maximum-likelihood fit of the normal cumulative distribution function to the empirical psychometric function and finding the 50%

response point on the fitted function. Thus, the PSE was the contrast of the test diamonds at which the observer was equally likely to choose diamonds A and B, and indicated the strength of the lightness illusion. Confidence intervals for the PSEs were obtained using a bootstrap procedure (Efron & Tibshirani, 1993).

We instructed observers to choose the "whiter" test patch. We found that some observers described test diamond B in the standard argyle as "more intense," "more luminous," or "popping out more" than diamond A, but nonetheless agreed that A seemed "closer to white" than B.

Classification-image experiment: Each observer completed seventeen 20-minute sessions of 600 trials that showed the noisy argyle stimulus, with a short break every 100 trials, for a total of 10,200 trials per observer. Observers participated in a maximum of four sessions per day, separated by breaks of at least 20 minutes, and the sessions were spread over a period of one to seven weeks. The sequence of events on each trial was the same as in the PSE experiment. At the beginning of the experiment, the contrast of the test diamonds was set to the PSE obtained in the noisy argyle condition of the PSE experiment. The contrast on each subsequent trial was chosen using a modified QUEST procedure (Watson & Pelli, 1983) that used the 50 most recent trials to make a running maximum-likelihood estimate of the observer's PSE throughout the experiment.

We calculated the classification image for each observer by taking the average noise field over all trials where the observer chose test patch A as appearing lighter, minus the average noise field over all trials where the observer chose B (Ahumada, 2002; Murray, 2011). Some elements of the argyle stimulus were larger than others (e.g., the diamonds were larger than the triangles; see Figure 1C). In order to compensate for the greater influence of larger patches on observers' responses, we divided the value of each classification-image patch by the number of pixels in that patch; see Appendix A for a proof that this is the appropriate correction for patch size under a linear observer model.

To determine the statistical significance of classification-image elements, we used a two-tailed z test. For classification images from individual observers, we controlled the family-wise error rate using a permutation procedure (Fisher, 1966; Nichols & Holmes, 2001). For the average classification images across observers, we controlled the family-wise error rate using Bonferroni correction. See Appendix B for details of these statistical analyses.

Results and discussion

Figure 2 shows PSEs for all observers. The top row shows the four observers who met the screening criteria (see Methods) and participated in the classification-image experiment. The middle row shows the three observers who met the screening criteria but did not participate in the classification-image experiment. The bottom row shows the four observers who did not meet the screening criteria. The height of each bar shows the amount by which the test diamond contrasts had to be adjusted in order for the observer to choose the two diamonds equally often. It is thus a measure of the strength of the lightness illusion.

The lightness illusion was about twice as strong in the standard argyle as in the broken argyle for the observers who passed screening ($M_s = 18.4$, $SD_s = 5.4$; $M_b = 9.2$, $SD_b = 4.0$; paired, two tailed, $t(6) = 6.3$, $p < 0.001$). For comparison, Adelson (1993) reports that the illusion was four times as strong in the standard argyle. It is not surprising that our stimuli generated a weaker illusion than Adelson's, since we used lower contrast to leave room for contrast noise in the noisy argyle condition.

For the four observers selected for the classification-image experiment, the illusion strength in the noisy argyle was not significantly different from in the standard argyle; adding noise to the argyle stimulus did not substantially weaken the lightness illusion ($M_s = 19.7$, $SD_s = 2.4$; $M_n = 18.6$, $SD_n = 1.3$; paired, two tailed, $t(3) = 1.1$, $p = 0.359$).

Figure 3 shows classification images for individual observers, as well as the average across all observers. The classification images were largely consistent with each other, indicating a common strategy across observers. The polarity of a patch (white or black) in a classification image shows how the noise fluctuations at that stimulus location were correlated with the observer choosing test patch A. A white patch in the classification image means that positive-contrast noise at that location made the observer more likely to choose A and negative-contrast noise made the observer less likely to choose A. A black patch means that the effects were in the opposite direction.

The classification images reveal a contrast-like effect in observers' lightness judgments. Test diamond A is surrounded by dark patches in the classification image, and B is surrounded by light patches. This means that observers were more likely to choose a test diamond as appearing lighter when it was surrounded by darker-than-usual elements.

This contrast-like effect had interesting spatial structure. First, it was localized: Lines, triangles, and diamonds neighboring the test diamonds had the strongest effect on observers' lightness judgments, whereas more distant stimulus elements had little or no effect. This is surprising, since illusions like the argyle illusion are often explained in terms of midlevel factors such as shadows and transparency, and so we might have expected an effect of distant but ecologically relevant stimulus elements. Particularly in the area-

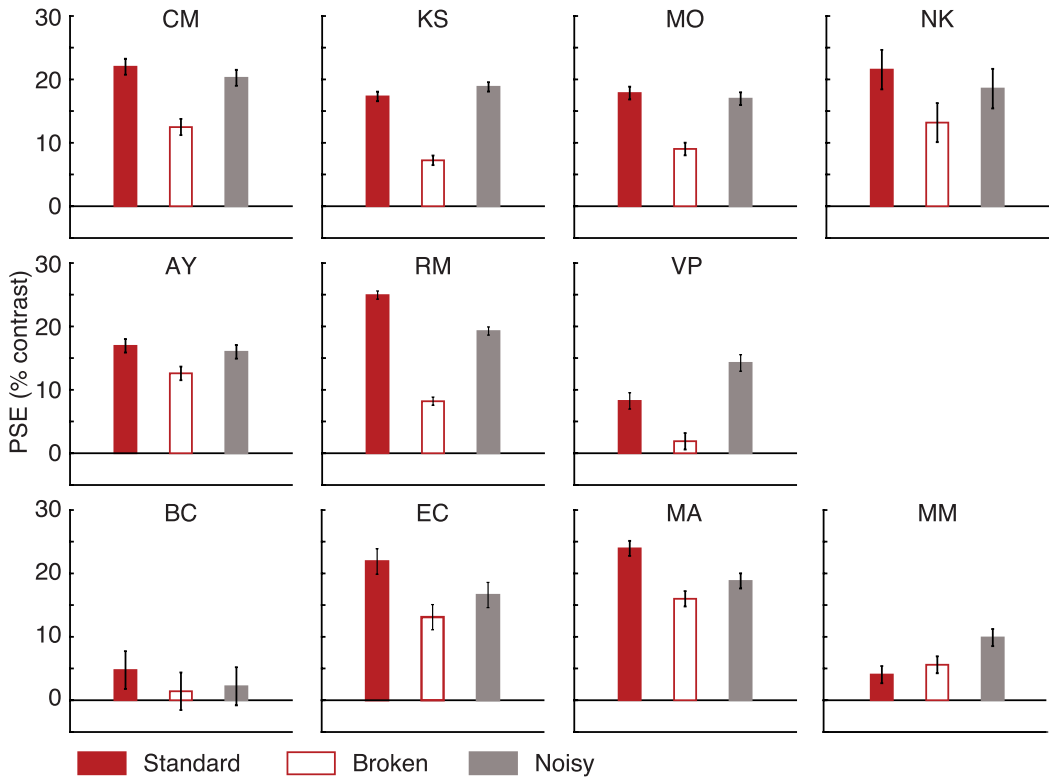

Figure 2. Points of subjective equality in Experiment 1. Bar height measures illusion strength. The top row shows observers who participated in the classification-image experiment. The middle row shows observers who passed screening but did not participate in the classification-image experiment. The bottom row shows observers who did not pass screening. Error bars are bootstrapped 95% confidence intervals (Efron & Tibshirani, 1993).

corrected classification images (Figure 3, middle row), it is striking how quickly influence falls off with distance. This is unlike other variants of the argyle illusion where there are long-range influences on perceived lightness (Flynn & Shapiro, 2014).

Second, the contrast-like effect was anisotropic: Neighboring diamond patches above and below the test diamonds were influential, whereas those to the left and right were not, even though they were the same distance away. The direction of the anisotropy suggests that the visual system is sensitive to the vertical structure of the argyle stimulus and recognizes the top and bottom diamonds as belonging to the same group as the test diamonds. These vertically displaced diamonds had the same polarity in the classification images as the lines and triangles neighboring the test patches, indicating that the diamonds also had a contrast-like effect on lightness. Our interpretation is that the grouping indicates that the visual system is sensitive to lighting frameworks, though gestaltlike grouping processes could also explain the data (e.g., grouping by similarity or by proximity).

A final remark is on using patch-wise noise rather than pixel-wise noise: Using patch-wise noise reduced the dimensionality of the classification image, but the resulting classification image does not reveal how different pixels within a patch contribute to the lightness judgments. Using pixel-wise noise would likely have shown a smooth falloff in the influence of each pixel as a function of distance from the test patch, but further experiments would be needed to confirm this hypothesis.

In the Modeling section, we examine classification images from four current models of lightness perception to see how well they account for the local, anisotropic, contrast-like effects we found in human classification images.

Experiment 2

In Experiment 1, the classification images were anisotropic, in that diamonds above and below the test diamonds significantly influenced observers' judgments but those to the left and right did not. Was this anisotropy due to the vertical lighting frameworks in the stimulus, or did it reflect a more general bias toward the vertical? To investigate, we ran a variant of Experiment 1 where we rotated the argyle stimuli clockwise by 90°. If the anisotropy was caused by a general bias toward the vertical, then even with rotated

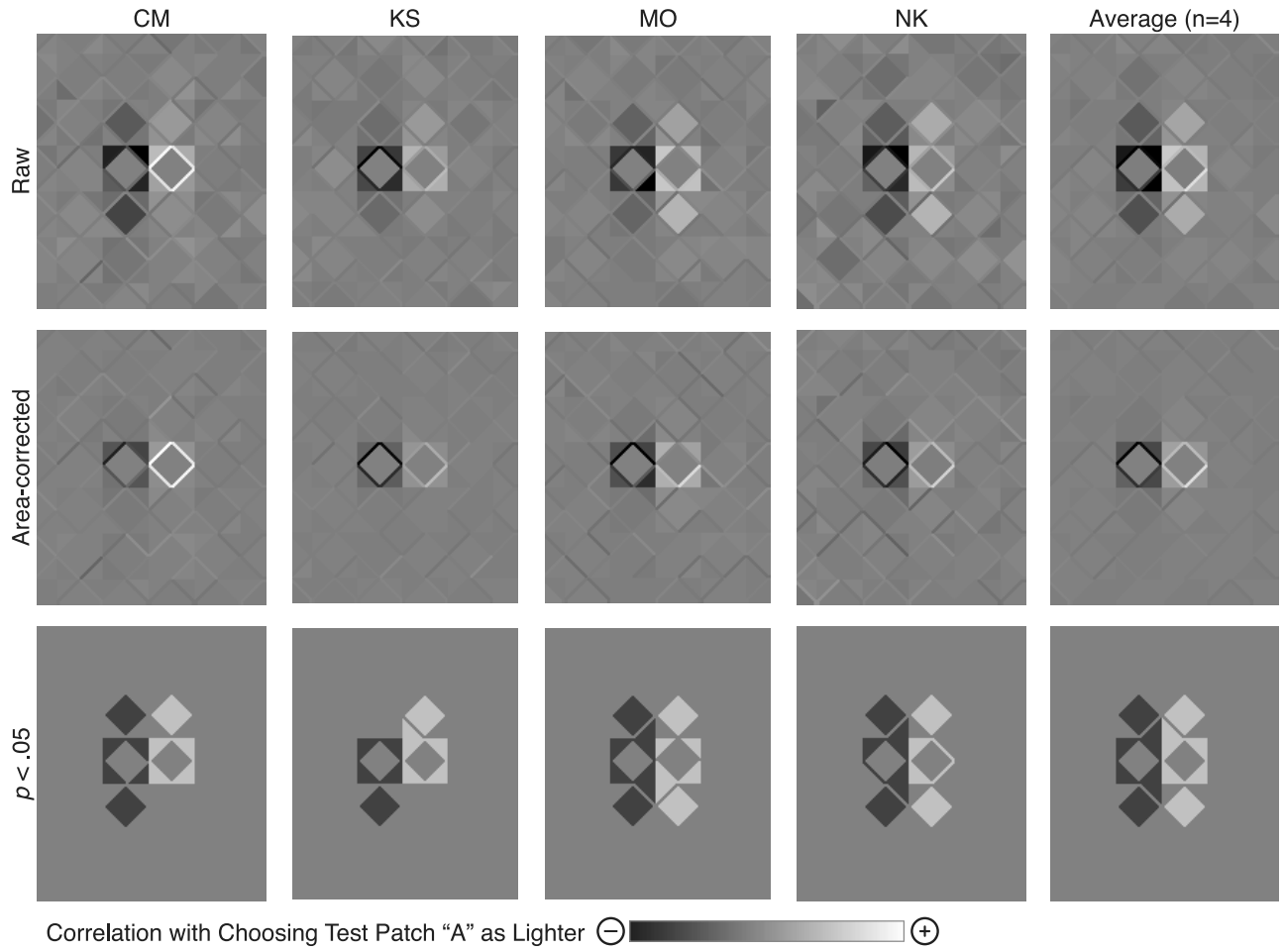

Figure 3. Classification images in Experiment 1. The polarity of each patch shows the direction of the influence of the corresponding stimulus patch on the probability that the observer chose test patch A as appearing lighter. The intensity of each patch shows the strength of the influence. Each observer's classification image has been scaled to fill the available black/white range, so values cannot be compared across observers in these images. The top row shows raw classification images. The middle row shows classification images where each image patch has been divided by the number of pixels it contains, to correct for patch size. The bottom row shows which image patches were statistically significant elements at $\alpha_{fw} = 0.05$. The family-wise error rate was controlled using a permutation procedure for the individual observers (Nichols & Holmes, 2001) and Bonferroni correction for the average across observers.

stimuli the classification images should show influences in the vertical direction and not the horizontal.

Methods

Observers

We recruited six observers from York University. None had participated in Experiment 1. All were unaware of the purpose of the experiment. We used the same PSE screening criteria as in Experiment 1 to choose participants for the classification-image experiment. Four observers met the screening criteria, and two of those participated in the classification-image experiment. Due to an error in the screening process, the two observers who did not meet the screening criteria also participated in the classification-image experiment, and we separately report their classification images here as well.

Stimuli

The stimuli were the same as in Experiment 1, except that they were rotated clockwise by 90°.

Procedure

Procedures were the same as in Experiment 1.

Results and discussion

Figure 4 shows PSEs for all observers. The top row shows PSEs for observers who passed the PSE

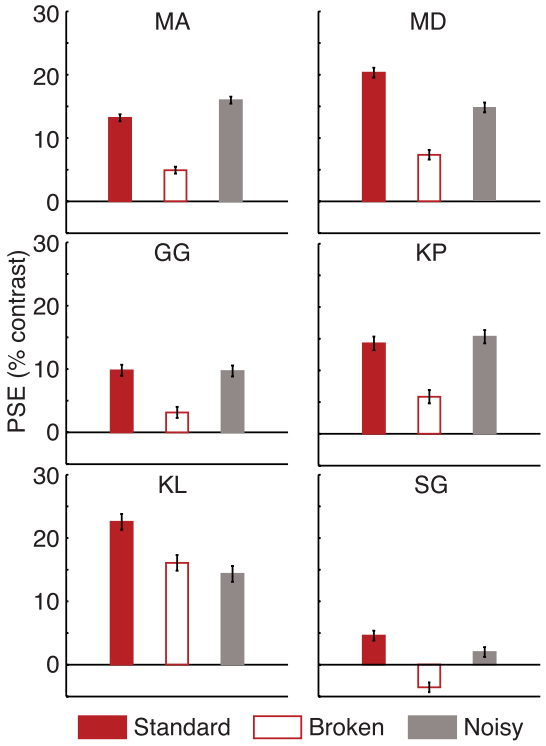

Figure 4. Points of subjective equality in Experiment 2. Bar height corresponds to illusion strength. The top row shows observers who participated in the classification-image experiment. The middle row shows observers who passed screening but did not participate in the classification-image experiment. The bottom row shows observers who did not pass screening; but nevertheless completed the classification-image experiment. Error bars are bootstrapped 95% confidence intervals.

screening criteria and participated in the classification-image experiment. The middle row shows PSEs for observers who passed screening but did not participate in the classification-image experiment. The bottom row shows PSEs for observers who did not pass screening. While observer SG passed screening according to the criteria laid out for Experiment 1, SG showed very small lightness illusions in all conditions, and furthermore showed a lightness illusion in the opposite of the expected direction in the broken argyle condition. Therefore, we consider SG to have failed the screening.

The left-hand side of Figure 5 shows classification images from observers who passed screening in the PSE condition (Figure 4, top row). These classification images were similar to those in Experiment 1 but were rotated 90° clockwise: The lines and triangles neighboring the test diamonds showed the strongest influence, followed by neighboring diamonds. Critically, the diamonds to the left and right of the test diamonds were more influential than the diamonds above and below, reflecting the orientation of the bright and dark strips in the argyle stimulus in this experiment. We conclude that the visual system is guided by the lighting frameworks in the argyle figure when making lightness judgments, rather than by a simple bias toward the vertical.

The right-hand side of Figure 5 shows classification images from the two observers who did not meet the screening criteria in the PSE condition (Figure 4, bottom row) but nevertheless participated in the classification-image experiment due to an error in the screening process. These observers' classification images were markedly different from those of observers with typical PSEs. For observer KL, the lightness illusion was no stronger in the noisy argyle condition than in the broken argyle condition (Figure 4), and this observer's classification image showed only a highly local contrast-like effect, with no significant influence of neighboring diamonds. For observer SG, the lightness illusion was weak in all three argyle stimuli, and reversed from the usual direction in the broken argyle stimulus (Figure 4). This observer's classification image was unusual in that it showed a significant effect of just two diamonds in the same bright and dark strips as the test patches, and a significant effect of a diamond located in a different strip than the test patches. Furthermore, the lines immediately adjacent to test patch A (top test patch in Figure 5) were positive contrast, whereas for all other observers in Experiments 1 and 2 they were negative contrast. We cannot draw strong conclusions, but it is noteworthy that observers with anomalous patterns of PSEs also show qualitatively different classification images.

Experiment 3

Why did only local stimulus elements influence observers' lightness judgments in Experiment 1? One possibility is that the noise in the noisy argyle stimuli weakens grouping between distant patches and the test diamonds. However, given that observers' PSEs for the noisy argyle were not significantly different from those for the standard argyle, this explanation does not seem likely.

Another possibility is that visual processing was limited by the falloff of acuity away from the central visual field. The 2×3 neighborhood of diamonds that showed the strongest influence in the classification image subtended $1.6° \times 2.4°$, which approximately fills the 2° diameter of the fovea. Therefore, it is possible that the set of influential patches was determined by poorer visual processing outside the fovea.

To test this explanation, we measured observers' PSEs in argyle stimuli scaled by factors of 2, 3, and 4. If the influence of stimulus elements is primarily limited

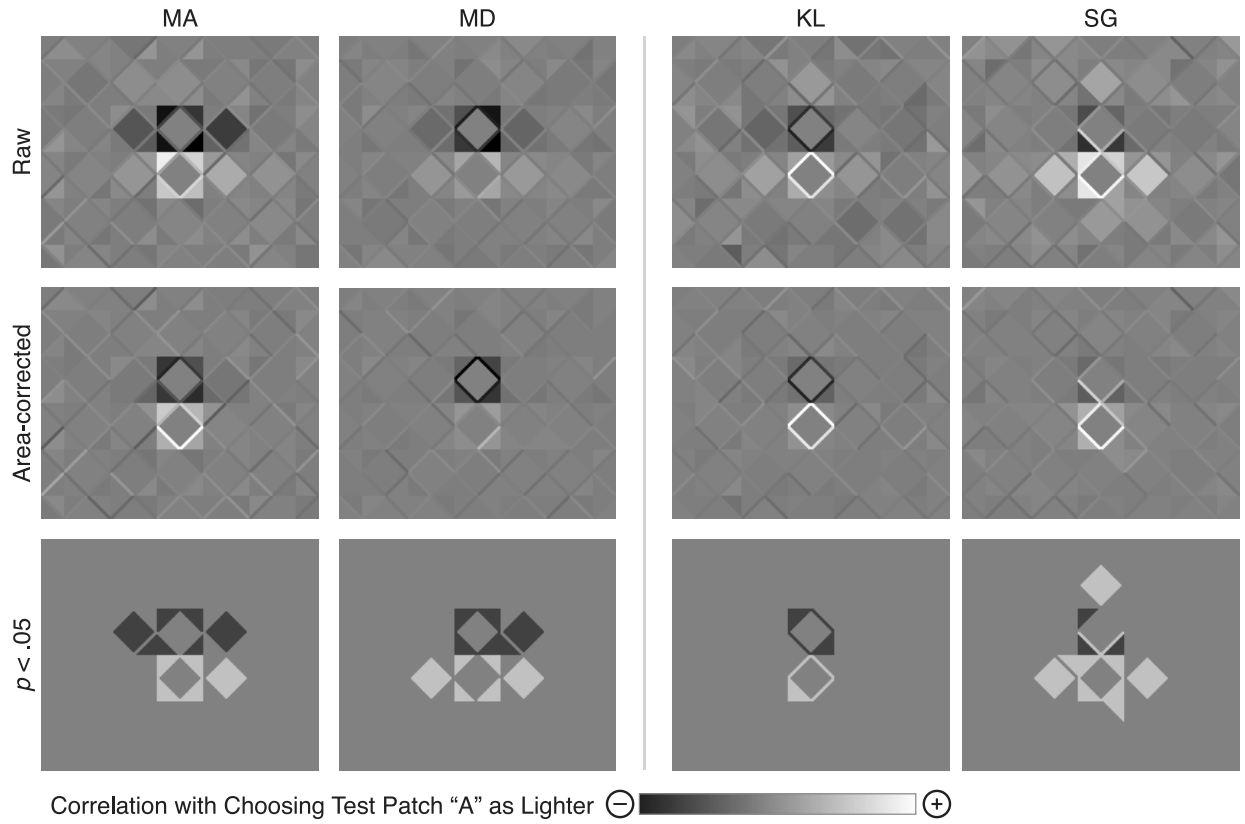

Figure 5. Classification images in Experiment 2. The two observers shown on the left met the screening criteria for points of subjective equality, and the two on the right did not. See Figure 3 for details.

by the falloff in acuity away from the fovea, then scaling up the stimulus to extend well beyond the fovea should greatly weaken the argyle illusion.

Methods

Observers

We recruited 12 observers from York University. None had participated in Experiments 1 and 2. All were unaware of the purpose of the experiment.

Stimuli

The stimuli were the same as in Experiment 1, except that they were scaled by a factor of 1, 2, 3, or 4. At a scale factor of 4, a single test diamond subtended 3.17° × 3.17°.

Procedure

Procedures were the same as in Experiment 1, except that to ensure that the center of the stimulus was presented to the center of each observer's visual field, each stimulus presentation was preceded by a brief fixation dot at the center of the display, and the stimulus duration was reduced from 1,000 ms to 200 ms to prevent observers from making saccades to scan the stimulus.

Results and discussion

Figure 6 shows that, at all scales, PSEs were about twice as large for the standard argyle as for the broken argyle, demonstrating that the argyle illusion is approximately scale invariant in the central 2° to 8° of the visual field. At a scale factor of 4, a single test diamond extended outside of the fovea. We conclude that the range of influential stimulus patches in Experiment 1 was limited not by poorer visual processing outside the fovea but by some other, higher level property of perceptual processing.

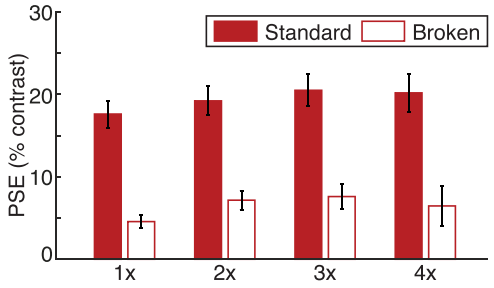

Figure 6. Points of subjective equality in Experiment 3. Error bars are standard error of the mean ($n = 12$). In Appendix C (Figure C1), we show points of subjective equality for individual observers.

Experiment 4

Classification-image analysis does not completely reveal an observer's strategy in a visual task. It is a form of generalized linear regression (Knoblauch & Maloney, 2008), so it captures some properties of the observer's stimulus-to-response mapping but not others. One conclusion we drew from the first three experiments is that only local stimulus elements have a substantial influence on observers' lightness judgments in the argyle task. But have classification images failed to capture the influence of more distant but nevertheless important image features? To test this hypothesis, we measured PSEs with a cropped argyle figure (Figure 7A) that showed only the central part of the stimulus that classification images in Experiments 1 and 2 showed to have a strong influence on observers' lightness judgments. For comparison, we also measured PSEs with the same standard and broken argyle figures as in the previous experiments. If distant image features play an important role in the argyle illusion, then PSEs should be much smaller with the cropped argyle figure than with the standard figure.

Methods

Observers

We recruited 12 observers from York University. Observers KL, KP, MA, and MD had also participated in Experiment 2. Observers BH, KP, LD, MA, MD, and MM had also participated in Experiment 3. All were unaware of the purpose of this experiment.

Stimuli

The stimuli were the standard and broken argyle figures used in the previous experiments, as well as a cropped argyle figure that showed only a central part of the standard argyle (Figure 7A). The cropped figure subtended 2.37° horizontally and 3.16° vertically.

Procedure

Procedures were the same as in the PSE condition of Experiment 1.

Results and discussion

Figure 7B shows that the PSE for the cropped argyle was about twice as large as for the broken argyle ($M_c = 17.2$, $SD_c = 6.9$; $M_b = 7.7$, $SD_b = 4.7$; paired, two tailed, $t(11) = 6.1$, $p < 0.0001$), and that the PSE for the cropped argyle was not significantly different from that for the standard argyle ($M_c = 17.2$, $SD_c = 6.9$; $M_s = 15.1$, $SD_s = 5.8$; paired, two tailed, $t(11) = 1.5$, $p = 0.152$). These findings support the conclusion that the classification images in Experiments 1 and 2 did not miss crucial, long-range influences of distant stimulus elements.

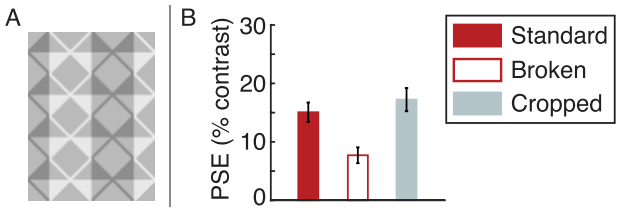

Figure 7. Experiment 4. (A) Sample cropped argyle. (B) Average points of subjective equality in the three conditions ($n = 12$). In Appendix C (Figure C2), we show points of subjective equality for individual observers.

Modeling

Classification images in these experiments showed local, anisotropic, contrast-like effects that matched the vertical structure of the stimulus. How well do current models of lightness and brightness perception account for these findings? To examine this question, we implemented four models of lightness and brightness perception and examined their performance in the same PSE and classification-image experiments that human observers participated in.

The first two models were a high-pass-filter model (Shapiro & Lu, 2011) and the oriented difference-of-Gaussians model (ODOG; Blakeslee & McCourt, 1999, 2012; Blakeslee, Cope, & McCourt, 2015). These are low-level models that do not explicitly represent lighting information; rather, they emphasize the role of contrast, and discount illumination changes using spatial filtering operations.

The other two models were an anchoring model (Economou, Zdravković, & Gilchrist, 2007; Gilchrist et al., 1999) and an atmospheric-link-function (ALF) model that estimates atmospheric transfer functions (Adelson, 2000; Metelli, 1974). These models rely on representations of lighting frameworks and are consistent with Adelson's (1993) explanation of the argyle illusion. However, they are incomplete in that they do not automatically identify lighting frameworks; rather, the modeler must manually segment the image. Extensions of these models that detect lighting boundaries are certainly worth pursuing, but are beyond the scope of this article.

High-pass filter

Illumination is often constant over large regions of a scene, and shadow boundaries are usually blurry, so lighting changes are often signaled by low-pass image features (Land & McCann, 1971). Shapiro and Lu's (2011) high-pass-filter model exploits this fact and removes low-spatial-frequency components from images in order to discount illumination and recover lightness. Given a luminance image $L(x, y)$ as input, the model calculates a lightness response $B(x, y)$ by subtracting out low-frequency components:

$$B(x, y) = L(x, y) - (\text{Box} * L)(x, y), \quad (1)$$

where $\text{Box}(x, y)$ is a 2-D box filter that finds the average luminance in a square region around its center, and $*$ is 2-D convolution.

Equation 1 is equivalent to convolving the input $L(x, y)$ with a high-pass filter (δ–Box), where δ is the Kronecker delta function. This filter has a positive peak in the center, surrounded by a square negative region. This is a simple center–surround receptive field, meaning that the model implements a kind of lateral inhibition. Unlike in classic lateral-inhibition models, the scale of the box filter varies flexibly according to the size of the test patch; Shapiro and Lu allow three orders of magnitude in their model, from 0.01° to 10° of visual angle, and they recommend matching the box-filter width to the width of the test patch. We used a filter width of 0.79° to match the test diamonds in the argyle stimulus.

We computed the model's response to the argyle stimulus on each trial of a lightness-judgment task, using the same stimuli and procedure as with human observers, except that we measured PSEs using 1,000 trials in each condition and classification images using 100,000 trials. On each trial, we took the model's mean pixel-wise response over each test patch to be its lightness response. The model chose the test patch with the higher lightness response as appearing lighter (i.e., the model observer used a difference rule; Pritchett & Murray, 2015). Using these procedures, we measured PSEs for the high-pass-filter model in the standard, broken, and noisy argyle conditions, and a classification image in the noisy argyle condition.

ODOG

The ODOG model (Blakeslee et al., 2015; Blakeslee & McCourt, 1999, 2012) is one of the most successful filter-based models of lightness and brightness perception. Despite having no explicit representation of lighting frameworks, ODOG accounts for some brightness phenomena that have been thought to involve midlevel factors such as shadows and transparency (e.g., White's illusion and Adelson's snake illusion; Blakeslee & McCourt, 2012), although it fails to capture others (e.g., the reverse contrast illusion; Economou et al., 2015).

The ODOG model's filters have a circular on-center region and a larger, elliptical off-surround region. It sums the responses of these filters across scales and orientations in a way that gives greater weight to higher spatial frequencies, normalizes the root-mean-square response at each orientation, and then sums responses across all orientations. We calculated the ODOG model's response to each test patch in the argyle stimulus as the mean pixel-wise response over the patch, and the model chose the test patch with the higher response as appearing lighter. Our implementation of ODOG is a direct MATLAB (MathWorks, Natick, MA) translation of Mathematica code provided as supplementary material by Blakeslee et al. (2015).

Anchoring theory

Anchoring theory states that the human visual system segments an image into lighting frameworks, and estimates lightness using the distribution of image luminances within each framework. We implemented the version of anchoring theory that Economou et al. (2007) used to examine the simultaneous contrast illusion. This model assigns a lightness of 0.90 (i.e., white) to the highest luminance patch in each lighting framework, and a lightness to every other patch in a framework based on the ratio of its luminance to the highest luminance: $l = 0.90\ L/L_{\max}$, where l is lightness, L is luminance, and $L_{\max}$ is the highest luminance in the framework. If the resulting lightness values in a framework span a range smaller than 30:1, then all lightness values less than 0.90 are adjusted downward to expand the range (Economou et al., 2007, equation 2). The final lightness value of each patch is a weighted average of the lightness calculated within its lighting framework and the lightness calculated for the same patch within the entire stimulus (the *global framework*).

The weighting parameter in the final step reflects how strongly the local lighting framework is perceptually segmented from the rest of the scene. If the local framework is strongly segmented, then the weight assigned to the local lightness estimate is near 1 and the weight assigned to the global lightness estimate near 0, so the rest of the scene has little influence on lightness estimates within the local framework.

In our model simulations, the local framework for each test patch was the vertical light or dark strip that contained the patch. The global framework was the whole stimulus. On each trial, the model computed lightness values for the two test patches and chose the lighter patch. In the standard and noisy argyle conditions we used a local-framework weighting parameter of 0.6, and in the broken argyle condition we

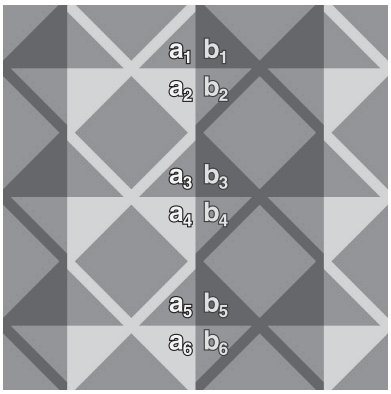

Figure 8. The ALF model. The figure shows a central section of the argyle illusion. If the central vertical boundary is due to a change in lighting, then each horizontally adjacent pair is a single reflectance patch seen in two different lighting frameworks.

used 0.2, as these values reproduced the mean illusion strength that we had found with human observers.

ALF

The ALF is a model that we implemented based on Adelson's (2000) suggestion that the human visual system estimates an *atmospheric transfer function* for each lighting framework—that is, an affine function that maps surface reflectance to image luminance. In this model, X-junctions have the twofold role of segmenting the argyle stimulus into lighting frameworks and providing information about how lighting conditions differ between two frameworks.

Figure 8 shows a vertical boundary in the argyle figure. If the central vertical line is a lighting boundary, then each horizontally adjacent pair of triangles (a_i, b_i) is a single reflectance patch seen in two lighting frameworks. That is, triangles a_i and b_i have the same reflectance but are seen under different lighting conditions. Metelli (1970, 1974) and Adelson (2000) point out that under many lighting conditions, including viewing through many kinds of transparent materials, image luminance L is related to surface reflectance R by an affine transformation, $L = mR + b$. From this it follows that the image luminances of triangles a_i and b_i are also related by an affine transformation, $a_i = pb_i + q$, where here we identify the triangles with their luminance, and the parameters p and q depend on the lighting conditions in the two vertical strips that meet at the boundary.

The ALF model used the relationship $a_i = pb_i + q$ for adjacent triangles in an X-junction and estimated the parameters p and q on each trial using a least-squares linear regression of the 12 a_i values against the corresponding 12 b_i values in the argyle stimulus. It then used this affine transform and the luminance of test diamond B to predict the luminance that B would have if viewed in the same lighting framework as test diamond A. If the predicted luminance of B was higher than the actual luminance of A, then the model chose B as appearing lighter; otherwise it chose A. We used this model with the standard and noisy argyle stimuli. For the broken argyle stimulus, the model simply compared the luminances of the two test patches. As mentioned before, the model did not identify lighting boundaries. Instead, we hard-coded the location of the boundary in the standard and noisy argyle stimuli.

Results and discussion

Figure 9 shows the average human PSEs from Experiment 1 (Figure 9A) in comparison to the PSEs of all models (Figure 9B through E). Figure 10 shows classification images for all models, and average classification images for human observers in Experiment 1.

High-pass filter

Figure 9B shows the high-pass-filter model's PSEs. We found that the model predicts a strong illusion in the standard argyle stimulus, consistent with Shapiro and Lu's (2011) findings. However, we also found that the high-pass-filter model predicts an even stronger lightness illusion with the broken argyle stimulus,

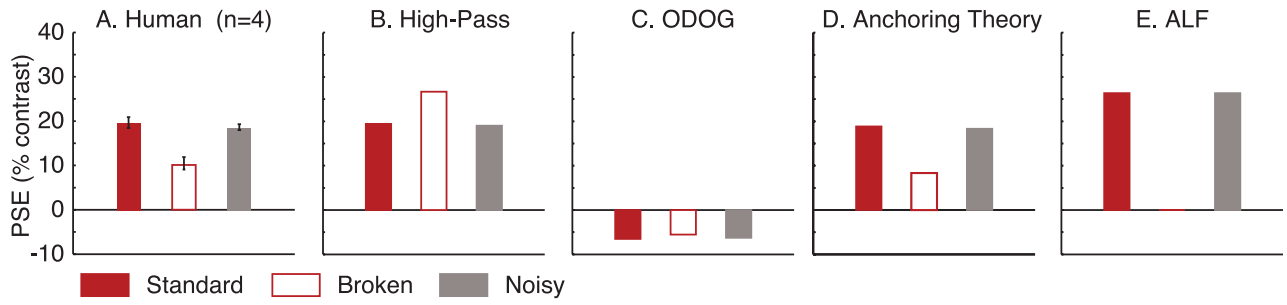

Figure 9. Points of subjective equality from lightness-model simulations, with mean results from human observers in Experiment 1 for comparison.

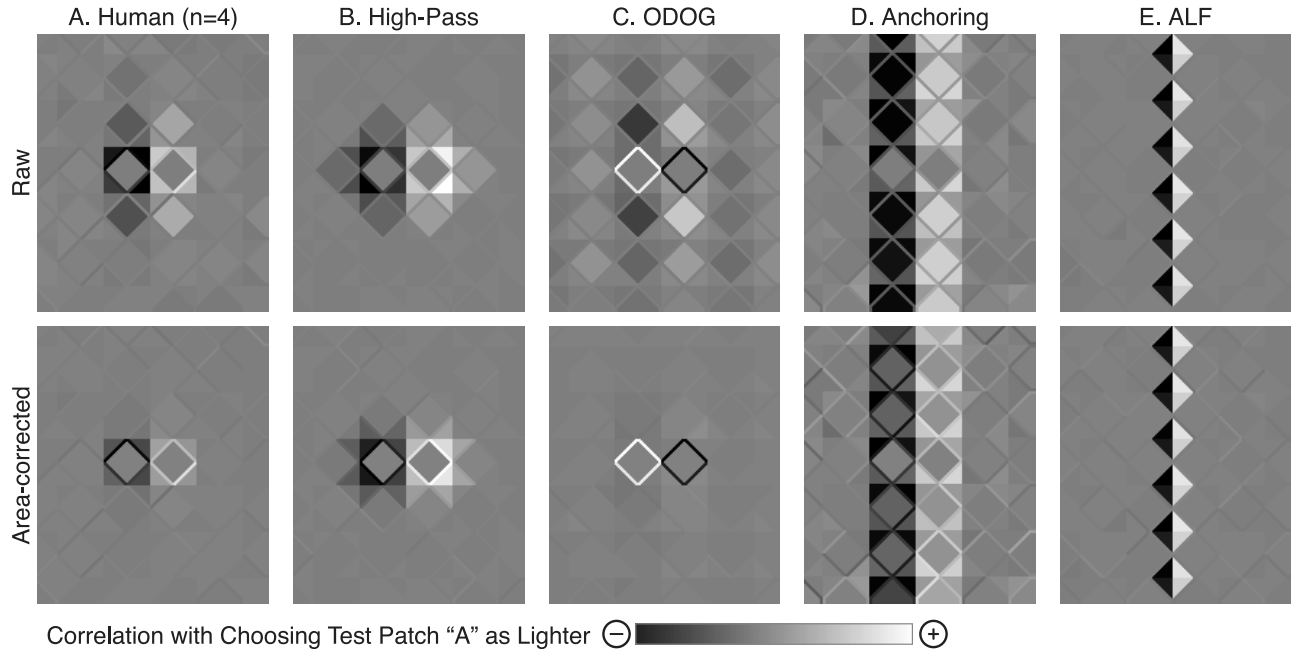

Figure 10. Classification images from lightness-model simulations, with mean results from human observers in Experiment 1 for comparison. See Figure 3 for details.

which Shapiro and Lu did not test. By comparison, human observers perceive a much stronger illusion in the standard argyle than in the broken argyle.

Why does the high-pass model predict this reverse illusion—that is, a stronger illusion in the broken argyle stimulus than in the standard argyle? It subtracts low-pass content from the original image, which tends to make large dark regions brighter and large bright regions darker. In the standard and broken argyles, the low-pass content mostly consists of the bright and dark vertical strips. In the standard argyle, the strips abut each other, and therefore the model's response inside the test diamonds is influenced by neighboring vertical strips. For example, the response of the model at test diamond A is determined by the high-pass filter's response to the diamond's own dark strip and to the two neighboring bright strips. The filter has a large inhibitory surround, so the neighboring bright strips inhibit the model's response in the test diamond. In the broken argyle, there are empty gaps at background luminance between the vertical strips, and these have a weaker inhibitory effect on the model's response at test diamond A. Thus the model has a higher response at A in the broken argyle stimulus than in the standard argyle. For similar reasons, it has a lower response at test diamond B in the broken argyle than in the standard argyle. As a result, the model predicts a stronger lightness illusion for the broken argyle than for the standard argyle. Indeed, accounting for the reduced lightness illusion in the broken argyle poses a challenge to other lateral-inhibition models as well.

Using a different filter size does not improve the PSE predictions of the high-pass model. We measured the model's PSEs for the standard and broken argyle stimuli, with filter sizes ranging from 0.07° to 5.0° of visual angle (Figure 11). At all filter sizes, the model predicted that broken-argyle PSEs would be as high as or higher than standard-argyle PSEs, except at large filter sizes (3° to 5°), where it predicted PSEs near zero.

The high-pass model's classification images (Figure 10B) replicated the influence of bordering lines and triangles that we found in human observers' classification images, but these local elements extended into neighboring vertical strips. The high-pass model is a simple linear filter that is not sensitive to scene

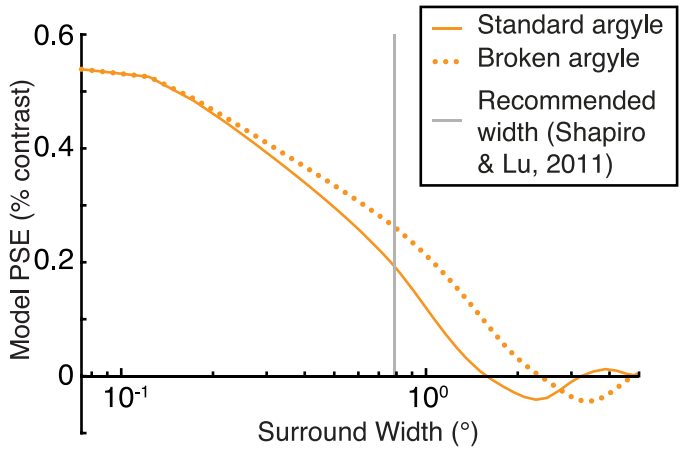

Figure 11. Points of subjective equality predicted by the high-pass-filter model as a function of filter width.

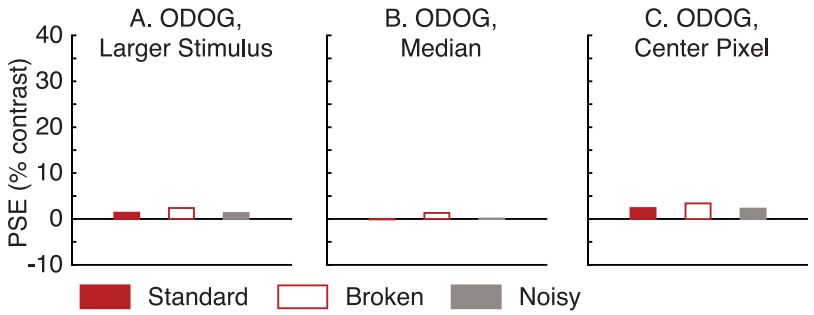

Figure 12. Points of subjective equality of oriented difference-of-Gaussians variants. In all cases, the model has a higher point of subjective equality for the broken argyle than for the standard argyle.

structure, and the influential regions are determined by the filter width. The filter is approximately isotropic (though square), so the influential regions are also approximately isotropic. This is critically different from human observers, who showed anisotropic classification images whose structure matched the structure of the stimulus.

ODOG

Figure 9C shows the ODOG model's PSEs, which differed from those of human observers in two important ways. First, the ODOG had negative PSEs in all three conditions, meaning that it predicted a lightness illusion in the wrong direction: The model chose test diamond B as appearing lighter, whereas human observers chose test diamond A. Second, its PSE was slightly greater in the broken argyle condition than in the standard argyle condition, whereas human observers' PSEs were greater in the standard argyle condition.

The ODOG model's classification images were also qualitatively inconsistent with our findings from human observers (Figure 10C). The lines immediately bordering test diamond A had positive contrast, meaning that positive-contrast stimulus noise in these locations made the model more likely to choose A as appearing lighter. The lines bordering test diamond B had negative contrast. For human observers, these bordering lines had the opposite polarity.

Why did neighboring lines in the ODOG classification images have the opposite polarity to those in the human classification images? The ODOG uses linear filters at six orientations, each with a small, circular excitatory center lobe and a much larger, elongated inhibitory surround lobe. When applied to the argyle stimuli, the excitatory center of the ODOG filters blurred the lines surrounding each test patch into the test patch itself. As a result, random fluctuations that increased the luminance of the surrounding lines increased the model's mean response in the corresponding test patch and made the model more likely to choose that test patch as appearing lighter.

Stimulus size: Blakeslee and McCourt (2012) have also examined the ODOG model's response to the argyle illusion. Their argyle stimulus was approximately five times as large as ours and had different numbers of rows and columns of image elements. Like us, they found that the ODOG incorrectly predicts a stronger illusion in the broken argyle than in the standard argyle, but unlike us they found that it at least correctly predicts that test diamond A appears lighter than test diamond B. To investigate this discrepancy, we ran the ODOG model in a simulation with our argyle stimuli enlarged by a factor of 5 (19° vertical extent). This simulation with larger stimuli gave results that were consistent with Blakeslee and McCourt's findings: With the larger stimulus, ODOG correctly predicts that test diamond A is seen as lighter than test diamond B but incorrectly predicts that the illusion is stronger in the broken condition than in the standard condition (Figure 12A). The difference between results with small and large stimuli shows that the illusion is not scale invariant for the ODOG model, whereas our Experiment 3 showed that the illusion strength is scale invariant over a large range for human observers. Furthermore, for both small and large stimuli, PSEs show that the argyle illusion is much weaker for the ODOG model (~5% contrast) than for human observers (~20% contrast).

The ODOG model's classification images were also different with the larger stimuli (Figure 13A) than with the smaller stimuli (Figure 10C). With the larger stimuli, image patches next to the test diamonds had a contrast-like effect, where bright neighboring patches made the test patches appear darker to the model; whereas with the smaller stimuli the neighboring patches had the opposite effect. Also, with the larger stimuli a much larger area of neighboring patches had a substantial effect on the lightness of the test patches, with a larger role for neighboring patches in the same frameworks as the test patches. Overall, ODOG's classification image with the larger stimuli was more

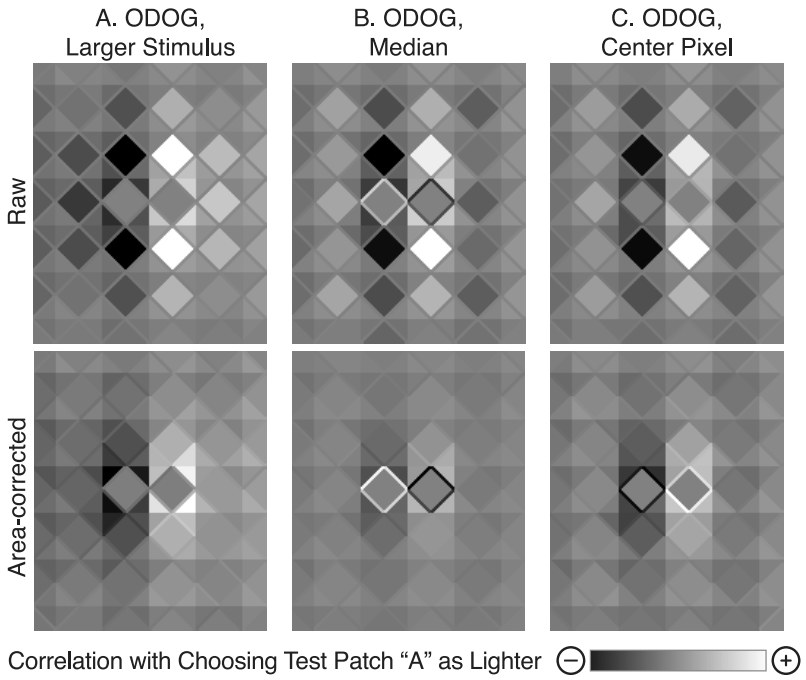

Figure 13. Classification images of oriented difference-of-Gaussians variants.

similar to those from human observers than the one with the smaller stimuli.

ODOG extensions

To examine the robustness of our findings, we tested variants of the ODOG model that used alternative methods of reading out the lightness response in the test diamonds.

In the first variant, the model based its decisions on the median lightness response in each test diamond instead of the mean. This model had near-zero, negative PSEs (Figure 12B), and its classification image still showed a strong effect of the lines neighboring the test diamonds, with the effect in the opposite direction as for human observers (Figure 13B).

In the second variant, the model based its decisions on the lightness response of a single pixel in the center of each test diamond. This model had near-zero, positive PSEs (Figure 12C), with a slightly larger lightness illusion for the broken argyle stimulus than for the standard argyle. This model's classification image was broadly consistent with those of human observers (Figure 13C). As explained previously, in the standard ODOG model the lines adjacent to the test diamonds were blurred into the test diamonds by the excitatory center of the ODOG filters, and thus classification images showed that these lines had an excitatory effect on the model's lightness judgments. In the center-pixel variant of the ODOG, though, the excitatory effect of the adjacent lines does not reach far enough into the test diamonds to affect the observers' responses, and thus the adjacent lines no longer have an excitatory effect. Nevertheless, because this model makes incorrect predictions for the PSEs, we reject it as a model of lightness perception in this task.

Anchoring theory

Figure 9D shows the anchoring model's PSEs. The model showed a larger lightness illusion with the standard argyle figure than with the broken argyle. This is expected, as we chose the local weighting parameter in the two conditions to produce the correct PSEs.

The anchoring model's classification image was very different from human observers' (Figure 10D). According to anchoring theory, the influence that an image patch has on the perceived lightness of a target patch in the same framework does not decrease with distance (e.g., Gilchrist, 2006, pp. 303–306), and the model's classification image shows that image patches along the entire vertical length of the lighting frameworks containing the target patches influenced lightness judgments. For human observers, only image regions near the target patches had a strong influence.

One way of improving the predictions of the anchoring model would be to weaken the influence of image patches with distance from the test patches. This would be a rational strategy if image patches far from the test patches were expected to convey little information about the lighting conditions at the test patches. We return to this point briefly in the General discussion.

ALF

Figure 9E shows that the ALF model correctly predicted a larger illusion with the standard argyle figure than with the broken argyle. This is not surprising, since we designed the model specifically to compensate for the lighting boundary between the two test patches in the standard argyle and not in the broken argyle.

The model's classification image was qualitatively different from human observers' (Figure 10E). The model showed a strong influence of image patches that formed X-junctions, whereas human observers did not. Conversely, human observers showed significant influences of a wider range of neighboring patches in the same framework as the test patches than the model did. The classification images for this model are largely what we would expect from the model's design. Nevertheless, the image patches in X-junctions contribute to the model's lightness judgments via the nonlinear process of estimating linear regression coefficients, so these classification images give a useful confirmation that X-junctions would appear in human observers' classification images if observers used them to compensate for the lighting boundary via an atmospheric link function of the kind we have modeled here.

As with the anchoring model, the ALF model's predictions would be more similar to human observers' classification images if the influence of key image features (here, X-junctions) declined with distance from the test patches. In the General discussion, we briefly outline how such behavior could emerge from a more complete model that estimates lighting boundaries from the image rather than having them artificially imposed.

General discussion

The classification images we measured from human observers showed a surprising mix of features that we might have expected from low-level and midlevel models of lightness perception. On the one hand, they showed local, contrast-like effects of nearby image elements on lightness judgments, as predicted by classic low-level lateral-inhibition models. On the other hand, the influence of surrounding image elements depended on the structure of the stimulus, in that lightness judgments were affected more strongly by elements in the same bright and dark vertical frameworks as the test patches than by similar elements in different frameworks. This is what we would expect from midlevel models that take account of lighting boundaries.

These findings rule out a broad class of linear lateral-inhibition models. According to such models, the psychophysical receptive field for a lightness judgment is the stimulus convolved with an isotropic linear filter. These models cannot block the influence of image elements that are separated from the target patch by stimulus features such as lighting boundaries.

However, the ODOG model is a good illustration of how sophisticated behavior can emerge from even fairly simple elaborations of linear filtering models. ODOG did not correctly predict the PSEs we found with human observers, and unlike with human observers, its PSEs varied substantially with stimulus size. Furthermore, at our baseline stimulus size its classification images were very different from those of human observers. However, with a larger stimulus and in a variant of ODOG that based its lightness judgments on the center pixel of the test diamonds, its classification images were much more similar to those from human observers, and even showed a larger influence of stimulus elements that belonged to the same lighting frameworks as the test patches. Our experiments, and many others, show that midlevel factors such as lighting boundaries play an important role in lightness perception; but it is not yet clear how flexible the perceptual mechanisms that process these features need to be. Although ODOG fails to predict the strength or direction of the argyle illusion, classification images show that under some conditions it is guided by image features that are similar to those that guide human observers, which suggests that a different model constructed from similar elements (e.g., linear filters, contrast normalization) may be able to deal more adequately with lighting boundaries when computing lightness.

Human observers' raw classification images (Figures 3 and 5, top row) show that diamonds above and below the test diamonds had a measurable influence on their lightness judgments, but area-corrected classification images (middle row) show that this influence was small and was measurable only because of the relatively large size of the diamonds. The raw classification images are useful for testing computational models, as they show that observers' lightness computations depend on lighting frameworks. However, the area-corrected classification images are also informative, as they show how very quickly the influence of neighboring elements falls off with distance. The cropped argyle condition in Experiment 4 gives some confirmation that the classification images did not fail to capture more distant but nevertheless important image features that strongly affected observers' responses.

A reasonable concern in classification-image experiments is whether the noise that is added to the stimulus changes how observers process the stimulus and make their responses. Murray and Gold (2004) and Murray

(2011) considered this problem and, based on observations such as the fact that noise-masking functions tend to be linear, concluded that in many tasks the stimulus noise does not substantially change how observers process stimuli. In the present experiments, we screened observers for the classification-image task by testing whether the lightness illusion was weakened in the noisy argyle condition. We found that for most observers, the illusion was about as strong in the noisy argyle condition as in the standard argyle condition; and furthermore, several of the observers who failed screening did not see the typical illusion even in the noiseless conditions—for example, in Experiment 1 observer MM perceived a stronger illusion in the broken argyle than in the standard argyle. In addition, we used patch-wise stimulus noise, and so the noise did not introduce any new luminance edges but only perturbed the luminance of existing stimulus regions. These observations suggest that the stimulus noise did not substantially disrupt normal visual processing and that the classification images we measured reveal useful information about how observers process the standard, noiseless argyle figure.

In this article we have not addressed the distinction between lightness, which is perceived reflectance, and brightness, which is perceived luminance. Observers can make distinct judgments of reflectance and luminance in natural scenes, and asking them to judge one or the other can lead to different experimental results (e.g., Arend & Spehar, 1993). Nevertheless, the relationship between lightness and brightness is not well understood. To take just one example, Adelson (2000) correctly describes the argyle illusion as a brightness illusion yet convincingly explains it as the product of a mechanism that computes lightness. Furthermore, recent work suggests that whether observers can distinguish lightness and brightness depends on the realism of the stimulus. Logvinenko and Maloney (2006) found that in achromatic paper stimuli with real shadows, multidimensional scaling revealed two perceptual dimensions, roughly corresponding to surface reflectance and local lighting intensity. However, Logvinenko, Petrini, and Maloney (2008) found that in the snake illusion (Adelson, 2000)—a simplified stimulus with no real lighting boundaries—achromatic patches had only one perceptual dimension. Here we do not propose a solution to the problem of how lightness and brightness are related, and we raise it here only to point out that although we have described our experiments as examining lightness perception, they bear on brightness perception as well. Logvinenko and Maloney's (2006; Logvinenko, 2015; Logvinenko et al., 2008) work on the perceptual dimensions of achromatic stimuli has made important progress in clarifying these issues.

When testing computational models of lightness perception, we found it useful to run the models in the same experiments as human observers and examine the models' performance using the same measures as for human observers, namely PSEs and classification images. A common alternative approach is to report models' responses to stimuli in arbitrary units—for example, the mean response of a linear filter in a region of interest. This can make it difficult to know how well the model actually accounts for some aspects of behavioral measurements from human observers, such as the absolute strength of an illusion. By measuring PSEs for computational models in full simulations of the experiments that human observers participated in, we were able to show, for instance, that some models predict lightness illusions that are about as strong as those seen by human observers, while others predict much weaker illusions (Figure 9).

In addition to testing computational models of lightness perception, our findings suggest possible directions for improving these models. The clear role of lighting frameworks in lightness computations shows that a broad class of linear lateral-inhibition models is inadequate. Elaborated low-level models like the ODOG have some promise, but a successful model will need to base lightness computations on image elements within lighting frameworks in order to avoid lightness illusions in the wrong direction due to contrast with surrounding frameworks. The atmospheric transfer function gives a useful way of describing local lighting regions, but our classification images showed that the ALF model that we built using this notion gives too large a role to X-junctions. Variants of this model that rely on image features other than X-junctions to estimate the atmospheric transfer function would produce very different classification images, though, so our findings do not rule out the idea that an estimate of the atmospheric transfer function plays a role in lightness perception.

The anchoring model is the most promising of the models we tested, with the substantial caveats that it does not account for the falloff in the influence of image elements at greater distances within lighting frameworks, and it requires the user to manually segment the stimulus into lighting frameworks and choose the weighting factor that links local and global frameworks. These points may be related. In that model, the luminance of one image patch affects the perceived lightness of another image patch only to the extent that they are judged to have common lighting conditions, either because they belong to the same local framework or because the weighting factor that links their frameworks is greater than zero. When implementing the anchoring model we assumed that the argyle stimulus is cleanly divided into vertical lighting frameworks, but this need not be the case. A fully computational anchoring model that estimates the lighting frameworks in a stimulus might judge some of the diagonal lines in the argyle stimulus to be weak lighting boundaries as well, and if so, then distant

image elements in a single vertical strip may be judged to belong to very different lighting frameworks. Thus, a fully computational anchoring theory may resolve both of these problems. In earlier work we found that a Bayesian model that makes simple assumptions about lighting and reflectance can account for many of the rules that make up anchoring theory, such as that perceived white should be anchored to the highest luminance (Murray, 2013; see also Allred & Brainard, 2013). Developing a Bayesian theory that identifies lighting frameworks and makes rational estimates of reflectances within each framework is a promising direction for future work on lightness models.

Keywords: lightness, argyle illusion, classification images

Acknowledgments

Commercial relationships: none.
Corresponding author: Minjung Kim.
Email: minjung.kim@tu-berlin.de.
Address: Fachgruppe Modellierung Kognitiver Prozesse, Technische Universität Berlin, Berlin, Germany.

References

Adelson, E. H. (1993, December 24). Perceptual organization and the judgment of brightness. *Science*, *262*(5142), 2042–2044.

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 339–351). Cambridge, MA: MIT Press.

Ahumada, A., Jr. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*(1):8, 121–131, https://doi.org/10.1167/2.1.8. [PubMed] [Article]

Allred, S. R., & Brainard, D. H. (2013). A Bayesian model of lightness perception that incorporates spatial variation in the illumination. *Journal of Vision*, *13*(7):18, 1–18, https://doi.org/10.1167/13.7.18. [PubMed] [Article]

Arend, L. E., & Spehar, B. (1993). Lightness, brightness, and brightness contrast: 1. Illumination variation. *Perception & Psychophysics*, *54*, 446–456.

Beck, J., Prazdny, K., & Ivry, R. (1984). The perception of transparency with achromatic colors. *Perception & Psychophysics*, *35*, 407–422.

Blakeslee, B., Cope, D., & McCourt, M. E. (2015). The oriented difference of Gaussians (ODOG) model of brightness perception: Overview and executable Mathematica notebooks. *Behavior Research Methods*, *47*, 1–7, https://doi.org/10.3758/s13428-015-0573-4.

Blakeslee, B., & McCourt, M. E. (1999). A multiscale spatial filtering account of the White effect, simultaneous brightness contrast and grating induction. *Vision Research*, *39*, 4361–4377.

Blakeslee, B., & McCourt, M. E. (2012). When is spatial filtering enough? Investigation of brightness and lightness perception in stimuli containing a visible illumination component. *Vision Research*, *60*, 40–50.

Economou, E., Zdravković, S., & Gilchrist, A. L. (2007). Anchoring versus spatial filtering accounts of simultaneous lightness contrast. *Journal of Vision*, *7*(12):2, 1–15, https://doi.org/10.1167/7.12.2. [PubMed] [Article]

Economou, E., Zdravković, S., & Gilchrist, A. L. (2015). Grouping factors and the reverse contrast illusion. *Perception*, *44*, 1383–1399.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Fisher, R. A. (1966). *The design of experiments* (8th ed.). New York, NY: Hafner Publishing Company.

Flynn, O., & Shapiro, A. G. (2014). A note concerning the relationship between the Adelson's Argyle illusion and Cornsweet edges. *Psihologija*, *47*(3), 353–358.

Gilchrist, A. L. (2006). *Seeing black and white*. New York, NY: Oxford University Press.

Gilchrist, A. L., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., ... Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review*, *106*, 795–834.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, *10*, 663–666.

Heinemann, E. G., & Chase, S. (1995). A quantitative model for simultaneous brightness induction. *Vision Research*, *35*, 2007–2020.

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, *8*(16):10, 1–19, https://doi.org/10.1167/8.16.10. [PubMed] [Article]

Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile? *Vision Research*, *44*, 1493–1498.

Land, E. H., & McCann, J. J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, *61*, 1–11.

Logvinenko, A. D. (2015). The achromatic object-

colour manifold is three-dimensional. *Perception*, *44*, 243–268.

Logvinenko, A. D., & Maloney, L. T. (2006). The proximity structure of achromatic surface colors and the impossibility of asymmetric lightness matching. *Perception & Psychophysics*, *68*, 76–83.

Logvinenko, A. D., Petrini, K., & Maloney, L. T. (2008). A scaling analysis of the snake lightness illusion. *Perception & Psychophysics*, *70*(5), 828–840.

Metelli, F. (1970). An algebraic development of the theory of perceptual transparency. *Ergonomics*, *13*, 59–66.

Metelli, F. (1974, April). The perception of transparency. *Scientific American*, *230*(4), 90–98.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision*, *11*(5):2, 1–25, https://doi.org/10.1167/11.5.2. [PubMed] [Article]

Murray, R. F. (2013). Human lightness perception is guided by simple assumptions about reflectance and lighting. *Proceedings of SPIE*, *8651*, 865106.

Murray, R. F., & Gold, J. M. (2004). Troubles with bubbles. *Vision Research*, *44*(5), 461–470.

Nagai, T., Ono, Y., Tani, Y., Koida, K., Kitazaki, M., & Nakauchi, S. (2013). Image regions contributing to perceptual translucency: A psychophysical reverse-correlation study. *i-Perception*, *4*, 407–428, https://doi.org/10.1068/i0576.

Nichols, T. E., & Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*, 1–25.

Pritchett, L. M., & Murray, R. F. (2015). Classification images reveal decision variables and strategies in forced choice tasks. *Proceedings of the National Academy of Sciences, USA*, *112*, 7321–7326.

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*(2), 147–166, https://doi.org/10.1207/s15516709cog2802_2.

Robinson, A. E., Hammon, P. S., & de Sa, V. R. (2007). Explaining brightness illusions using spatial filtering and local response normalization. *Vision Research*, *47*, 1631–1644.

Shapiro, A., & Lu, Z.-L. (2011). Relative brightness in images can be accounted for by removing blurry content. *Psychological Science*, *22*, 1452–1459.

Shimozaki, S. S., Eckstein, M. P., & Abbey, C. K. (2005). Spatial profiles of local and nonlocal effects upon contrast detection/discrimination from classification images. *Journal of Vision*, *5*(1):5, 45–57, https://doi.org/10.1167/5.1.5. [PubMed] [Article]

Volterra, V. (1930). *Theory of functionals and of integral and integro-differential equations*. London, UK: Blakie.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*, 113–120.

Wetherill, G. B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *The British Journal of Mathematical and Statistical Psychology*, *18*, 1–10.

Appendix A: Patch size

In a model where the signal is represented as a vector of pixel values $\mathbf{s} = (s_i)$ and stimulus noise is represented as a vector $\mathbf{n} = (n_i)$, the decision variable of a template-matching observer with late noise is

$$d = \sum_{i=1}^{N} (s_i + n_i) t_i + z, \quad (2)$$

where $\mathbf{t} = (t_i)$ is the template and z is the internal noise. According to the most common model of performance in simple two-alternative tasks, the observer makes one response when the decision variable d is at or below a criterion value and the other response when d is above the criterion. Classification-image methods exploit the fact that the expected value of each noise element n_i, conditional on the observer model (Equation 2) giving one of the two possible responses, is proportional to the corresponding element t_i of the template (Ahumada, 2002; Murray, 2011; Volterra, 1930).

Suppose that the stimulus is divided into N patches (i.e., groups of pixels), where the ith patch has M_i pixels. We can represent the signal as s_{ij}, the stimulus noise as n_{ij}, and the template as t_{ij}, where $i \in \{1, \ldots, N\}$ is the patch number and $j \in \{1, \ldots, M_i\}$ is the pixel number within the patch. Then the decision variable is

$$d = \sum_{i=1}^{N} \sum_{j=1}^{M_i} (s_{ij} + n_{ij}) t_{ij} + z. \quad (3)$$

If the signal, noise, and template are constant within each patch ($n_{ij} = n_i$, $s_{ij} = s_i$, $t_{ij} = t_i$), this becomes

$$d = \sum_{i=1}^{N} \sum_{j=1}^{M_i} (s_i + n_i) t_i + z \quad (4)$$

$$= \sum_{i=1}^{N} M_i (s_i + n_i) t_i + z. \quad (5)$$

Comparing Equations 2 and 5 shows that an experiment with N stimulus patches is equivalent to an experiment with N stimulus pixels and template values

$M_i t_i$. Thus, the expected value of each classification-image element in an experiment with patch-wise noise is proportional to $M_i t_i$, and the expected value of each classification-image element divided by M_i is proportional to t_i. This is the correction for patch size that we used when calculating classification images.

Appendix B: Significance tests

In a two-tailed z test of a single classification-image element with value c_i, we could calculate the z score z_i of c_i under the null hypothesis that the expected value of c_i is zero, set a criterion $z_{crit} = 1.96$, and consider c_i to be significantly different from zero if $|z_i| > z_{crit}$. This would produce a type I error rate of $\alpha = 0.05$. However, our classification images contained 280 elements, so following this procedure for each element individually would produce approximately $0.05 \times 280 = 14$ false positives per classification image. To correct for multiple comparisons, we used a higher criterion z_{crit}, chosen so that the probability of even a single type I error in a classification image was $\alpha = 0.05$ (Nichols & Holmes, 2001).

We calculated z_{crit} as follows:

$$P(\text{ any } |z_i| > z_{crit} \mid H_0) = P(\max(|z_i|) > z_{crit} \mid H_0) \quad (6)$$

$$= P(\min(z_i) < -z_{crit} \vee \max(z_i) > z_{crit} \mid H_0) \quad (7)$$

$$\simeq 2P(\max(z_i) > z_{crit} \mid H_0). \quad (8)$$

The last equality is only approximate, because the maximum and minimum of the z_i values are not statistically independent, but with 280 classification-image elements they will be very close to independent:

$$= 2(1 - P(\max(z_i) \leq z_{crit} | H_0)) \quad (9)$$

$$= 2(1 - F_z(z_{crit})^n), \quad (10)$$

where F_z is the cumulative distribution function of the z score of each classification-image pixel under the null hypothesis that the expected value of each classification-image element is zero; and n is the number of classification-image pixels. F_z is just the standard normal cumulative distribution function $\Phi(x)$:

$$= 2(1 - \Phi(z_{crit})^n). \quad (11)$$

For a type I error rate of $\alpha = 0.05$ and $n = 280$ classification-image pixels, this evaluates to $z_{crit} = \Phi^{-1}((1 - 0.05/2)^{1/280}) = 3.74$. Alternatively, Nichols and Holmes (2001) describe a method for choosing z_{crit} using Monte Carlo simulations.

To calculate the z score for each classification-image element, we need its standard deviation under the null hypothesis that the expected value of the element is zero. Under the null hypothesis, the classification image is just a weighted sum of stimulus noise samples that are uncorrelated with the observer's responses, so we can calculate the standard deviation of the classification-image elements using elementary formulas for the standard deviation of arithmetic functions of random variables. For example, if the stimulus noise contrast has patch-wise standard deviation $\sigma = 0.20$, and if the observer gives response A on 5,000 trials and response B on 5,200 trials, then under the null hypothesis the variance of each classification-image element is $0.20^2/5{,}000 + 0.20^2/5{,}200$.

Appendix C: Subjective equality for individual observers

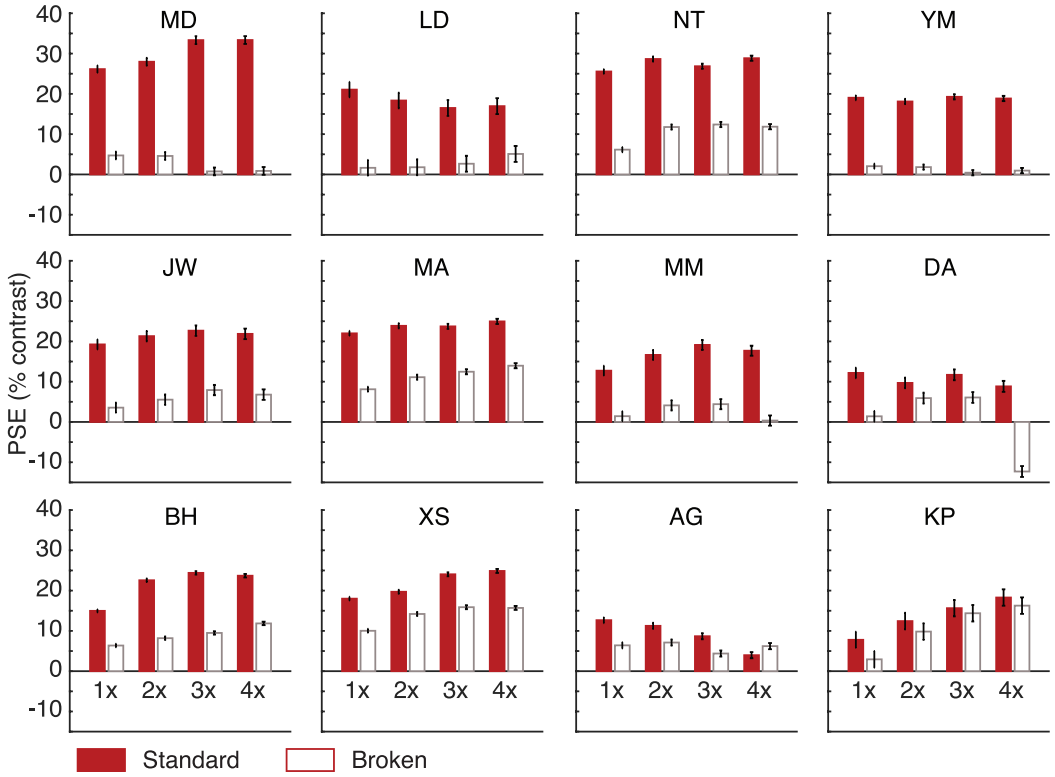

Figure C1. Individual observers' PSEs in Experiment 3. Each bar shows an observer's PSE with the argyle figure enlarged by a factor of one, two, three, or four. Error bars are bootstrapped 95% confidence intervals.

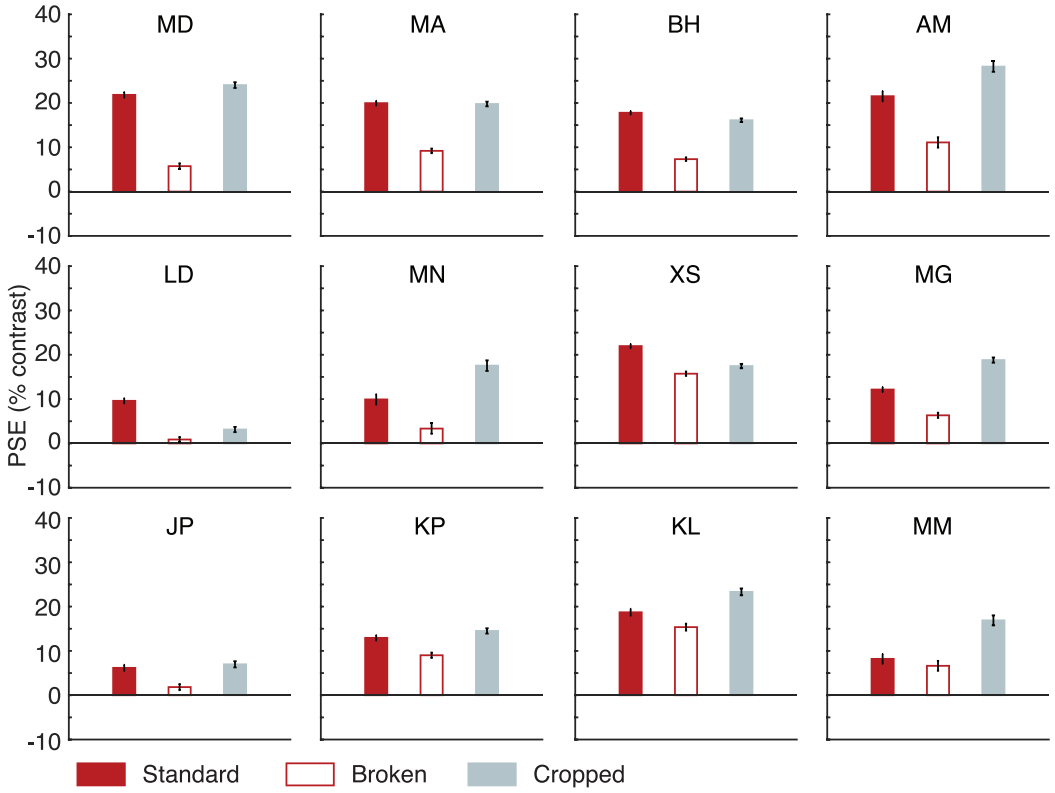

Figure C2. Individual observers' PSEs in Experiment 4. Each bar shows an observer's PSE with the cropped, standard, or broken argyle figure. Error bars are bootstrapped 95% confidence intervals.