Spatial relationships between contours impact rapid scene classification

John Wilder University of Toronto, Toronto, Ontario, Canada

Sven Dickinson University of Toronto, Toronto, Ontario, Canada

Allan Jepson University of Toronto, Toronto, Ontario, Canada

Dirk B. Walther University of Toronto, Toronto, Ontario, Canada

Photographs and line drawings of natural scenes are easily classified even when the image is only briefly visible to the observer. Contour junctions and points of high curvature have been shown to be important for perceptual organization (Attneave, 1954; Biederman, 1987) and have been proposed to be influential in rapid scene classification (Walther & Shen, 2014). Here, we manipulate the junctions in images, either randomly translating them, or selectively removing or maintaining them. Observers were better at classifying images when the contours were randomly translated (disrupting the junctions) than when the junctions were randomly shifted (partially disrupting contour information). Moreover, observers were better at classifying a scene when shown only segments between junctions, than when shown only the junctions, with the middle segments removed. These results suggest that categorizing line drawings of real-world scenes does not solely rely on junction statistics. The spatial locations of the junctions are important, as well as their relationships with one another. Furthermore, the segments between junctions appear to facilitate scene classification, possibly due to their involvement in symmetry relationships with other contour segments.

Introduction

Humans can effortlessly categorize the objects and scenes around them. The mechanisms enabling this categorization are still not completely understood. According to one of the dominant theories of object recognition, categorization relies on recovering 3D parts of objects and their spatial relationships, which in turn relies heavily on analyzing the contour junctions in the image (Biederman, 1987; Biederman & Cooper, 1991). Biederman and Cooper (1991) argued that junctions are more useful than contour segments between junctions. They suggested that missing portions of contours between two vertices are more easily filled in, whereas it is more difficult to complete a contour when junctions are missing. Similarly, Attneave (1954) showed that an object is easily recognized by connecting points of high-curvature (L-junctions) with straight segments. Even though many of the details of the true bounding contour are lost, the object is still easily recognized.

In everyday life, we do not see isolated contours of objects, but complex arrangements of many objects and surfaces. Even though such scenes are more complex than an isolated object, humans are still able to process them rapidly (Potter & Levy, 1969; Thorpe, Fize & Marlot, 1996; VanRullen & Thorpe, 2001). The rapid speed of processing scenes led researchers to suspect that easily extracted summary statistics of visual features may underlie scene categorization (Oliva & Schyns, 2000; Torralba & Oliva, 2003; Delorme, Richard, & Fabre-Thorpe, 2000; Wichmann, Drewes, Rosas, & Gegenfurtner, 2010). Loschky et al. (2007) and Loschky and Larson (2008), however, have shown that summary statistics over the entire image are not adequate for explaining scene categorization, and instead that sufficient localization of image features is necessary.

Just like photographs of scenes, line drawings of real-world scenes can be rapidly categorized (Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). Also, as with photographs, summary statistics of image features can be extracted from these line drawings. Walther and Shen (2014) showed that non-accidental relationships between contours are more influential in scene categorization than are unary features, such as contour length and orientation. Junctions are useful in the determi-

Citation: Wilder, J., Dickinson, S., Jepson, A., & Walther, D. B. (2018). Spatial relationships between contours impact rapid scene classification. *Journal of Vision*, *18*(8):1, 1–15, https://doi.org/10.1167/18.8.1.

https://doi.org/10.1167/18.8.1 Received September 29, 2017; published August 1, 2018 ISSN 1534-7362

nation of border ownership (Sajda & Finkel, 1995; Craft, Schütze, Niebur, & Von Der Heydt, 2007), which may be more important in a complex scene than in the case of isolated objects, further strengthening the hypothesis that junctions may be useful for scene categorization.

Contour junctions have been shown to be instrumental for categorizing objects (Biederman, 1987) and scenes (Walther & Shen, 2014). In contrast, Kennedy and Domander (1985) found middle segments between two junctions to be more important than junctions for object recognition. A similar result was found by Panis, De Winter, Vandekerckhove, and Wagemans (2008), using a more comprehensive object dataset. This suggests that spatial relationships other than junctions may be relevant for recognition as well.

In this paper we aim to resolve these apparently contradictory accounts of scene perception. We do this with two experimental manipulations, one where the spatial locations of the contours are disrupted (Experiment 1) and the other where spatial locations are maintained (Experiments 2 and 3). In Experiment 1, we either maintain or destroy junction distributions by randomly shifting the contours or junctions about the scene. This manipulation disrupts spatial location information. In Experiments 2 and 3, either the junctions are entirely removed, leaving only the middle segments between junctions, or the middle segments are removed, leaving only the junctions. This manipulation preserves spatial location information.

These manipulations allow us to determine (a) the role of contour junction summary statistics versus other summary statistics, and (b) if the relative spatial location of contours carries category-relevant information about complex, real-world scenes. In particular, we find that junction summary statistics, while important, are insufficient to explain human scene categorization. Additionally, unshifted contour segments between junctions, presumably through their longer range spatial interactions, carry category-specific information. This indicates that categorization of complex real-world scenes is not easily reduced to a single aspect of scene structure, but that a more inclusive feature set needs to be taken into account.

Experiment 1

Due to limited exposure to the stimulus at which humans are able to classify natural scenes, even when the presentation is followed by a mask, researchers theorized that scene classification must depend on easily extracted features, for which no neural feedback is required (Oliva & Schyns, 2000; Torralba & Oliva, 2003). Summary statistics of simple line-drawing features can be rapidly extracted from visual input, and have been shown to be related to the scene classification of line drawings (Walther et al., 2011; Walther & Shen, 2014). Walther and Shen (2014) showed that a computer model can accurately classify scenes using only extracted feature histograms of either contour length, contour curvature, contour orientation, the angle between junctions of contours, or the label of the junctions between contours (e.g., T-junction), but they found that only a classifier trained using only the curvature, junction type, or junction label histograms had similar error patterns as human observers.

Walther and Shen (2014) randomly and independently translated contours within an image, thereby combining a change in junction statistics with a change in spatial relations. In this experiment, we directly test the effect of spatial shuffling of contours or junctions on scene perception. Participants were tested on three different conditions: (1) intact line drawings, (2) contour-shifted, and (3) junction-shifted line drawings. Whereas the first condition served as a control, conditions 2 and 3 were designed to disambiguate the role of perturbing junction statistics and the locations of contours and their junctions within the image. In condition 2, contours were randomly translated, thereby destroying existing junctions and spatial relations of contours, while keeping constant all summary statistics of contours themselves, such as contour length, orientation, and curvature. In condition 3, contours were split into individual junction regions, which were subsequently spatially shuffled. This manipulation retained the summary statistics of the junctions at the cost of perturbing the distribution of contour length and, to a smaller extent, curvature. Note that all three conditions contained the same number of contour pixels.

Methods

Participants

Participants were 19 undergraduate students (17 female, two male) who received course credit for their participation. Ages ranged from 18 to 25, with a mean of 19. The study was approved by the University of Toronto Research Ethics Board (REB) and written informed consent was given by each participant prior to beginning the experiment.

Stimuli

Line drawings of real-world scenes were obtained from artists tracing the most salient outlines in a set of photographs (Walther et al., 2011). The images were highly typical examples of beaches, forests, mountains, city streets, highways, and offices (Torralbo et al., 2013).

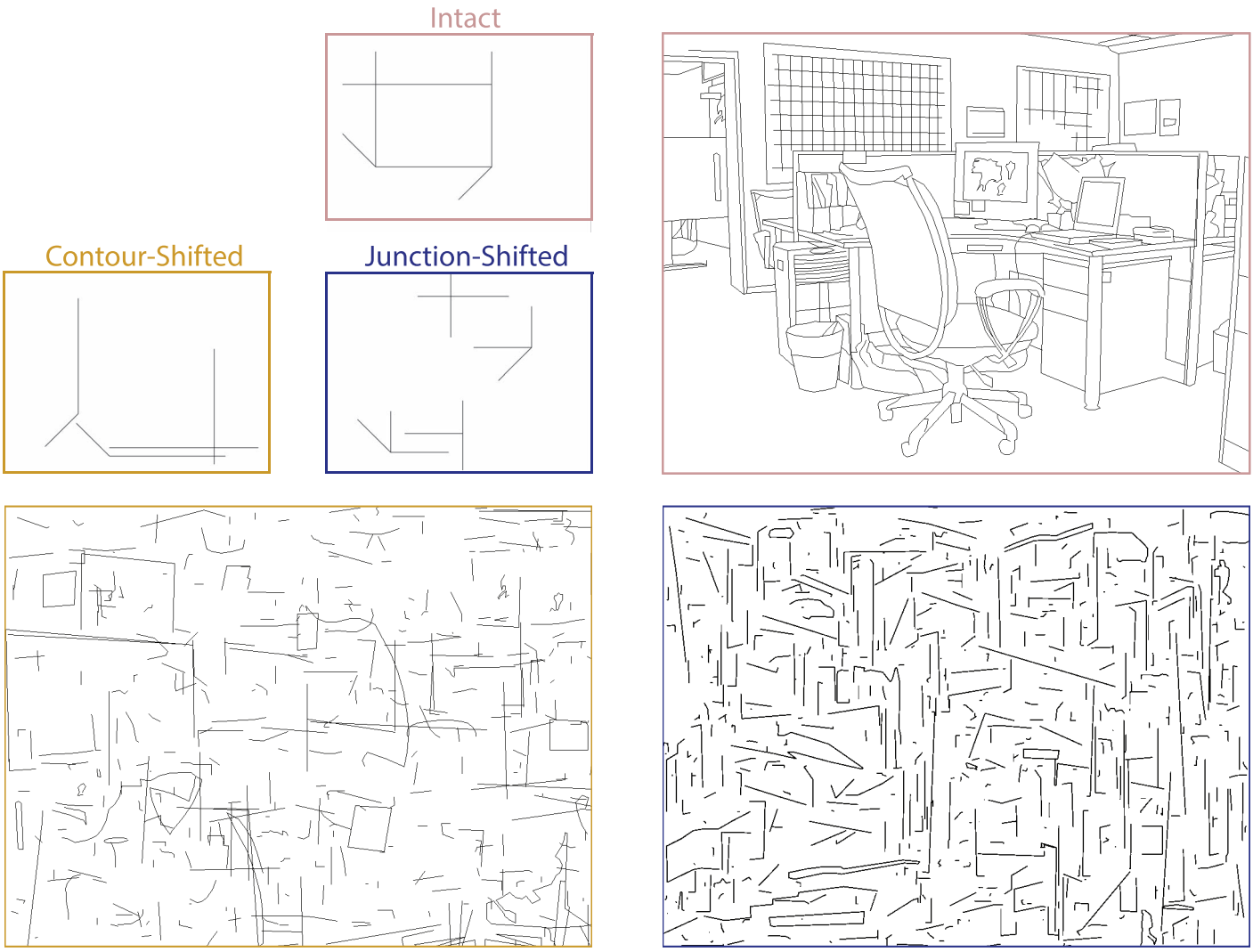

Figure 1. (Upper left) Three panes showing a simplified example of the image manipulations. The larger three panes (showing full scene examples) have the same layout as the smaller three panes: (upper right) intact line drawing; (lower left) contour-shifted; (lower right) junction-shifted.

The three image conditions were: (1) intact line drawings, (2) contour-shifted, and (3) junction-shifted.

The intact line drawings are the original line drawings of the artists (see Figure 1, top left), shown full-screen at a resolution of 1,024 by 768 pixels. When the artists traced a scene on a graphics tablet, a vector representation of the line drawing was obtained (Walther et al., 2011). A contour was defined as the line created from the point where an artist pressed down the graphics pen until the point where she or he lifted the pen. The artists were given the instruction: "For every image, please annotate all important and salient lines, including closed loops (e.g., boundary of a monitor) and open lines (e.g., boundaries of a road). Our requirement is that, by looking only at the annotated line drawings, a human observer can recognize the scene and salient objects within the image." For the contour-shifted images, all contours were randomly translated such that the contours were still entirely contained within the image. Contours had a root mean squared translation of 359 pixels, or an average unsigned translation of 242 pixels horizontally and 172 pixels vertically, with an average signed translation of 3 pixels horizontally and −4 pixels vertically. See Figure 1, lower right, for an example stimulus. Because each contour was translated independently, the original junctions of the image were completely destroyed, and new junctions were created, with the exception of junctions caused by self-intersection, which were not destroyed and were randomly translated with the contour (in total, 11% of junctions were due to self-intersections, and thus were still present when the contours were translated). Contours were only translated, never rotated or sheared. This process was repeated independently for each participant, so, with

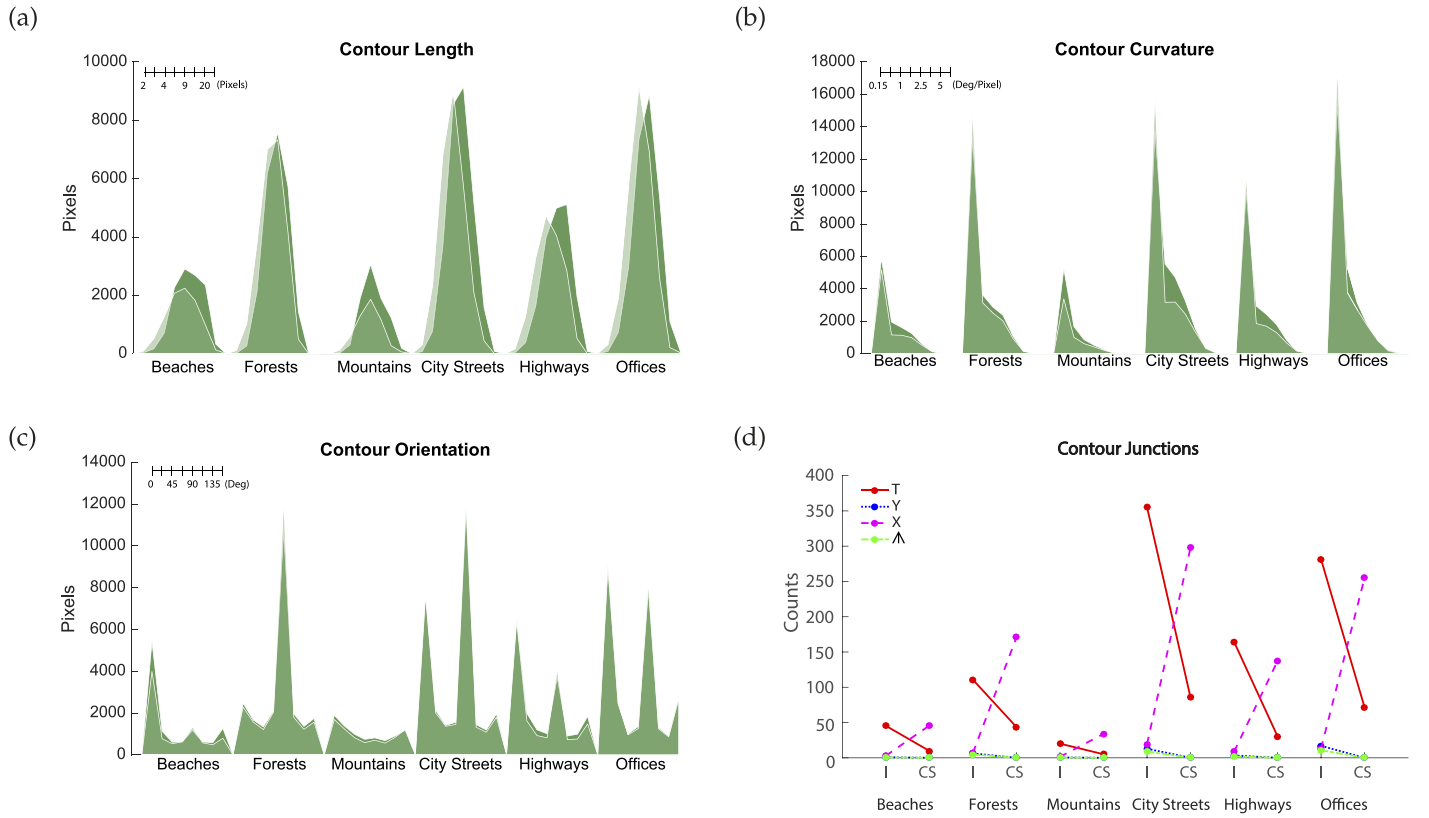

Figure 2. Distributions of contour features: (a) contour length, (b) contour curvature, (c) contour orientation, and (d) contour junctions. In (a), (b), and (c), the dark green distributions are from intact scenes, and light green are from the junction shifted scenes. Distributions are not shown for the contour-shifted scenes, as they are identical to the intact scenes. Note that in (a) contour length is shown on a log scale. In (d), for each category, the count for each junction type is shown for the intact scenes (I) and contour-shifted scenes (CS). For the junction-shifted scenes, the junctions counts are identical to the intact scenes.

high probability, no two participants saw the same identical contour-shifted images.

To create the junction-shifted condition, we first located all junctions between two contours (locations where the contours intersect). If a contour was involved in more than one junction, we located the midpoint between two adjacent junctions and broke the contour at that point. The result was a set of shorter contours, each involved in at most one junction. These junctions were then randomly translated within the image (as with shuffling contours, maintaining the orientation). When placing the broken-up contour sections, we ensured that no new junctions were created by first placing the largest segment at a random location. We then positioned the second-largest segment and randomly positioned it with at least a 3-pixel margin from the previously placed first segment. We continued by taking the largest remaining junction and randomly choosing a position in the image that did not bring that junction within three pixels of any previously placed junction. The result is that junctions had a root mean squared translation of 384 pixels on average, or an average unsigned translation of 256 pixels horizontally and 180 pixels vertically, with an average signed translation of 0 pixels horizontally and −15 pixels vertically. A result of this procedure for one scene is shown in Figure 1, lower left. As with the contour-shifted images, this process was repeated for each participant, resulting in a new set of stimuli for each participant.

These manipulations affected some of the distributions of contour features. The contour-shifted images have identical contour length, orientation, and curvature distributions, while the contour junction distributions are changed. The junction-shifted images have identical contour junction distributions, while the contour length, orientation, and curvature distributions have been modified. The changed distributions can be seen in Figure 2. Generally, the histograms for contour length have been shifted down toward shorter contours. Curvature distributions show fewer high-curvature contours. Orientation histograms are largely unchanged. The changes due to the shuffling of the contours are dependent upon the artist choice of how to draw the lines.

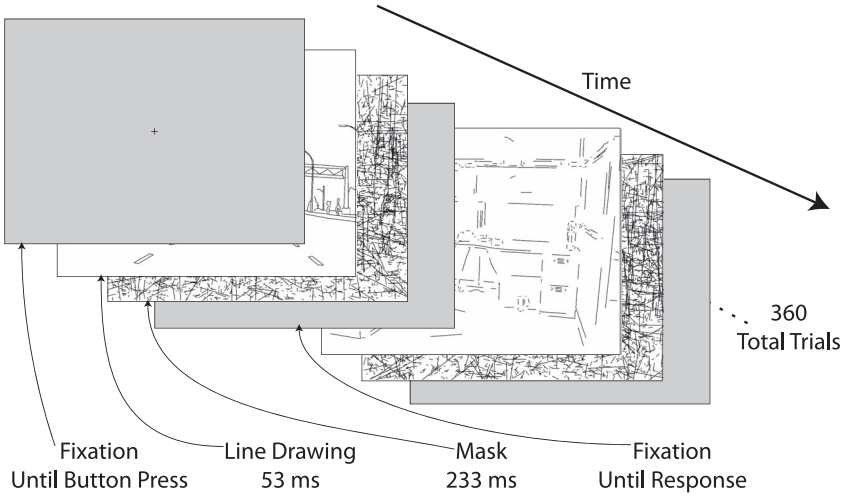

Figure 3. A schematic of the experiment. The sequences consisted of a stimulus (53 ms in the test phase, longer in the training and ramping phases) followed by a mask (500 ms) and then a blank screen until the participant responded with a key press.

Apparatus

The experiment was conducted using PCs running Windows 7 using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) on MATLAB R2014b (MathWorks, Natick, MA). The displays were CRT monitors running at 120 Hz at a resolution of 1,024 by 768 pixels.

Participants sat roughly 57 cm away from the display, in a dark room. Head position was unconstrained. The stimuli were shown at full screen, subtending approximately 39° of visual angle.

Design and procedure

The experiment was divided into three phases: Training, Ramping, and Test.

In each phase, participants were asked to decide which scene category they saw. The mapping between keys (S, D, F, J, K, and L) and categories was randomized for each participant and kept the same for each phase of the experiment. Prior to each phase of the experiment, the pairing of categories with keys was displayed to familiarize/re-familiarize participants with their mapping. Prior to the training phase, the participants memorized the mapping between key and category. If they expressed difficulty with the memorization, the experimenter suggested rehearsing the category name while pressing the corresponding finger, starting with the left-most response finger and moving toward the right. After the first training trial, but prior to giving a response, several participants asked to see the key mapping again. The experiment was restarted and they were allowed to study the key mapping again and begin the experiment. No participant asked to see the key mapping after the first training trial, and none expressed difficulty remembering the key mapping throughout the experiment.

In the training phase, the participants were shown a fixation mark until they pressed any key to begin the experiment. Then they were shown an intact line drawing for 233 ms. This was immediately followed by a mask for 500 ms. The mask was created by sampling contour segments from the set of contours of all line drawings in all categories. After the mask, the screen was blank, with the exception of a small central fixation mark, until the participant responded with a key press. Following the key press, feedback was given. A high tone indicated a correct response, and a low tone indicated an incorrect response. The training phase terminated when the participant responded correctly on at least 17 of the last 18 trials, or a maximum of 72 trials.

The ramping phase was similar to the training phase, with the exception that the stimulus duration was decreased as the phase progressed. For the first four trials, the duration was 200 ms. Every four trials, the duration was decreased by two frames (17 ms). The final two trials had a stimulus duration of 33 ms, resulting in a total of 54 trials. As in the training phase, only intact line drawings were shown, and feedback was given. The mask was always displayed for 500 ms.

In the test phase, a new set of stimuli was presented to avoid effects of familiarity with specific scenes. Every image was only ever shown once during the test phase, either in the intact, the contour-shifted, or the junction-shifted conditions. Conditions were intermixed randomly. Stimulus onset asynchrony (SOA) was fixed to 53 ms, and there was no feedback. In this phase, there were 20 line drawings per condition per category, for a total of 360 trials.

A schematic of the experiment is shown in Figure 3.

The stimulus duration was chosen so that in the intact case, participants would make at least some

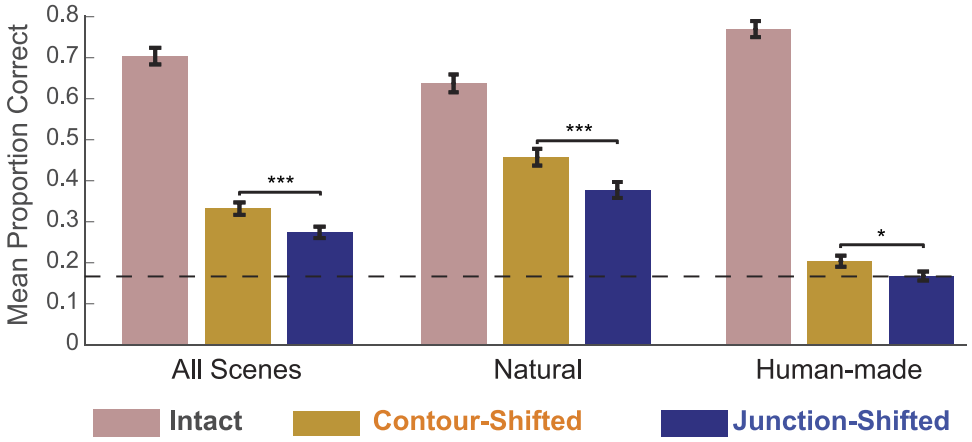

Figure 4. Proportion correct categorization performance averaged over the 19 participants in Experiment 1, for (left) all scenes, (middle) natural scenes, and (right) human-made scenes. The * denotes significance at the $p < 0.05$ level, and *** for $p < 0.001$.

classification errors. In the manipulated images, where scene information is disrupted, we expect performance to be worse than for the intact scenes. Since errors are expected in all conditions, we are able to examine these errors in confusion matrices and determine if the types of errors were different for the different conditions. To test this, we computed the correlation coefficients of the off-diagonal elements of each pair of confusion matrices. Additionally, we are interested in determining if the different image manipulations lead to an overall performance difference from one another. We compared overall performance (proportion correct) between those two conditions. We used a paired samples *t* test to test for significant performance differences between the two conditions with manipulated images across participants.

Results

Participants performed above chance (16.7%) in all three conditions (see Figure 4). Participants performed worse for both contour-shifted (33.2%) and junction-shifted (27.4%) than for intact scenes (70.4%). Performance in the contour-shifted condition was significantly better than in the junction-shifted condition (paired-samples *t* test, $t = -4.04$, $df = 18$, $p = 7.6 \times 10^{-4}$).

The participants' confusions (combined across all 19 participants) can be seen in Figure 5. The confusion matrices reveal a strong disadvantage for the categorization of human-made scenes, reducing the performance to chance in the junction-shifted condition. This preference does not appear in the intact images. The confusion matrices reveal that this failure to classify

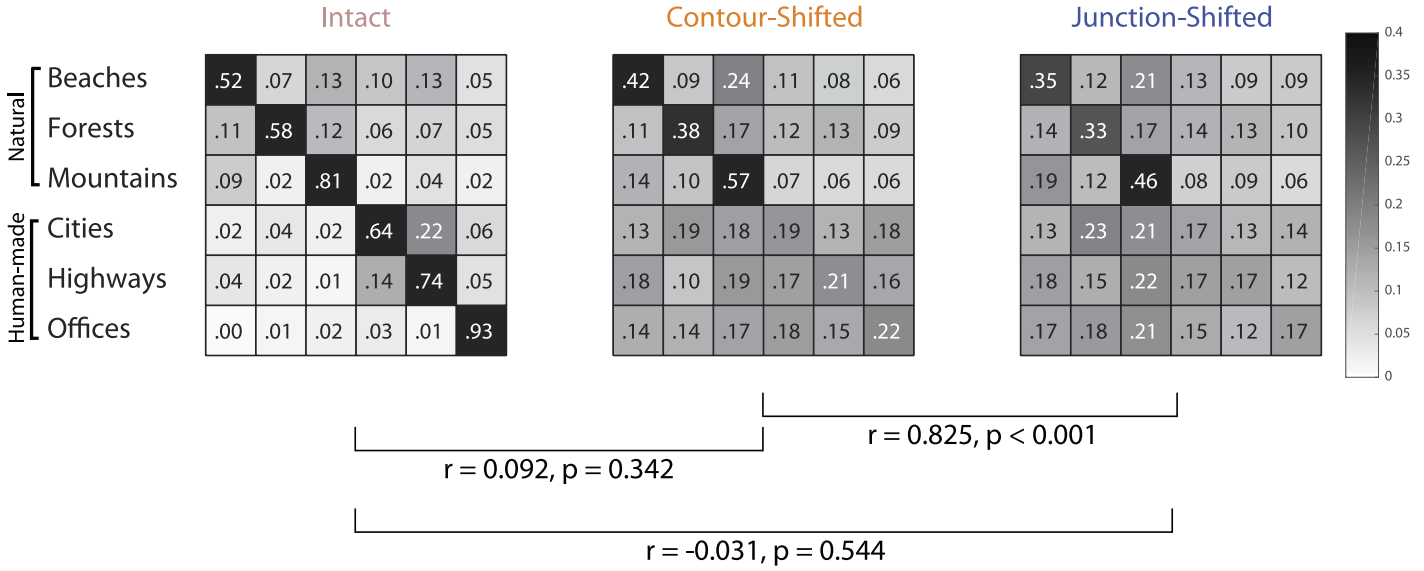

Figure 5. Confusion matrices for Experiment 1. Rows correspond to ground truth, and columns are the participants' responses. Error correlations (shown below the confusion matrices) are computed using only the off-diagonal entries.

human-made scenes is not a result of the participant classifying the scene as a human-made scene and then selecting randomly from the three human-made scene classes. Instead, participants simply respond randomly for human-made scenes, evenly distributing errors across all scene classes. Errors (off-diagonal entries in the confusion matrices) in the intact scenes were not correlated with the errors in either the contour-shifted condition ($r = 0.09$, $p = 0.34$) or the junction-shifted condition ($r = –0.03$, $p = 0.54$). The two experimental conditions did have a high error correlation with each other ($r = 0.83$, $p < 0.001$); the high error correlation suggest that there was no difference in the strategy when viewing either of the shifted scenes.

Due to the difference between performance in the natural scenes and the human-made scenes, we looked at the performance difference between the different conditions separately human-made or natural scenes (see Figure 4 middle and right). First, of note, is that for intact scenes. Performance for both contour-shifted and junction-shifted was highest for the human-made scenes, with a proportion correct = 0.769 versus 0.639 for the natural scenes. However, as mentioned, performance for both the contour-shifted and junction-shifted conditions was highest at 0.458 and 0.379, respectively, for natural scenes, versus 0.205 and 0.169 for human-made scenes. For human-made scenes, shifting the junctions resulted in performance at chance level, suggesting that junctions and their spatial locations within the image are very important for scene classification, specifically for human-made scenes.

Shifting either contours or junctions clearly hinders classification performance, but there is still enough information present to classify the scenes above chance in natural scenes. By disrupting the spatial layout of the scene, all of the spatial relationships between contours, such as junction opposedness and contour parallelism, were disrupted (with the exception of the two contour segments at a junction). Only the histograms of contour features or junction features were maintained. We wish to determine how much of the performance gap between the intact line drawings and the shifted line drawings was due to the removal of these spatial relationships.

Experiment 2

Here, we directly tested the importance of junctions and the spatial relationships between either junctions or between contour segments in the classification of real-world scenes. Beginning with a line drawing, we either removed the junctions or the portions of contours between junctions, resulting in two complementary half-images, each containing half of the contours of the original line drawing (similar to the contour-deleted images of (Biederman & Cooper, 1991).

Methods

Participants

Twenty-three undergraduates (18 female, five male) participated for credit for psychology courses. Ages ranged from 19 to 24 years, with a mean of 19.8. All participants had normal or corrected-to-normal vision. The experiment was approved by the University of Toronto REB under the same protocol as Experiment 1. Participants all gave written informed consent prior to participation.

Stimuli

Stimuli were constructed from the same set of line drawings as in Experiment 1.

The original line drawings were manipulated in two different ways, one version removing contours at and around junctions, called the *junctions-removed* condition, and the other version removing contiguous contour sections in-between junctions, called the *middle-removed* condition. To this end, locations of junctions were determined as points where two contours crossed, points where the angle between adjacent contour elements was less than 130°, and endpoints of contours. We then marked breakpoints at 25% and at 75% of the distance between two junctions. This allowed us to remove junctions by removing, from all contours, the sections between 0 and 25% and between 75% and 100% of the distance between two adjacent junctions and to remove middle parts by erasing the sections between 25% and 75% of the distance (see Figure 6, top right for an illustration). Note that both manipulated versions are complements of each other, containing half of the total contour content, only having overlap at the end points of the middle segments and junction segments. An example of these manipulations on an office scene is shown in Figure 6. These manipulations change the distributions of contour features. For example, the number of pixels is halved, so for each feature, the distribution is scaled. Additionally, because the contours were separated into smaller pieces, the length distributions are shifted. A comparison of the distributions of contour features for the intact and the manipulated image are in Figure 7. There is no figure for how the junctions were changed, because either they were maintained (and thus identical to those in the intact scene) or completely removed. By design, the length distributions are shifted to the left (shorter) because the contours were separated into smaller

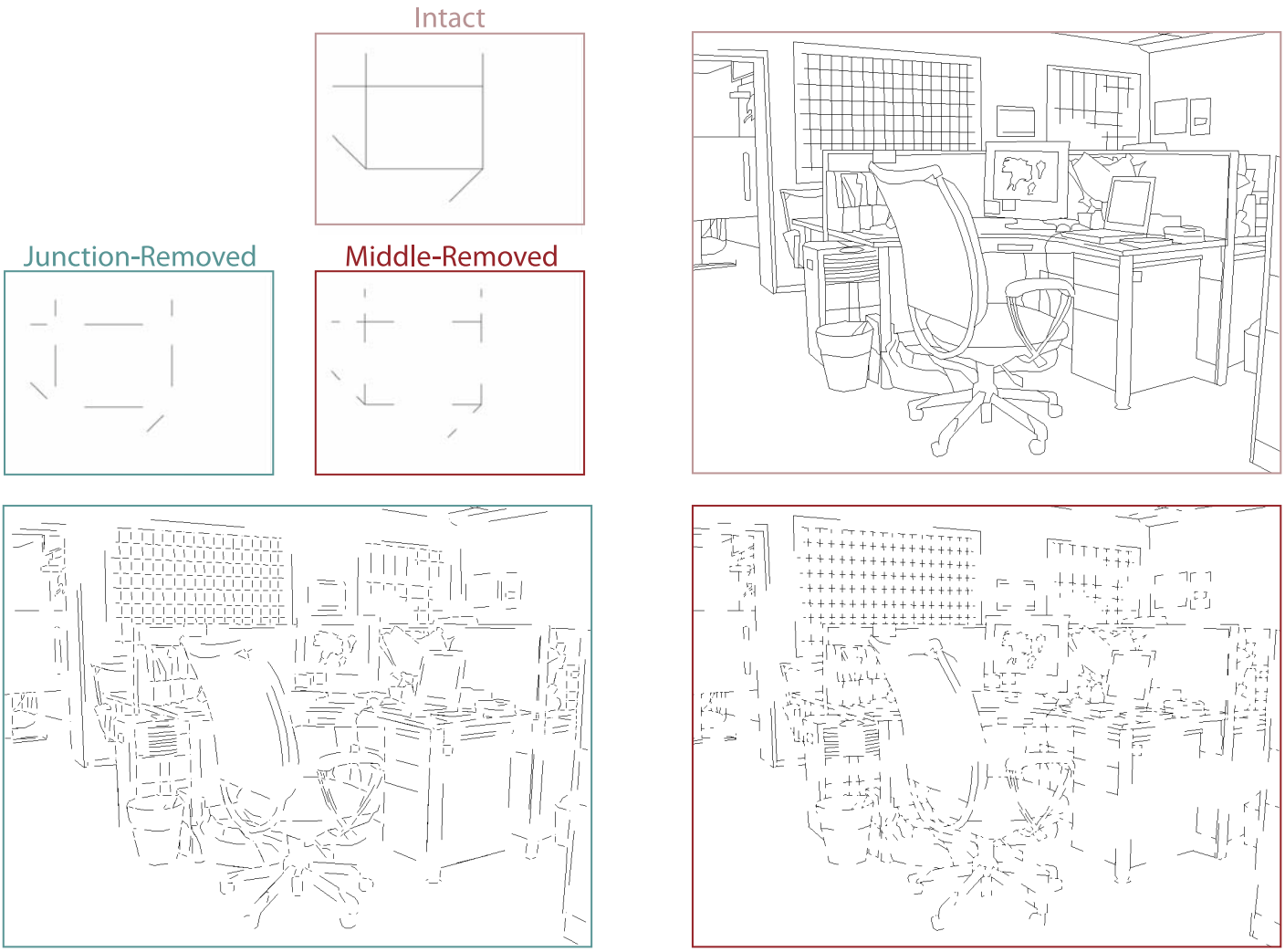

Figure 6. (Upper left) Simple example of the experimental conditions. (Upper right) An intact office scene. (Lower left) The office scene with junctions-removed. (Lower right) The office scene with middle-removed.

pieces. Also, with fewer pixels than in the intact scenes, the overall distributions are smaller. The changes between the distributions are similar for all scene categories.

Unlike the stimuli in Experiment 1, these stimuli are not randomly shifted in any way. Thus, the stimulus generation is deterministic, and the images will not differ between two participants who are shown the same scene exemplar in the same experimental condition.

Design and procedure

The apparatus, design, and procedure were identical to Experiment 1, except that in the test phase, we used the experimental conditions junctions-removed and middle-removed. As in Experiment 1, no image from the ramping and training was used in the test phase, and no image in the test phase was used in more than one of the three experimental conditions. Data analysis was performed the same way as for Experiment 1.

Results

As in Experiment 1, participants performed above chance (16.7%) in all three conditions (see Figure 8). Intact scenes were most easily classified, with participants performing at 68.7% correct. The scenes with the junctions-removed were classified at 47.5% correct. Scenes with the middle-removed were classified at 42.2% correct. This difference in proportion correct is statistically significant (paired-samples t test, $t = 4.53$, $df = 22$, $p = 1.65 \times 10^{-4}$). Individually, 18 of the 23 participants performed better in the junctions-removed condition than in the middle-removed condition.

In addition to analyzing proportion correct, we looked at the patterns of errors in each of the conditions. Figure 9 shows the confusion matrices for

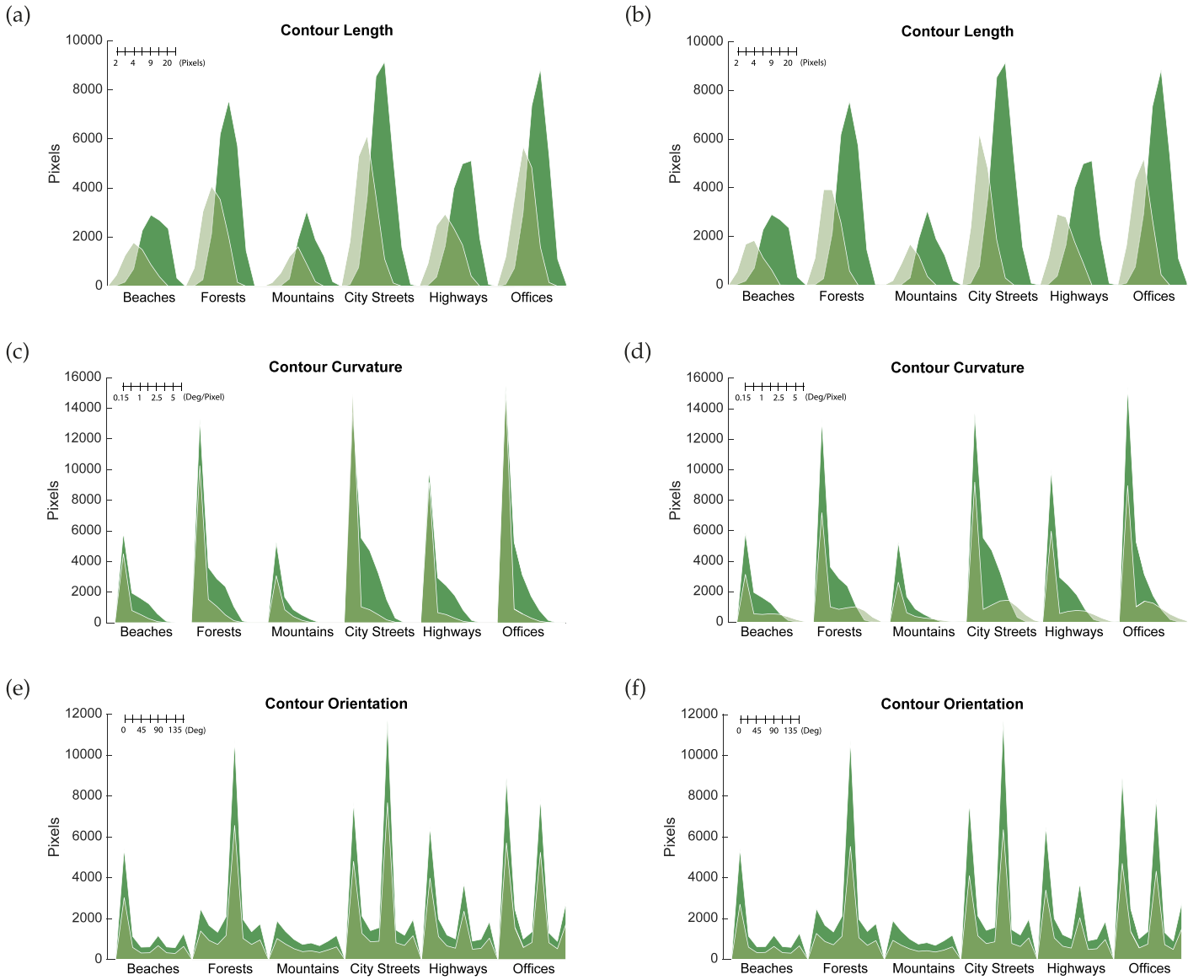

Figure 7. Distributions of contour features: (a) and (b) contour length, (c) and (d) contour curvature, (e) and (f) contour orientation. The dark green distributions are from intact scenes, and light green are from the junctions-removed scenes (a, c, and e) or the middle-removed scenes (b, d, and f). Note that in (a) and (b) contour length is shown on a log scale. Unlike in Experiment 1, the contour junction distributions are not shown, as there are either no junctions are present or they are identical to the intact scenes.

each condition. For intact scenes, we see a strong diagonal, signifying that each scene category was usually correctly classified. The most common mistakes were misclassifying "cities" as "highways." In both of the manipulated conditions, there was a weaker diagonal. The error patterns (off-diagonal entries in the confusion matrices) were fairly similar between the two experimental conditions, and they introduce some errors that were uncommon with the intact scenes. For example, we see that "offices" are much more commonly called "cities" in the experimentally manipulated scenes than in the intact scenes. The types of errors between the difference conditions were all very similar. The correlation of the error patterns was significant between all conditions (intact/middle-removed: $r = 0.59$, $p < 0.01$; intact/junctions-removed, $r = 0.77$, $p < 0.001$; middle-removed/junctions-removed, $r = 0.91$, $p < 0.001$). Also note that for this experiment there is a strong diagonal for the human-made scenes, unlike in the first experiment. The improvement in overall performance in the experimental conditions over the performance in the first experiment is driven almost entirely by the improvement for human-made scenes. This suggests that the spatial layout of the scene is more important for human-made scenes, whereas for natural scenes the histograms of contour or junction features carry almost as much information.

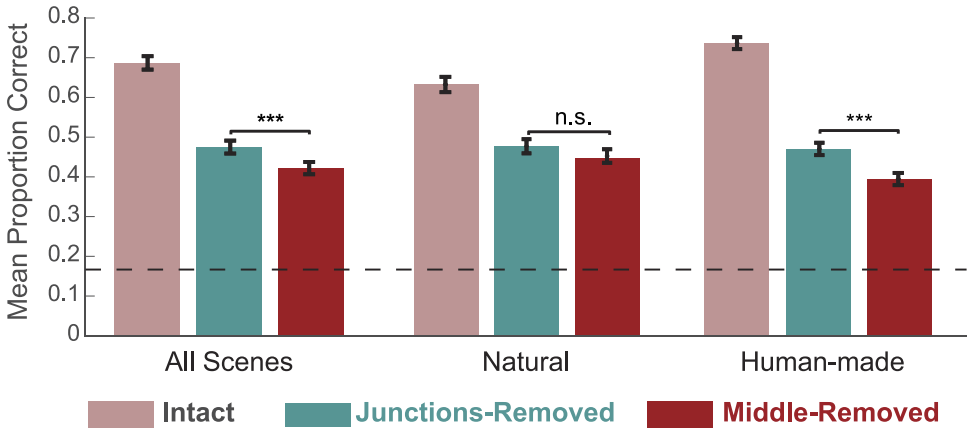

Figure 8. Categorization accuracy in Experiment 2 for (left) all scenes, (middle) natural scenes, and (right) human-made scenes. The *** denotes significant with $p < 0.001$.

In Figure 8 (middle and right), we see that for the junctions-removed condition, the performance is almost equal between human-made (proportion correct = 0.478) and naturally occurring scenes (0.473). The difference, made apparent by the figure, is that removing the middle segments appeared to have a larger effect in the human-made scenes than in the naturally occurring scenes. In fact, the significant difference we found between the junctions-removed and the middle-removed images was almost entirely driven by the human-made scenes (paired-samples t test, $t(22) = -4.66$, $p = 1.2 \times 10^{-4}$), although the non-significant effect is in the same direction with naturally occurring scenes (paired-samples t test, $t(22) = -1.67$, $p = 0.11$). One hypothesis is that in the human-made scenes there are more parallel segments that are being removed in the middle-removed condition; however, it is difficult to measure this, due to computational complexity. For naturally occurring scenes, this effect would be prominent in the forest scenes, but less common in beach and mountain scenes. We also hypothesize that it is more difficult for the visual system to complete the contours between the junctions in our scenes. One possible reason for this difficulty is that junctions have a more ambiguous orientation. To address this possibility, we computed the variance in local orientation over each junction or middle segment. The junctions have a higher variation in their orientation statistics (mean orientation variance = 8.2°) than the middle segments (mean orientation variance = 2.4°). With unlimited viewing duration it may feel that completing the contours is trivial. The more ambiguous

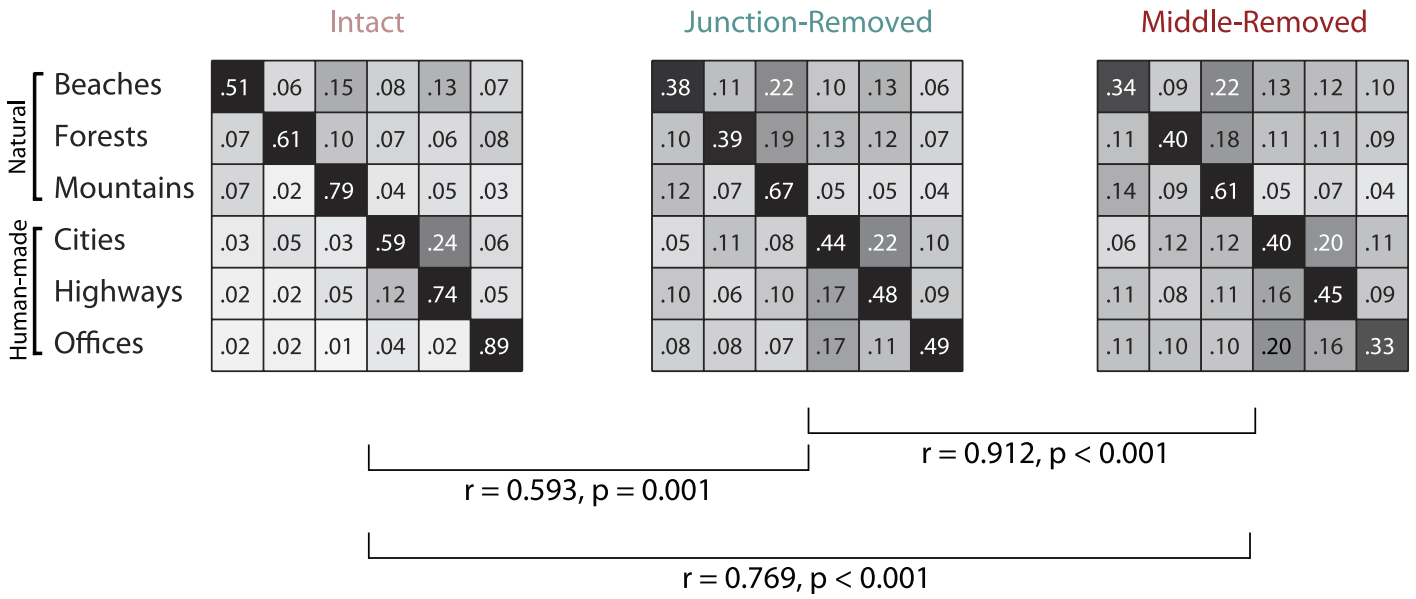

Figure 9. Confusion matrices for Experiment 2. Rows correspond to ground truth, and columns to the participants' responses. Below the confusion matrices are error correlations. Error correlations are computed using only off-diagonal entries.

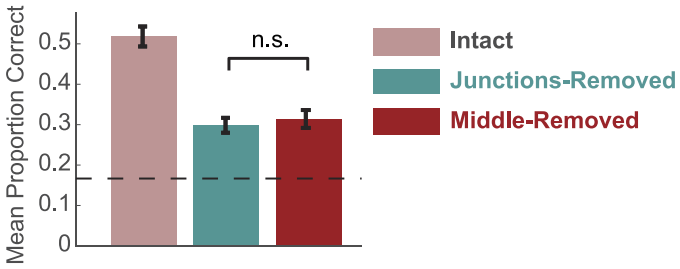

Figure 10. Mean categorization performance (proportion correct averaged across 16 participants) for Experiment 3.

orientation at junctions may contribute to the difference in accuracy for brief and masked presentations by making rapid contour interpolation more difficult.

Experiment 3

Self-reports from several participants in Experiment 2 indicated that they believed the images with "longer" contours were easier to classify. The middle segments are, in fact, physically longer than the junction segments, on average, because contour end points are treated as junctions. As a result, there are small segments equal to 25% of the length to the next junction at the end of contours. In this experiment, we equate the average contour length by adjusting the break-point on the contours so that the physical length is equated. Note that due to this manipulation, total pixel count is no longer equated between the two manipulations.

Method

Participants

Sixteen undergraduates (10 female, six male) from introductory level psychology courses at the University of Toronto participated for course credit. Ages ranged from 18 to 21 years, with a mean of 18.4. Participants gave informed consent prior to participation. All participants had normal or corrected-to-normal vision.

Stimuli

The stimuli are the same as those in Experiment 2 with a change in the location at which the intact contours were separated. Here, the segment between two junctions was not split at one quarter and three quarters of the entire length of the segment. Instead, we computed the distance that would result in the means of the length distributions being equal for the two experimental conditions. For beaches, this means that a middle segment was 46% of the segment between two junctions as opposed to 50% in Experiment 2. For all other categories, a middle segment was 42% of the length of the segment.

Because the middle-removed condition removed a smaller total portion of the line, the number of pixels in the image was no longer controlled. This means there were more black pixels in the middle-removed than in the junctions-removed images.

Design and procedure

The apparatus and procedure were identical to that of Experiment 2, and was approved under the same REB protocol. The same analysis used in Experiments 1 and 2 will be used for this experiment.

Results

All participants performed best in the intact condition, with average proportion correct = 0.52 (see Figure 10). Performance in the middle-removed condition was the same as in the junctions-removed condition (average proportion correct = 0.31 and 0.30, respectively, paired-samples t test, $t = -1.37$, $df = 15$, $p = 0.19$). Even with more of the contour content visible, participants did not perform significantly better in the junctions-removed condition, suggesting that the middle segments contain important information that aids participants in categorizing scenes. The participants for Experiment 3 were recruited later in the semester than the participants for Experiment 2. We generally observe lower performance later in the semester, which may be due to self-selection effects in the student population. Therefore, a direct comparison of the two experiments' performance in the intact condition is not justified. However, notice that even though the performance on the intact condition was lower in Experiment 3, relative to Experiment 2, the relative performance difference between the intact condition and the manipulated conditions in each experiment is roughly the same (near 0.3). In addition to data being collected at a different time in the semester, Experiment 3 may be perceived as more difficult overall, because fewer pixels were presented in the middle-removed condition of that experiment. This can result in the participants becoming slightly less motivated.

General discussion

Our results show that shuffling junctions within a scene results in worse performance than shuffling entire contours. This suggests that distributions of junctions are not solely responsible for scene classification. While

performance was higher when entire contours were shuffled, it was still very low, suggesting that the distribution of contour statistics is insufficient for rapid scene classification. We hypothesize that spatial layout plays an important role in rapid scene perception, because it reveals the spatial relationships between surfaces and objects, which are bounded by the contours and their junctions, especially for human-made scenes. This is demonstrated by the near floor performance for human made scenes when contours or junctions are shifted within the scene. This is consistent with previous work using photographs, where it was shown that disrupting Fourier phase while maintaining the amplitude spectrum hinders scene classification (Loschky et al., 2007).

Additionally, our results show that shifting the contours within the scene hinders performance more for human-made scenes than for natural scenes. One popular model of scene classification has argued that a superordinate classification (e.g., natural vs. human-made) occurs prior to classification at a basic level (Oliva & Torralba, 2001). Loschky and Larson (2008) argue that this is because simple features must be localized well enough to allow for grouping into larger configurations. Before a basic level classification can be made, these configurations allow for the natural/human-made classification. Thus, if the classification process is interrupted, or if there is not sufficient information for basic level categorization, the visual system may only be able to make classifications at the superordinate level.

We were surprised to find that removing middle segments hurt performance more than removing junctions in Experiment 2. Experiment 3 equated the means of the contour length distributions, at the cost of an imbalance in the number of contour pixels in the images, and images containing junctions were still not easier to classify. Middle segments and their spatial relationships appear to be more important for human-made than natural scenes, presumably due to the high degree of geometric design inherent in human-made artifacts.

These results are not what we would expect based upon the results of Biederman (1987) and Attneave (1954), who suggest that junctions and points of high curvature may be the most important features for the classification of objects. DeWinter and Wagemans (2008) directly tested the hypotheses of Attneave (1954), by creating closed shapes that connected points of maximum curvature, or points halfway between two points of high curvature with straight lines. They found that objects that were connected by points of high curvature were more easily recognized, but that there was considerable variability between shapes. Biederman and Cooper (1991) argued that junctions are useful because a missing contour between two vertices can be filled in accurately through a local process, while it is not easy for a local process to fill in a missing junction. An example of this difficulty is that it is not clear whether a missing L-junction is actually an L-junction, or if, instead, one contour extends farther, resulting in a T-, Y-, or Arrow-junction. Additionally, the junctions are useful for locating part boundaries (Hoffman & Richards, 1984), and this allows an observer to determine where object parts are located and how they interact (Biederman, 1987).

Other previous work has led to a different conclusion, apparently consistent with our results. Kennedy and Domander (1985) found that maintaining middle segments led to better classification of objects than maintaining junctions. Green and Courtis (1966) created a demonstration using Attneave's cat (Attneave, 1954), showing that versions of the cat with only the line segments around the junctions were subjectively just as easily categorized as with only the line segments around the middle. Similar to Kennedy and Domander (1985) and directly related to our Experiment 2, Panis et al. (2008) had participants categorize fragmented objects, where the fragments were located at salient points (which tended to be at locations of high curvature) or at points at the midpoint between salient points. Participants had an easier time identifying objects with fragments at the midpoints, and needed roughly 33% larger fragments in order to achieve the same performance in the salient point condition.

This previous work was related to the categorization of objects. Often the shapes are very simple with few internal contours. The exceptions to this, Kennedy and Domander (1985) and the demonstration in Green and Courtis (1966), used more complex shapes with more internal contours. Similarly, we are working with real-world scenes, where objects are not isolated from the background and from each other. In a complex scene, with many objects and surfaces, junctions have been shown to help resolve the three-dimensional structure of the scene (Anderson & Julesz, 1995; Clowes, 1971; Guzmán, 1968; Huffman, 1971; Mackworth, 1976; Malik, 1987). So, on the one hand, we have the suggestion that the interpretation of a complex scene involves analysis of the junctions, but on the other hand junctions appear to be less useful when objects are more complex. We find that the relationships between contour middle segments dominate junctions in scene categorization. We hypothesize this may be due to ambiguity about which line segments should connect to one another. Biederman and Cooper (1991) suggested that junctions are useful because they more easily allow for contour completion; in our stimuli, the completion problem is more ambiguous when only junctions are present. In this situation, it appears that the junctions are less important than the segments between the

junctions. Enns and Rensink (1991) looked at the role of junctions in interpreting the 3D structure of line drawings of simple objects. They showed that in order to rapidly determine the 3D orientation of a cube, the junctions needed to be physically connected. Completion required a much slower, and error prone, process. Their stimuli were simpler than scenes, so the connections of the different junctions were less ambiguous than in our full scenes. Still, participants seemed unable to rapidly fill in the missing contours, suggesting that presentation duration may also play a role in the amount of contour completion that can occur.

What aspects of middle segments help convey category-relevant information? Panis et al. (2008) observed that object recognition was more accurate when based on middle line segments than line segments around junctions.[1] There is an abundance of possible spatial relationships between contours that affect perceptual grouping of contour elements, and could help explain our results. The contour integration literature has shown that the distance between elements and orientation of one element relative to a neighbor (i.e., curvature) is useful (Field, Hayes, & Hess, 1993; Geisler, Perry, Super, & Gallogly, 2001). Contour segments can be related through collinearity, as in the Gestalt principle of good continuation, or through parallelism, as with the Gestalt principle of symmetry (Wertheimer, 1938; Koffka, 1935; Metzger, Spillmann, Lehar, Stromeyer, & Wertheimer, 2006; Wagemans, Elder, et al., 2012; Wagemans, Feldman, et al., 2012).

When inspecting the junctions-removed images from Experiment 2, we noticed that many opposing pairs of contours were retained and presumably aided in perception. Even though removal of junctions creates uncertainty about the local relationships between surfaces, middle segments may reveal longer-range relations within an image, such as the relative position of these elements. Inter-contour spatial relationships were destroyed by the shuffling in Experiment 1, and this may account for the significant loss of performance observed there. In Experiment 2, the middle segments may provide stronger inter-contour grouping cues than the junctions alone. The relative position of image features (such as middle segments) appears to be of importance for intercontour grouping, which is important for determining scene category.

Conclusion

Scene perception relies on a range of features that encode the spatial structure of a scene. Summary statistics of junction information alone is insufficient to fully explain scene categorization. Rather, spatial relationships between junctions play an important role. Furthermore, the middle segments of contours between junctions are more important for interpreting spatial structure of complex scenes than simple objects. In simple objects, contours define the outer boundary of the shape of the object or self-occlusion boundaries of object parts. In complex arrangements of objects and surfaces, contours additionally define spatial relationships between parts of a scene. Longer-range interactions between elongated sections of contours, such as parallelism and symmetry, may therefore play a more important role in such complex stimuli. More research is needed to better understand the role that different structural image features play in grouping, organizing, and recognizing visual information in complex, real-world scenes.

Keywords: scene perception, scene categorization, contour junctions, perceptual grouping

Acknowledgments

DBW was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2015-06696, a Sony Faculty Research Award, and SSHRC Insight Development Grant 430-2017-01189. DBW was also supported by NSERC Discovery Grant 498390 and Canadian Foundation for Innovation 32896. SD acknowledges funding through NSERC and AJ acknowledges NSERC Discovery Grant support.

Commercial relationships: none.
Corresponding author: John Wilder.
Email: jdwilder@cs.toronto.edu.
Address: University of Toronto, Toronto, Ontario, Canada.

Footnote

[1] They found the opposite pattern of results when using dots rather than line segments: A single dot at the high curvature points was more informative than a single dot along the middle segments.

References

Anderson, B. L., & Julesz, B. (1995). A theoretical analysis of illusory contour formation in stereopsis. *Psychological Review*, *102*(4), 705.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115.

Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, *23*(3), 393–419.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Clowes, M. B. (1971). On seeing things. *Artificial intelligence*, *2*(1), 79–116.

Craft, E., Schütze, H., Niebur, E., & Von Der Heydt, R. (2007). A neural model of figure–ground organization. *Journal of Neurophysiology*, *97*(6), 4310–4326.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Research*, *40*(16), 2187–2200.

De Winter, J., & Wagemans, J. (2008). The awakening of Attneave's sleeping cat: Identification of everyday objects on the basis of straight-line versions of outlines. *Perception*, *37*(2), 245–270.

Enns, J. T., & Rensink, R. A. (1991). Preattentive recovery of three-dimensional orientation from line drawings. *Psychological Review*, *98*(3), 335.

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local association field. *Vision research*, *33*(2), 173–193.

Geisler, W. S., Perry, J. S., Super, B., & Gallogly, D. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724.

Green, R., & Courtis, M. (1966). Information theory and figure perception: The metaphor that failed. *Acta Psychologica*, *25*, 12–35.

Guzmán, A. (1968). Decomposition of a visual scene into three-dimensional bodies. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part i* (pp. 291–304).

Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, *18*(1-3), 65–96.

Huffman, D. (1971). Impossible objects as nonsense sentences. *Machine Intelligence 6*, 295–324.

Kennedy, J. M., & Domander, R. (1985). Shape and contour: The points of maximum change are least useful for recognition. *Perception*, *14*(3), 367–370.

Koffka, K. (1935). *Principles of gestalt psychology, international library of psychology, philosophy and scientific method*. New York: Harcourt Brace.

Loschky, L. C., & Larson, A. M. (2008). Localized information is necessary for scene categorization, including the natural/man-made distinction. *Journal of Vision*, *8*(1):4, 1–9, http://doi.org/10.1167/8.1.4. [PubMed] [Article]

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1431.

Mackworth, A. K. (1976). Model-driven interpretation in intelligent vision systems. *Perception*, *5*(3), 349–370.

Malik, J. (1987). Interpreting line drawings of curved objects. *International Journal of Computer Vision*, *1*(1), 73–103.

Metzger, W., Spillmann, L. T., Lehar, S. T., Stromeyer, M. T., & Wertheimer, M. T. (2006). *Laws of seeing*. Cambridge, MA: MIT Press.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*(2), 176–210.

Oliva, A., & Torralba, A. (2001, 05). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. Retrieved from http://myaccess.library.utoronto.ca/login?url=https://search.proquest.com/docview/1113597138?accou (Copyright - Kluwer Academic Publishers 2001; Last updated - 2012-10-20)

Panis, S., De Winter, J., Vandekerckhove, J., & Wagemans, J. (2008). Identification of everyday objects on the basis of fragmented outline versions. *Perception*, *37*(2), 271–289.

Pelli, D. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10.

Sajda, P., & Finkel, L. H. (1995). Intermediate-level visual representations and the construction of surface perception. *Journal of Cognitive Neuroscience*, *7*(2), 267–291.

Thorpe, S., Fize, D., & Marlot, C. (1996, June 6). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: computation in neural systems*, *14*(3), 391–412.

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater bold activity. *PLoS One*, *8*(3), e58594.

VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655–668.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). A century of gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, *138*(6), 1218.

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, *108*(23), 9661–9666.

Walther, D. B., & Shen, D. (2014). Nonaccidental properties underlie human categorization of complex natural scenes. *Psychological Science,* https://doi.org/10.1177/0956797613512662.

Wertheimer, M. (1938). *Laws of organization in perceptual forms*. London, UK: Kegan Paul, Trench, Trubner & Company.

Wichmann, F. A., Drewes, J., Rosas, P., & Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision*, *10*(4):6, 1–27, https://doi.org/10.1167/10.4.6. [PubMed] [Article]