TECHNISCHE UNIVERSITÄT BERLIN
Faculty IV - Electrical Engineering and Computer Science
Institute of Computer Engineering and Microelectronics
Dept. Computational Psychology

First Examiner: Prof. Dr. Marianne Maertens
Second Examiner: Prof. Dr. Guillermo Gallego
Supervisor: Lynn Schmittwilken

Do the Presumed Mechanisms Underlying Human Edge Perception Translate to Natural Images?

Submitted by:
Anna Lucia Haverkamp
Matriculation Number: 391195
M.Sc. Computer Science

March 28, 2025

Statutory Declaration

I hereby declare that the thesis submitted is my own, unaided work, completed without any external help. Only the sources and resources listed were used. All passages taken from the sources and aids used, either unchanged or paraphrased, have been marked as such.
As for generative AI tools, I have used ChatGPT (OpenAI, GPT-4) and DeepL Write (DeepL SE) for checking my writing regarding grammar, spelling, punctuation and improving readability. I am fully responsible for the selection, adoption, and all results of the AI-generated output I use.
I have taken note of the Principles for Ensuring Good Research Practice at TU Berlin dated 15 February 2023. https://www.static.tu.berlin/fileadmin/www/10002457/K3-AMBl/Amtsblatt_2023/Amtliches_Mitteilungsblatt_Nr._16_vom_30.05.2023.pdf
I further declare that I have not submitted the thesis in the same or similar form to any other examination authority.

Berlin, March 28, 2025

.....

Signature

Abstract

Edges are fundamental visual features that define the boundaries of objects and play a crucial role in the early visual processing of humans. While many studies have investigated edge sensitivity using controlled stimuli, it remains unclear how these findings generalise to natural stimuli. This thesis examines the relationship between edge sensitivity for simple stimuli and contour perception of natural scenes by masking stimuli with 2D noise.

We compare the results from a 2-AFC task by Schmittwilken et al. (2024), in which participants detected the location of Cornsweet edges, with a contour-tracing task we conducted. The contour-tracing task involved participants using a drawing tablet to draw all visible contours in natural images from the Contour Image Database by Grigorescu et al. (2003). For both tasks, the stimuli were masked with noise patterns, consisting of three narrowband noises with different spatial frequencies (0.5, 3, 9 cpd) and three broadband noises (white, pink, brown) to investigate the spatial frequency selectivity of cells responsible for edge detection.

Findings reveal that pink noise and narrowband noise of 3 cpd had the strongest impact on both edge detection and contour perception, whereas low-frequency noise (0.5 cpd, brown) had minimal effects. These results suggest that the mechanisms underlying edge sensitivity in simple stimuli translate well to natural images. On the one hand, this observation underscores the relevance of controlled psychophysical studies for understanding natural vision. On the other hand, it also highlights the potential for studying visual processes using tasks that are more behaviourally relevant.

Abstract - German Version

Kanten sind wichtige visuelle Merkmale – sie definieren die Grenzen von Objekten und spielen eine entscheidende Rolle in der frühen visuellen Verarbeitung des Menschen. Da zahlreiche Studien Kantensensitivität nur mithilfe kontrollierter Reize untersucht haben, bleibt derzeit noch unklar, inwieweit sich diese Forschungsergebnisse auf natürliche Reize übertragen lassen. In dieser Arbeit untersuchen wir daher die Beziehung zwischen der Kantensensitivität für einfache, kontrollierte Stimuli und der Konturwahrnehmung in natürlichen Szenen. Dazu verwenden wir Stimuli, die mit 2D-Rauschen maskiert sind.

Wir vergleichen die Ergebnisse eines 2-AFC Experiments von Schmittwilken et al. (2024), in dem die Teilnehmer die Position von Cornsweet-Kanten bestimmen mussten, mit einer von uns durchgeführten Konturensegmentierungsaufgabe. In dieser Aufgabe zeichneten die Teilnehmer mithilfe eines Zeichentabletts alle sichtbaren Konturen in natürlichen Bildern ein. Die verwendeten Bilder stammen aus der Bilddatenbank von Grigorescu et al. (2003). In beiden Experimenten wurden die Stimuli mit verschiedenen Rauschmustern maskiert: Drei schmalbandige Rauschmuster mit unterschiedlichen Raumfrequenzen (0,5, 3 und 9 cpd) sowie drei breitbandige Rauschmuster (weißes, rosa und braunes Rauschen). Ziel war es, die räumliche Frequenzselektivität der für die Kantenerkennung verantwortlichen neuronalen Mechanismen zu untersuchen.

Unsere Ergebnisse zeigen, dass sowohl pinkes Rauschen als auch Rauschen mit einer Raumfrequenz von 3 cpd den größten Einfluss auf die Wahrnehmung von Kanten sowohl bei kontrollierten als auch bei natürlichen Stimuli haben. Im Gegensatz dazu zeigten niederfrequente Rauschmuster (0,5 cpd, braunes Rauschen) nur minimale Auswirkungen auf die Wahrnehmung. Diese Funde legen nahe, dass die Mechanismen, die der Kantensensitivität bei einfachen Stimuli zugrunde liegen, auch für die Verarbeitung natürlicher Bilder relevant sind.

Diese Erkenntnis unterstreicht zum einen die Bedeutung psychophysikalischer Studien mit kontrollierten Stimuli für die Erlangung wertvoller Erkenntnisse bezüglich des natürlichen Sehens. Zum anderen wird durch die Ergebnisse die Relevanz verhaltensnäherer Aufgabenstellungen für die Untersuchung visueller Prozesse unterstrichen.

Contents

1. Introduction

Edges have long been a critical feature in vision research. An edge is a discontinuity in luminance – i.e., the amount of light that reaches the eye – that typically occurs at the boundary of an object. The foundational work of Hubel and Wiesel (1962) on edge detection revealed that most cells in the early visual system respond selectively to edges of specific orientations within the visual field. This discovery highlighted the importance of edges in the initial stages of human visual processing. Since then, extensive research has been conducted to further investigate the mechanisms underlying edge detection.

An important advance in understanding these processes came from Campbell and Robson (1968). They found that, in addition to orientation and location (as shown by Hubel and Wiesel (1962)), cells in the early visual system are also selectively responsive to a limited range of spatial frequencies. Spatial frequency (SF) refers to the number of repeating sine waves within a certain distance to the retina. SF is typically measured in cycles per degree (cpd), i.e., the number of times the pattern repeats in one degree of the visual field.

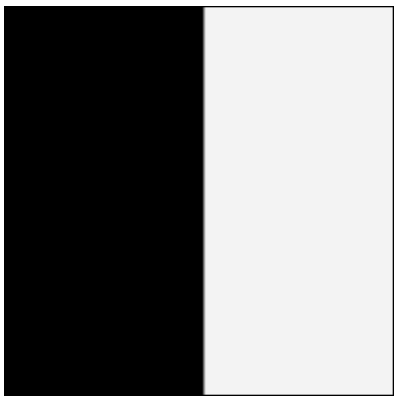
Further research explored which SFs are most critical for edge detection. Solomon and Pelli (1994) examined the impact of visual noise on edge perception, where noise refers to a random signal containing specific frequencies. When used to mask a stimulus, noise can hinder perception by interfering with similar frequencies present in the stimulus itself. The degree to which visual noise disrupts perception reflects the visual system's sensitivity to specific SFs. Their study demonstrated that noise with a SF of 3 cpd was especially effective in impairing perception. This suggests that cells responsible for edge detection are most sensitive to this particular frequency, to a greater extent than to either higher or lower frequencies. This finding is corroborated by additional research, which also identified peak sensitivity for edge detection around 3 cpd (Shapley and Tolhurst, 1973; Foster et al., 1985).

Collectively, these studies support the current assumption that the visual system contains multiple "SF-selective channels" (Graham, 2011) with SFs around 3 cpd playing a particularly crucial role in human edge perception. The term *channel* refers to a set of cells that exhibit similar response properties. Nevertheless, a recurrent theme among these studies is that they predominantly employed highly controlled, artificial stimuli – such as sinusoidal gratings, step edges, or bars – to examine edge processing. This approach has been criticised for its limited validity, as these simple stimuli may not accurately reflect how the visual system processes edges in the more complex, real-world environments that humans typically encounter (Touryan and Dan, 2001; Olshausen and Field, 2005). As previously outlined, these studies have provided valuable insights; however, it remains unclear whether their findings for the processes involved can be generalised to natural vision.

Are spatial frequencies around 3 cpd equally important for edge detection when analysing entire scenes rather than isolated edges? In order to address this question and determine whether these findings are applicable to natural visual behaviours, it is essential to test them using other, more ecologically relevant stimuli. In this context, using natural images is useful because they display a broader SF spectrum than simple stimuli and are more representative of real-world visual experiences (Field, 1987).

When considering edge detection in natural images, we shift from discussing isolated edges to the broader concept of contours. In this context, we define contours as visible edges that often occur at the boundaries of objects or their elements in an image. Their representation is limited to the outlines of these elements, excluding the internal details such as fine-grained textures. Figures 1.1a and 1.1b illustrate the generalisation we seek to achieve.

The central objective of this thesis is to examine the transferability of research findings on edge sensitivity with simple, controlled stimuli, such as the one depicted in Figure 1.1a, to the contours of natural images, as shown in Figure 1.1b.



1.1(a): Simple edge

1.1(b): Natural image

To achieve this, we build on a recent experiment investigating human edge sensitivity in the presence of different 2D noise patterns (Schmittwilken et al., 2024). We selected this study to specifically challenge the SF-selective mechanisms underlying human edge sensitivity in natural scenes. The stimuli used in their experiment consisted of Cornsweet edges with varying SF properties, masked by different noise patterns. Cornsweet edges exhibit an abrupt change in luminance that gradually smooths out to the mean luminance on both sides of the transition. Participants were asked to indicate whether they perceived an edge above or below a marked midline via a button press. Figure 1.2 provides an example of a stimulus used in the experiment.

In accordance with previous research, their findings confirmed peak edge sensitivity around 3 cpd. They observed that noise of 3 cpd reduced the visibility of edges across all SFs and that pink noise had the strongest overall impact on edge sensitivity. Conversely, their results showed that low SF noises (0.5 cpd and brown noise) did not affect performance for any of the edges.

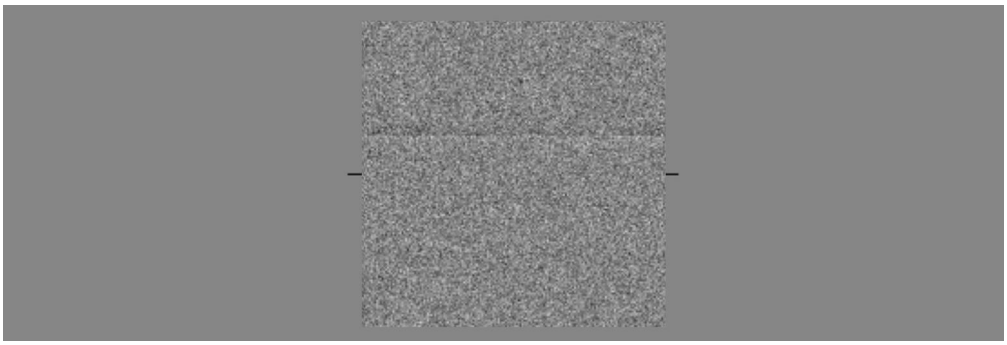


Figure 1.2.: Stimulus of the experiment by Schmittwilken et al. (2024), a 3 cpd edge masked with white noise. The correct response would be that the masked edge is above the marked midline.

Based on these findings, they concluded that SFs between 1 and 10 cpd are critical for edge detection, while noise with a SF below 1 does not affect edge sensitivity.
In this thesis, we aim to investigate whether the presumed mechanisms underlying human edge perception translate to contour perception in natural images. To address this question, we conducted a psychophysical experiment that measures contour perception in natural images, examining whether these effects align with the findings for isolated edges reported by Schmittwilken et al. (2024). In our experiment, participants were presented with stimuli comprising natural images with varying image contrast levels, that were masked with the same 2D noise patterns as those employed in the reference study. They were instructed to trace all the contours they perceived using a drawing tablet. To evaluate performance, we calculated the level of overlap between contour traces in the presence and absence of noise for each stimulus.
In order to determine whether the noise patterns that impact human edge perception of isolated edges also affect contour perception in natural images, we conducted a comparative analysis on the effects of noise and image contrast. The numerical outcomes of this analysis were then compared with those of the reference study. Their results suggested that pink noise and narrowband noise of 3 cpd have the strongest impact on human edge perception, while low SF noises such as narrowband noise of 0.5 cpd and brown noise have a minimal influence. If we observe similarly strong effects for pink noise and 3 cpd noise, along with a low impact of low SF noise on contour perception, we can conclude that the presumed mechanisms translate to natural images.
Additionally, we examined inter-observer variability in contour tracing. It is essential to ascertain whether participants consistently identified and prioritised the same contours for assessing both the reliability of our findings and potential disparities in task interpretation. The analysis of the similarity of participants' contour traces in the absence of noise allowed the determination of whether they perceived and prioritised the same contours. Depending on the observed variability, this analysis provides insight into both the robustness of our results and the validity of contour-based approaches.

Beyond individual differences, we also considered other potential sources of performance variability, such as image-specific factors. To assess this, we examined whether certain images were found to have a consistently positive or negative impact on performance levels. If we identified systematic differences, we could infer that specific image characteristics affect segmentation performance. This would suggest that perceptual contrast and other image properties play a role in contour perception.

Giving an outlook on the results, our findings align with the trends observed in the edge sensitivity experiment, indicating that the same noise patterns produced comparable effects on contour perception in natural images as on edge detection in simple, controlled stimuli. This suggests that the presumed mechanisms underlying human edge perception translate to natural images.

Furthermore, our analysis demonstrated that individual natural images influenced participant performance, suggesting that perceptual contrast and other image characteristics have a significance in contour perception. Finally, the consistency in participants' responses, where they generally prioritised the same contours, validates our experimental task as a reliable method for studying contour perception. This consistency underscores the task's effectiveness in capturing contour perception and highlights its potential for investigating visual processes using more behaviourally relevant approaches.

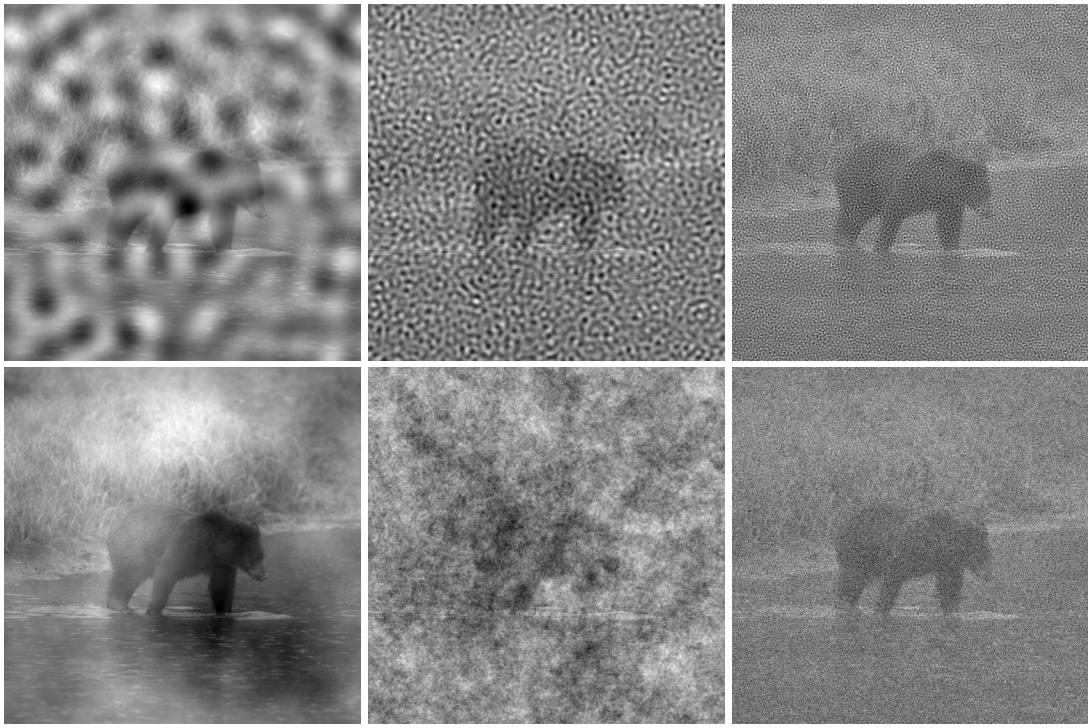
2. Methods

We aim to investigate whether the impact of different noise patterns on edge sensitivity observed in simple stimuli extends to natural stimuli. This section outlines the steps required to design an experiment that allows us to measure contour perception in natural images, ensuring comparability with the study by Schmittwilken et al. (2024).
We describe the natural stimuli used, the experimental task, and the methods to assess and analyse contour perception, along with the specific procedure followed in the experiment. In addition, we present the results of our pilot study, which we conducted for initial insights, and have based key experimental decisions on. Finally, we describe the experimental setup and apparatus used.

2.1. Stimuli

The stimuli consisted of natural images that were each masked with different noise and image contrast conditions, as well as a no-noise condition with one fixed image contrast. In our experiment setup, each stimulus spanned an area of 11.64×11.64 degree visual angle. They were centred on top of a grey background. The background and the mean luminance of the stimuli were $100 cd/m^2$.
A preliminary study evaluated various natural image data sets for the task of contour tracing (Sørensen, 2023), and it was determined that the data set of Grigorescu et al. (2003) was the most suitable. Consequently, it was decided to utilise this data set in our experiment. The data set contains 40 greyscale images, all being sized 512×512 pixels[1]. They depict different natural scenes, predominantly featuring animals within their natural habitat, as well as a limited number of artificial objects, such as cars.
Since our primary data question concerned the impact of noise on contour perception, we selected six different noise types as a key experimental variable. To ensure comparability between natural images and the effects of noise on edge sensitivity in simple, controlled stimuli, we adopted the same noise conditions used in Schmittwilken et al. (2024). This allowed for direct data comparison and enabled us to examine whether the observed trends align across both stimulus types.
The noises chosen were three narrowband noises (NB) with spatial frequencies of 0.5, 3 and 9 cpd and three broadband noises, brown, pink and white noise. Figure 2.1a–2.1f presents an exemplary image of the data set, with these six noise conditions applied.
While narrowband noise contains only a limited range of frequencies, broadband noise covers a wide range with different power densities in different frequency areas. The analysis of specific SFs in terms of their importance in contour detection is facilitated by narrowband noise.

[1] The images are available at http://www.cs.rug.nl/~imaging



2.1(a): NB of 0.5 cpd 2.1(b): NB of 3 cpd 2.1(c): NB of 9 cpd

2.1(d): Brown noise 2.1(e): Pink noise 2.1(f): White noise

Figure 2.1.: Overview of the six noise conditions being applied to the same image.

Meanwhile, the detection of non-linear effects when multiple frequency channels are stimulated simultaneously is enabled by broadband noise (Schmittwilken et al., 2024). The selected range of the three NB noises was adequate for exploring the differing impact of low, mid and high SFs on contour perception.

For the broadband noises, pink noise proved to be of particular interest, given the similarity of its power distribution to that of natural images (Field, 1987). Its power density decreases with increasing frequency, proportional to $\frac{1}{f}$. Brown noise's power density also decreases with increasing frequency, proportional to $\frac{1}{f^2}$. While its power spectrum is similar to pink noise, it has more power at frequencies under 0.1 cpd compared to pink noise. White noise, in contrast to pink and brown noise, has an equal power density across the entire frequency spectrum. Consequently, it is useful to provide additional insight into perception when all spacial frequencies are affected equally.

Figure 2.1 shows the visual impact of each noise. As the noise itself was generated randomly, the noise pattern obscured certain image features, which therefore could not be drawn by participants. To prevent the same contours being affected for each participant, we created five different noise masks for each noise. We randomly chose which one was used for each stimulus, although each participant saw each noise mask of each noise type once. As the noise mask mappings were stored, this also provided the opportunity to exactly replicate a stimulus if needed for analysis. Apart from the varying noise applied to the images, the contrast was the second variable we manipulated.

In this context, the term 'contrast' refers to the Root Mean Square (RMS) contrast, which is mathematically expressed as the ratio of standard deviation to mean luminance. This describes the variability of the luminance of each pixel relative to its mean. A high RMS contrast therefore is indicative of highly dispersed data. This, in turn, results in an image with strong contrasts, which also results in a higher visibility of contours.
For the stimuli, contrast can be applied both to the image and to the noise. To investigate contour detection performance, we fixed the noise contrast to 0.10 and varied the image contrast. We chose the same noise contrast to have equal amounts of noise in each noise condition as expressed by the noise power (proportional to the RMS contrast).
An example of how image contrast influences the visibility of contours can be seen in Figure 2.2. We chose five concrete image contrast values based on prior piloting per noise (see Section 2.5).

Figure 2.2.: Increasing RMS image contrast on an image with white noise of RMS = 0.10.

In summary, the stimuli were comprised of five different image contrasts and six different noises, resulting in 30 stimuli. We decided to apply these stimuli to different natural images instead of always using the same one. The rationale behind this approach was to eliminate the possibility of any learning effects that might be caused by drawing the same picture multiple times.
The attribution of effects to the actual perception of contours would be impossible in such cases, since it would be difficult to ascertain whether participants drew only what they saw, or whether they also drew what they remembered. Each participant saw each of these stimuli once in the first part of the experiment and the same images with a fixed contrast of 0.16 and no noise in the second part of the experiment.

2.2. Task

In vision research, experimental designs typically employ straightforward approaches to measure observer performance. For instance, the experiment in our comparison paper used a Two-Alternative Forced Choice (2-AFC) task, a widely applied method to assess sensitivity to a stimulus. In 2-AFC tasks, participants are presented with two answer choices and must select one, thus facilitating rapid data collection.

Given that each response is completed within a matter of seconds, it is possible to conduct multiple trials for each stimulus condition without the necessity of concern regarding the potential effects of learning. Additionally, because responses are either correct or incorrect, performance can be quantified simply by calculating the percentage of correct answers, enabling easy comparison across conditions.

In our experiment, however, there is no single correct answer. Instead, we want to assess how well participants perceive contours in the presence of different noise patterns, but we do not have a predefined correct answer that would tell us whether they perceived everything accurately. To address this, we asked participants to trace all the contours they could see. They viewed natural images overlaid with noise on a monitor and traced contours using a drawing tablet. In a preliminary study (Sørensen, 2023), participants were given the option of using either a mouse or a tablet, but for consistency and to eliminate potential extraneous variables, we restricted our experiment to the drawing tablet. To standardise the task, all participants received the same definition of what qualifies as a contour:

"We define contours as visible edges that often occur at the boundaries of objects or their elements in an image. Other examples include discontinuities at or between surfaces, or abstract features such as shadows or the horizon line. Importantly, contours represent only the outlines of these elements – they do not include internal details like fine-grained textures. Furthermore, contours do not need to be closed or continuous."

Figure 2.3 illustrates this with an example image. While details such as individual blades of grass or branches in the bush could technically be perceived as contours, we excluded them, classifying them as texture instead. This decision was made to ensure that the task remained feasible within a reasonable time frame. A too high level of detail would have made the tracing process too time-consuming and difficult for participants.

Figure 2.3.: Image with contour traces highlighted in orange, that participants were given as part of the task description.

Participants were instructed to trace only the contours they could actually see in the stimulus and to avoid any logical completions or continuations. We emphasised this distinction because our aim was to assess how noise affected edge detection, rather than how well participants could infer or reconstruct partially obscured images.

By providing concrete task specifications, we aimed to ensure that participants would generally trace the same contours, allowing us to analyse how noise specifically affects the perception of these contours.
However, asking participants to draw perceived contours manually introduces potential confounding factors, such as differences in drawing ability. Some participants might be able to perceive all the contours but struggle to accurately translate their perception into a drawing. To quantify contour perception while accounting for these variations, we chose an approach that compares contour traces drawn in the presence of noise to those drawn in the absence of noise. We extracted the contours participants traced for each stimulus, generating segmentations using a tool developed in a preceding study by Sørensen (2023). This allowed us to analyse how the presence of noise influenced the contours that participants were able to perceive and replicate.

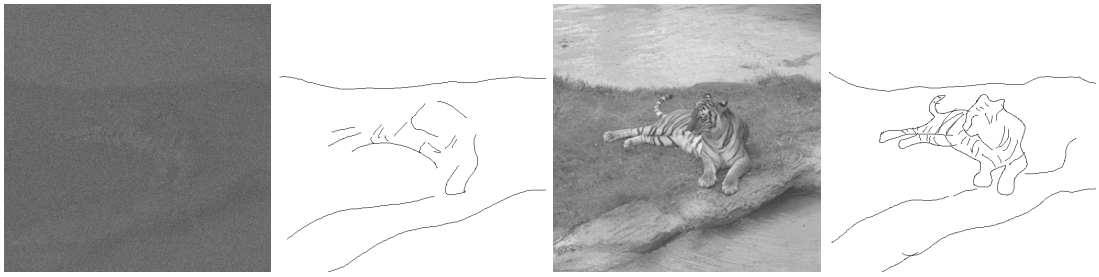


Figure 2.4.: Left: Image masked with white noise and a low image contrast, and its segmentation; Right: Image without noise and its segmentation.

In Figure 2.4, we first see a natural image masked with noise and next to it, we see a participant's segmentation, representing the contours they traced for this specific stimulus condition. To assess how much this stimulus condition affected the participant's performance, we then asked them to trace the same image, but this time without any noise and with high image contrast, as in the image on the right side in the Figure. These noise-free reference segmentations we call ground truths. For this particular observer, the resulting ground truth segmentation is shown on the outer right-hand side.

2.3. Measurements

To quantify the similarity between two segmentations, we need a method to calculate how much they overlap. Since the drawn contour lines are quite thin, it is nearly impossible for participants to trace them with pixel-perfect accuracy. To account for this, we introduce an error margin by using a dilated version of the ground truth. This allows for a more flexible comparison, where nearby pixels can still be considered as overlapping. While this does not qualitatively affect the results (Sørensen, 2023), it helps to scale them more meaningfully. Based on the findings of Sørensen (2023), we applied an error margin of ten pixels for all analyses in this paper, meaning that any contours that overlapped within this range were counted as a match.

Figure 2.5 illustrates this process. We eroded both the observer's segmentation where noise was present (in red) and the ground truth segmentation (in black) by ten pixels. The third image visualises their overlap: green areas indicate true positives (contours correctly traced in both segmentations), black areas show false negatives (contours present in the ground truth but missing from the participant's segmentation), and red areas represent false positives (contours traced by the participant but absent in the ground truth). This method provides a structured way to measure how well participants could perceive and replicate contours under different noise conditions.

Figure 2.5.: From left to right: eroded no-noise segmentation, eroded noise segmentation, overlay of the two (green marks overlapping segments).

From a mathematical perspective, we can quantify the visualised results in the image above by calculating a performance score ranging from zero to one, based on the number of pixels classified into each of the three categories.

$$P = \frac{|E|}{|E|+|E_{FP}|+|E_{FN}|}$$

Specifically, $|E|$ represents the set of correctly detected contour pixels, while $|E_{FP}|$ denotes the number of false positives (pixels present in the noise segmentation but not in the ground truth), and $|E_{FN}|$ represents the number of false negatives (pixels present in the ground truth but absent from the noise segmentation). This performance measure, designed to assess the similarity between two segmentation maps of an image, was originally introduced by Grigorescu et al. (2003).

As previously mentioned, we decided to compare each participant's segmentations against their own reference segmentations for each image. We chose this participant-specific score calculations to account for individual differences in contour perception. Not all participants may identify the same contours as equally relevant, particularly when it comes to the level of detail in e.g. foliage. In addition, this approach allows us to analyse variability in contour selection in the absence of noise across participants. This provides further insight into how effectively our experimental task captures contour detection.

2.4. Analysis

Now that we have established a method for quantifying participants' performance in contour tracing in the presence of noise, we aim to evaluate the differing effects of the six noise patterns, presented in Section 2.1 on contour perception. This analysis allows us to compare our findings with those of the comparison paper and determine whether specific noise patterns have a similar influence on contour perception in natural images as they do on edge detection in simple stimuli. To achieve this, we fit and analyse psychometric functions that model the relationship between a physical stimulus (x-axis) and the corresponding measured responses of human observers (y-axis).

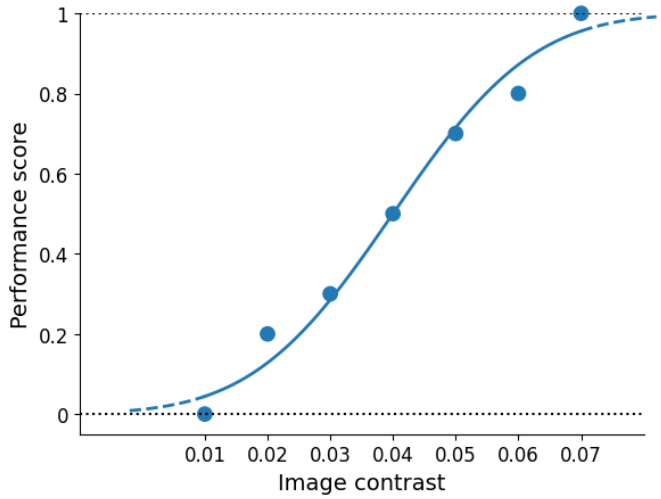


Figure 2.6.: Exemplary psychometric function with randomised data.

In our study, the psychometric functions describe the relationship between natural image contrast (x-axis) and participants' performance (y-axis) under different noise conditions, see the exemplary Figure 2.6. These functions estimate the level of image contrast required for participants to achieve a certain performance score in the presence of a noise pattern, based on the data collected in our experiment. This approach enables us to assess how contrast thresholds vary across different noise types and quantify the relative impact of each noise condition on contour perception.

In order to fit to our functions, we had to decide on a number of parameters: For the psychometric function, we selected the cumulative Gaussian distribution as the sigmoid model. This choice is in line with the approach used in the comparative study and is appropriate for our response data, which is continuous between zero and one.

We used 'equal asymptotes' to fit our functions, assuming symmetrical upper and lower bounds. Given the nature of our task, where participants either perceive a contour and draw it or do not see it and therefore do not draw anything, this approach is appropriate. Stimulus-independent errors are assumed to be equally likely in both cases. Since guessing is not applicable in this task and we assume that participants did not systematically guess, the guess rate is set to zero.

A participant randomly guessing and drawing contours would likely result in a performance score of zero. Similarly, the lapse rate, representing stimulus-independent incorrect responses, was set to zero. As the experiment included an undo option, we assume that participants corrected unintended errors.
In the comparison paper, the psychometric functions relate the edge contrast (x-axis) to the percentage of correct responses (y-axis). Since they performed a 2-AFC experiment, the lower asymptote was fixed to $1/n$, with n = 2, meaning that an observer randomly guessing would achieve 50% accuracy. Their upper asymptote was free to vary. The psychometric functions for both experiments were fitted using Psignifit, a Python library for Bayesian psychometric function estimation (Schütt et al., 2016).
Before we could start our experiment, however, we needed to establish a set of image contrasts. In order to actually fit psychometric functions for our experiment, we needed contrasts that effectively captured participants' performance across the different noises. To ensure the selected contrast levels were representative, we conducted a pilot study to guide this decision.

2.5. Piloting

We conducted two pilot studies to finalise outstanding decisions for the experiment. The first aimed to determine an appropriate set of image contrasts for each noise condition, while the second focused on selecting a subset of natural images to be shown to participants in the main experiment. For the first pilot study, we decided to select five image contrasts per noise condition, resulting in a total of 30 stimulus conditions (6 noise types × 5 image contrasts). This number was chosen as it provided a manageable task workload for participants, ensuring they could complete all required segmentations without fatigue, and approximately within one hour.
To effectively compare performance across noise conditions, we needed to select contrast levels that captured a broad range of performance scores while avoiding ceiling and floor effects. If contrast levels were too high, participants might consistently perceive all contours across conditions, making it difficult to assess the impact of noise. Conversely, if the contrast levels were too low, performance might remain consistently poor, preventing meaningful comparisons. Using the same contrast levels for all noise conditions could have resulted in some noise types always allowing full contour visibility, while others completely obscured them. To ensure comparison, we needed to determine the contrast levels for each noise condition individually.
To address this, we initially selected a larger set of 16 RMS contrast values ranging from 0.01 to 0.16. The goal was to collect data on how contour visibility changed across different contrasts using the previously introduced performance measure, which quantifies the similarity between segmentations drawn in the presence of noise and the corresponding ground truths drawn without noise.

To minimise potential biases introduced by the natural images, we randomised the assignment of natural images to stimulus conditions (i.e., contrast level × noise type) and ensured that no image was used more than once per noise type. The pilot was carried out by one observer, drawing contour segmentations for each stimulus condition, meaning one segmentation for each of the 16 image contrasts for each of the six noises, resulting in 96 different segmentations in total. They also drew their own ground truths once for each image, with no noise present and a high image contrast of 0.16.

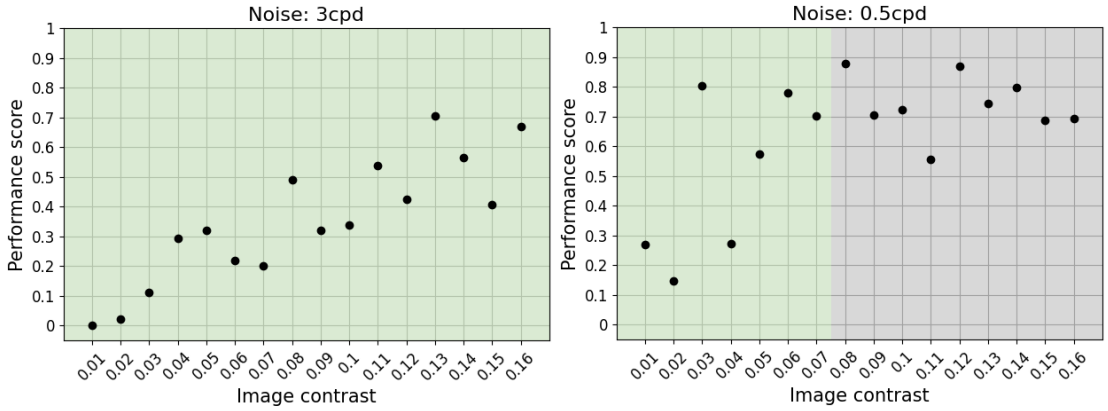


Figure 2.7.: Exemplary results of the first pilot for two noise conditions: NB noise of 3 cpd and 0.5 cpd. Results for the remaining four noise patterns are provided in the Appendix (Figure A.1). Each data point represents a single segmentation and the corresponding performance score achieved at a given image contrast for a specific noise condition. The differently coloured areas show the range from which we sampled our final contrast set.

The data in Figure 2.7 reveal different performance trends for the two noise conditions. For NB noise of 3 cpd, performance increased steadily with rising image contrast, peaking at the highest image contrast of 0.16. In contrast, for NB noise of 0.5 cpd, performance improved more rapidly, reaching its peak earlier and stagnating at around half of the total contrast range. The green-shaded areas in the plots indicate the contrast range where performance was still increasing, while the grey areas mark the point at which performance appeared to stabilise.

Since our goal was to capture an even distribution of performance levels with our final contrast selections, we based our decisions on these observed trends. Specifically, we aimed to include contrasts where participants would be unable to perceive any contours, where performance would be optimal, and three intermediate levels between these extremes. First, we selected the lowest contrast value based on the minimum performance scores observed in the pilot.

Generally, this was 0.01 for all noise conditions, except for NB noise of 0.5 cpd, which had its lowest performance at 0.02. However, we considered this an outlier rather than representative of the overall trend. Additionally, for NB noise of 0.5 and 9 cpd, the lowest contrast did not result in a performance score of zero, which we aimed to achieve.

To address this, we chose an even lower minimum contrast of 0.001 for the actual experiment. Next, we determined the highest contrast level.
As indicated in the plots, we aimed to identify the contrast at which performance no longer increased significantly. For NB noise of 3 cpd, this point was not reached within our tested contrast range, so we selected the highest available contrast of 0.16. The same applied to pink noise, as shown in Figure A.1. In contrast, for NB noise of 0.5 cpd, performance stagnated much earlier, at approximately 0.07. A similar pattern was observed for brown noise and NB noise of 9 cpd, both of which reached peak performance at lower contrast levels. Consequently, we selected 0.07 as the highest contrast for these three noise types. For white noise, which showed intermediate behaviour, we chose a maximum contrast of 0.12.
Finally, we selected three additional contrast levels for each noise condition by computing evenly spaced values between the lowest and highest contrasts. For the drawing of the ground truths (absence of noise), we decided to use the highest contrast applied in the noise-masked conditions, 0.16. The final image contrast choices are represented in Table 2.1.

Noise	**Image Contrasts**
0.5 cpd, 9 cpd, brown	0.001, 0.018, 0.036, 0.053, 0.07
3 cpd, pink	0.001, 0.04, 0.08, 0.12, 0.16
white	0.001, 0.03, 0.06, 0.09, 0.12
none	0.16

Table 2.1.: Decision on final set of image contrasts per noise.

The plots indicate that even with a sample of data from a single observer, the distribution of performance scores across noise conditions aligns with our expectations based on the findings from our main comparison paper (Schmittwilken et al., 2024). Specifically, NB noise of 3 cpd and pink noise showed the slowest increase in performance across contrast levels, whereas the other noise conditions showed a relatively rapid improvement within the first few contrast levels. This serves as an initial indication that our study design is well-suited for investigating contour perception in natural images.
Another noticeable trend in the pilot results is the variability in performance scores. For example, in the plot for NB noise of 3 cpd (Figure 2.7), some scores are lower than the preceding ones despite an increase in image contrast. The influence of the specific image being traced is a likely explanation. As previously mentioned, we randomised image assignments across stimulus conditions to prevent learning effects from repeatedly drawing the same image. However, this also introduced variability due to differences in the individual images. Even when two images share the same noise pattern and RMS contrast, their perceived contrast, which we cannot measure directly, may differ.

This can result in some images appearing more or less contrasted to the human eye and therefore cause variations in performance scores. To investigate this further and minimise the influence of image identity on performance scores, we conducted a second pilot study.

The data set by Grigorescu et al. (2003), which we chose for our experiment, contains 40 different natural images (see Section 2.1). Since our experiment used six noise patterns and five image contrasts, we required only 30 of these images. The goal of the second pilot study was to identify and exclude the ten images most likely to introduce variability in performance scores due to perceptual contrast differences. By doing so, we aimed to ensure that participants' performance scores in the main experiment were primarily influenced by the stimulus conditions rather than by image characteristics.

To assess the impact of image identity, the second pilot involved a single observer drawing segmentations for all 40 images, each masked with the same stimulus condition. Based on the results of the first pilot, we selected NB noise of 3 cpd with an image contrast of 0.08. We chose a noise where we had a relatively even distribution of data points across the image contrasts in the first pilot, and then the contrast in the middle of the contrast range where the increase in performance should be greatest.

Additionally, this pilot study was intended to help us decide whether each stimulus condition should be fixed to specific images for all participants or randomly assigned to images for each individual. Fixing stimulus conditions to specific images would allow us to analyse participant-specific drawing patterns, revealing how much variation exists in performance scores for exactly the same stimulus across different participants. However, this approach would be problematic if image characteristics significantly influenced performance scores. If a particular image had a lower perceptual contrast and was consistently more difficult to trace, it would systematically receive lower performance scores across participants, making it unclear whether observed effects were due to stimulus conditions or the image itself. While the first pilot already suggested that image identity could affect performance scores, the second pilot aimed to provide a direct comparison to quantify this variability and see whether our initial suspicions would be confirmed.

Figure 2.8 shows the results of the second pilot study. They confirm that the individual image strongly influences performance scores, supporting our observations from the first pilot study. Most scores have a pairwise difference of less than 0.2, however, the difference between the highest and lowest score exceeds 0.5. Examining the images at these extremes further illustrates the role of perceptual contrast. Figure 2.9 presents both the original and noise-masked versions of two images. The upper image, *hyena*, received a performance score of 0.75, while the lower image, *gnu_2*, scored 0.21. The *hyena* image features dark animals against a light background, making contours more distinct despite the noise.

In contrast, *gnu_2* has more varied greyscale levels and softer edges, reducing contour visibility. By excluding the ten images most affected by perceptual contrast, we aimed to minimise the influence of image characteristics on performance scores in the main experiment. While the remaining images seemed to be more consistent in perceptual contrast, some variability was still present. To further reduce potential bias, we opted for the random assignment of stimulus conditions to images, rather than fixing them for all participants.

We decided to use two of the ten images for the training session preceding the experiment. This will be detailed in the next section.

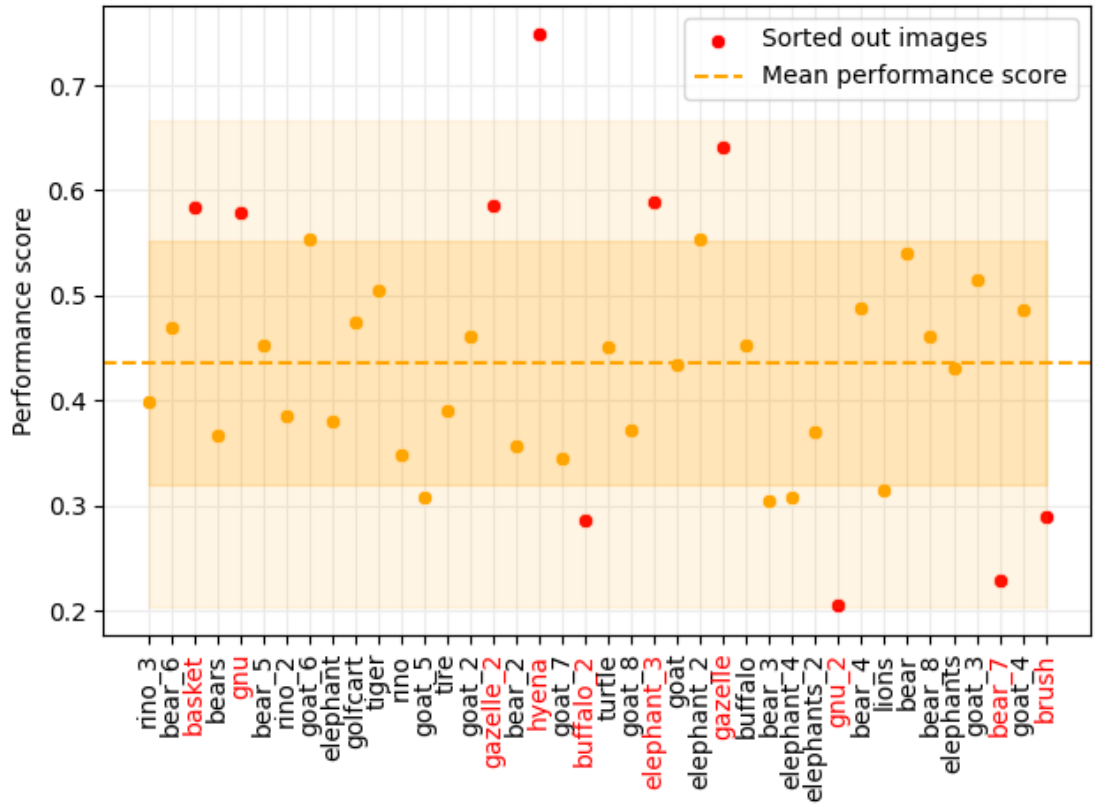


Figure 2.8.: Performance scores for all natural images for the same stimulus condition of NB noise of 3 cpd and RMS = 0.08. The y-axis represents the performance score, while the x-axis lists the 40 natural images. Each data point corresponds to the performance score for a specific image.
The dashed line in the plot marks the mean performance score across all images at 0.435. The darker orange background around this line represents one standard deviation from the mean, while the lighter orange area extends to two standard deviations.
Red data points indicate the ten images we excluded, chosen based on the largest deviations from the mean. Specifically, we removed six images with above-average, and four with below-average scores.

Figure 2.9.: Original images and masked with NB noise of of 3 cpd and the same RMS (stimulus condition of the pilot).

2.6. Procedure

To conduct our actual experiment, we needed a well-controlled setup that would ensure the reliability and reproducibility of our findings. We designed our procedure to be consistent across participants and to minimise confounding factors, so that any effects we observed would not be not due to methodological inconsistencies.

Ten observers took part in our experiment in December 2024. The experiment began with a short demographic questionnaire that collected information on age, vision correction, and handedness. The latter was included to assess whether handedness might influence the use of the drawing tablet.

Participants then familiarised themselves with the setup and operation of the graphics tablet in a training session. The training was done to reduce the likelihood of mistakes that could result in unusable segmentation maps. This step was crucial to ensure that any missing contours in the actual experiment were due to noise interference rather than user errors. During training, participants viewed a stimulus on the monitor and were asked to trace all the contours they perceived using a drawing tablet. The pen functioned like a mouse, with a cursor displayed on the monitor corresponding to the pen's hovering position. This feature helped participants accurately place their pen on the stimulus, ensuring they could see where they were drawing.

When pressed against the tablet, the pen registered input as a black line drawn over the stimulus on the screen. This input was layered directly onto the stimulus to further help participants to ensure that they had traced the intended contours correctly. We customised the tablet buttons available to suit the experiment and provide the following options:

1. Undo a line
2. Switch the display method
3. Continue to the next stimulus

(1) was intended to correct unintended mistakes and reduce potential errors. When changing the display method with button (2), the participant could chose to either see the stimulus with their input added on top, or toggle the view to a side by side comparison, showing the stimulus on the left and their segmentation in black on a white background to the right. The intention was to help participants spot missing gaps in the segmentation that might be difficult to see on top of the stimulus, which might especially occur while drawing onto darker stimuli.

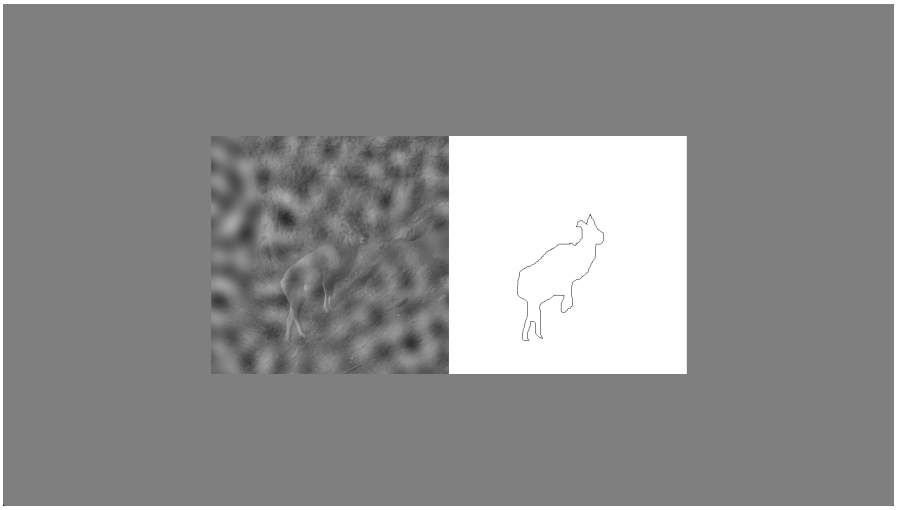


Figure 2.10.: Toggled split view of the stimulus (left) and the segmentation (right).

If the participant continued to the next stimulus with button (3), the segmentation was saved as a binary image, showcasing the drawn contours as black lines on a white background, identical to the right panel of the split view.
The training session used two natural images with an RMS contrast of 0.16 and no added noise, allowing participants to practice in a controlled setting and get used to the tools. To prevent learning effects, these training images were not part of the experimental image set but were instead selected from the rejected image pool of the second pilot study. All participants received the same training stimuli to maintain consistency across subjects. They could repeat the training session if they wished.

Following training, participants proceeded to the main experiment, which consisted of 30 stimuli, with each stimulus presented once. Other than that, the setup remained the same as described above. One experiment session took about one hour. The experiment was conducted in two separate sessions. In the first session, participants traced contours on images masked with different noise conditions. In the second session, they traced the same images but without noise. The procedure was identical between sessions, with the exception that the demographic information was collected only during the first session.

2.7. Apparatus

The stimuli were displayed on a ViewPIXX 3D monitor ($523 \times 293mm, 1920 \times 1080px$, $120Hz$) using the presentation software HRL [2]. The ViewPIXX is an LED monitor. We used its 'standard backlight' mode, where all LEDs are constantly illuminated (up to $250cd/m^2$).

The experiment was conducted in a darkened room with a fixed headrest positioned $70cm$ from the monitor to ensure a consistent viewing distance across trials. The spatial resolution at this distance was 44 pixels per degree.

The input device was a Wacom Intuos pen tablet used together with the appropriate Wacom Pen. The tablet had no screen – it provided the drawing surface, while the input it registered was displayed on the monitor.

[2] http://github.com/computational-psychology/hrl

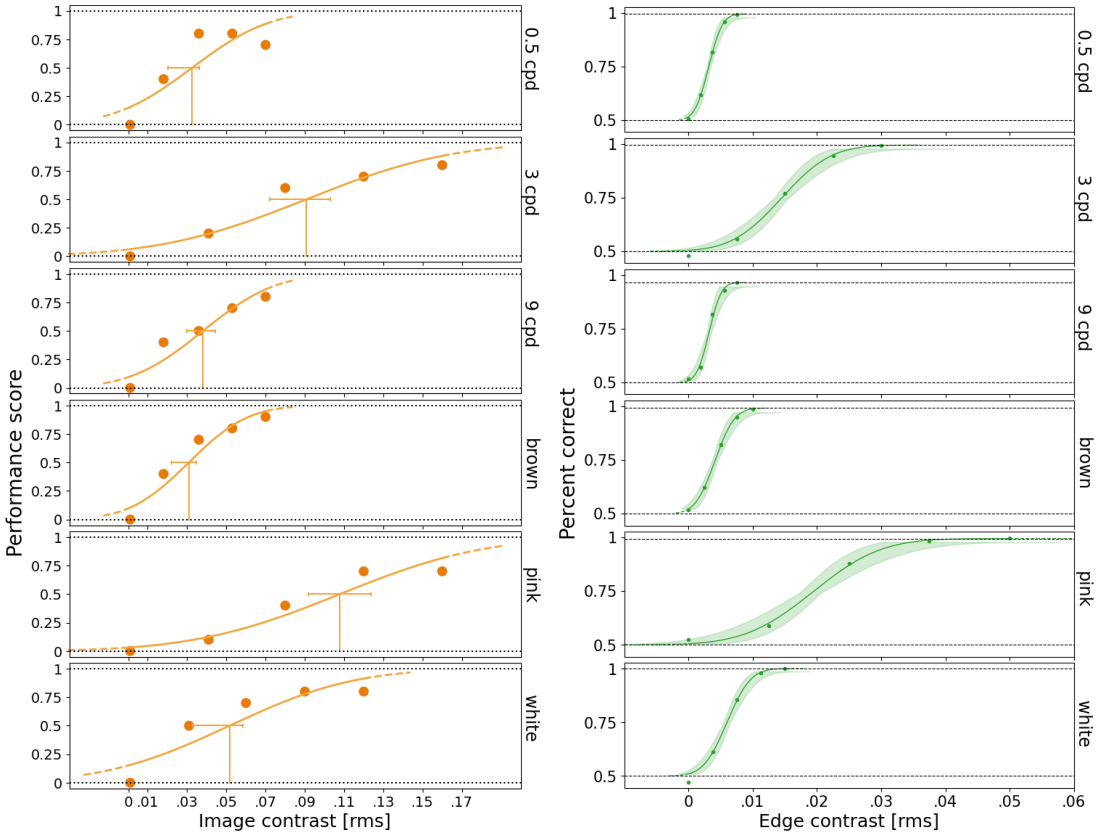
3. Results

This chapter presents the results of our experiment and is structured according to our different data questions. We tested 10 subjects (6 female, age range between 18 and 34), all of them being right-handed. All subjects had normal (N = 4) or corrected-to-normal (N = 6) vision.
We found no evidence that demographic factors systematically influenced contour perception in our study. Thus, we can exclude their influence on the following analyses.

3.1. Impact of Noise

Our main data question examines whether noise affects human edge perception for controlled stimuli similarly to contour perception for natural scenes. To investigate this, we analyse how different noise patterns and image contrasts influence contour perception in natural images and compare these findings to results from experiments on simple edges. For this comparison, we reference the study by Schmittwilken et al. (2024), which used a two-alternative forced-choice (2-AFC) task with simple stimuli.
In their experiment, observers detected a horizontal Cornsweet edge by indicating its position (above or below a midline) via a corresponding button press. They tested three Cornsweet edges with different peak spatial frequencies (SF) of 0.5, 3 and 9 cpd. Since natural images predominantly consist of low SFs (Field, 1987), as described in Section 2.1, we focus on the 0.5 cpd edge, as it provides the most relevant comparison.
Noise masking conditions were identical across experiments, consisting of three narrowband (NB) noises (0.5, 3, and 9 cpd) and three broadband noises (brown, pink, and white). To compare our results, we fitted psychometric functions to our experimental data and evaluated how the observed trends align with those reported in Schmittwilken et al. (2024).
To account for variability in individual performance, we scaled the performance scores per participant before fitting. Each participant's scores were adjusted relative to their highest performance to minimise the influence of individual differences, such as varying familiarity with drawing on a graphic tablet. This normalisation ensured a more consistent comparison across participants.
Figure 3.1a presents the psychometric functions for our contour perception experiment, and Figure 3.1b shows those for the edge detection experiment by Schmittwilken et al. (2024). First, we can observe that the set of image contrasts per noise condition, determined during the pilot phase, proved to be a suitable choice. Since the selection was based on data from a single participant, it was uncertain whether these values would generalise well to other participants.

However, the fact that we were able to plot psychometric functions for all noise conditions without reaching ceiling or floor effects confirms that the chosen contrasts were appropriate.



3.1(a): Contour perception experiment

3.1(b): Edge detection experiment

Figure 3.1.: Comparison of psychometric functions relating contrast to performance. The shaded area in the comparison plots represents the 68% credible interval, indicating the precision of the function fit. In our plots, this is presented by the width of the horizontal line at threshold values. The credible interval shows the range within which the psychometric function is expected to fall with 68% probability.

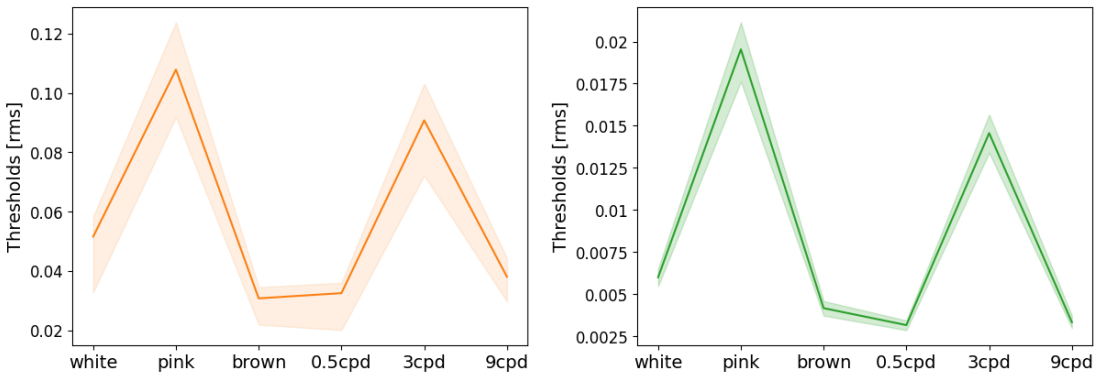
There is still room for refinement, particularly for NB noise of 0.5 cpd. The highest performance was already reached at the third contrast level, suggesting that the highest contrast value could have been set lower. Similar seems to be the case for white noise, though less pronounced.

If we now examine the psychometric functions obtained from our experiment, we observe that both pink noise and NB noise of 3 cpd had the strongest impact on contour perception performance. These conditions resulted in the shallowest psychometric curves, with pink noise leading to slightly lower average performance scores. In contrast, NB noise of 0.5 and 9 cpd, and brown noise produced the steepest curves, indicating minimal disruption to contour perception. White noise exhibited an intermediate effect, with a shallower curve than the least disruptive conditions but steeper than those

for pink noise and 3 cpd noise. When comparing the psychometric functions from both experiments, we observe similar trends. Conditions with steep (0.5 cpd, 9 cpd, brown noise) and shallow (3 cpd, pink noise) curves align across experiments, with white noise falling in between.

To compare the psychometric functions in more detail, we analyse their thresholds across all noise conditions. Thresholds represent the contrast levels at which the psychometric function estimates that participants, on average, achieve a specific level of performance. In this analysis, we focus on the contrast level at which participants reach 50% performance, as it marks the transition from chance-level to above-chance perception. For the contour perception experiment, we define the threshold as the contrast level at which the performance score reaches 0.5. This corresponds to 50% of the maximum performance, assuming both the guess rate and lapse rate are zero. In the edge perception experiment, the threshold is set at 75% correct responses. This is due to the 2-AFC design, where the lower asymptote of the psychometric function is fixed at 50% due to the chance level of guessing. The threshold is therefore defined as the midpoint between the upper and lower asymptotes.

Although the absolute threshold performance levels differ between the two experiments, they conceptually represent the same point: the contrast level at which participants achieve halfway between chance performance and their maximum performance. This equivalence allows for a meaningful comparison of thresholds across experiments, despite differences in task structure.



3.2(a): Contour perception experiment

3.2(b): Edge detection experiment

Figure 3.2.: Comparison of contrast thresholds to achieve 50% / 75% performance per noise condition, corresponding to the psychometric functions shown earlier. The shaded areas represent the 68% credible intervals for the estimated thresholds.

The threshold distribution in Figures 3.2a and 3.2b shows the extent to which different types of noise affect perception to varying degrees. For example, in both experiments, a pink noise masked stimulus requires a significantly higher contrast level to reach 50% threshold than a brown noise masked stimulus.

Examining the distribution of thresholds, we observe that in both experiments, pink noise results in the highest contrast threshold, followed by NB noise of 3 cpd and then white noise. The contour perception experiment shows the smallest contrast threshold for brown noise, followed by 0.5 cpd noise and then 9 cpd noise. In the edge detection experiment, the pattern of lower thresholds follows a slightly different order: 0.5 cpd noise results in the lowest threshold, 9 cpd noise follows, with brown noise slightly higher. However, in both cases, the absolute differences between these lower thresholds are minimal – in the contour perception experiment, for example, the RMS thresholds for brown noise and 0.5 cpd noise differ by around 2.57% relative to the maximum available contrast of 0.07 for both noises.
We must also consider that the threshold for NB noise of 0.5 cpd in the contour perception experiment would likely have been lower if our psychometric functions were fitted better. The distribution of performance scores across image contrasts for NB noise of 0.5 cpd suggests that the psychometric function would have been steeper with a more optimal functional fit.
Looking at the 68% credible intervals of the thresholds, we observe that they are more widespread for our experiment. This indicates a greater uncertainty in estimating the thresholds, likely due to the not ideal fits of the psychometric functions for our experiment. This could be, for one, improved by more carefully choosing the parameters for the psychometric function. For another, the sample size could be increased. Since each participant completed only one trial per stimulus condition, limited by the lengthy duration of the task, our data set consists of only ten samples per condition. In contrast, the 2-AFC experiment consisted of 200 trials, which allowed for a better prediction of the threshold values.

3.2. Impact of the Image Identity

In the second pilot experiment, we confirmed that the individual image impacted the contour perception performance, and we tried to minimise that impact by selecting a subset of the natural images with similar perceptual contrast. However, when examining the performance scores in our experiment data, we still noticed a considerable variation between data points for the same stimulus conditions, as shown in Figure 3.3. Looking at the image contrast of $RMS = 0.036$, performance ranges from a score of 0.1 to nearly 0.7. Since the image assignment was randomised, participants were mostly drawing different images under the same noise and contrast conditions.
We want to investigate how much this variation can be attributed to the natural image each participant viewed for the stimulus condition, or, more specifically, how much the selection of image influences the contour perception, beyond the effects of noise and contrast. Although all images were adjusted to have the same RMS contrast, their perceptual contrast, meaning how contrasted they appear to the human eye, still varies

and is a factor we cannot directly quantify. Some images may naturally appear more or less contrasted despite having the same RMS contrast, which can affect segmentation performance.

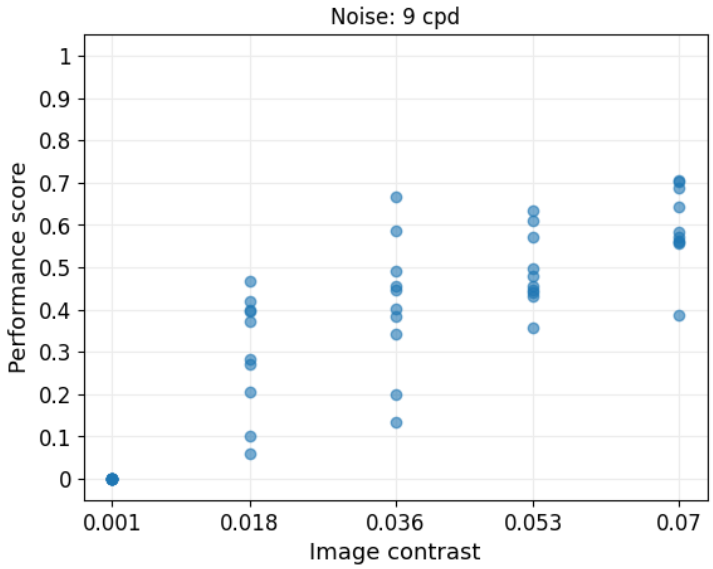


Figure 3.3.: Original performance scores[1] of all participants for stimuli masked with NB noise of 9 cpd.

To examine these variations, we compared participants' actual performance on each image to the predicted performance based on the fitted psychometric functions (see Section 3.1). Specifically, we computed the difference between each participant's performance score for a given image and the expected performance at the corresponding noise and contrast level. As each image was seen by all participants (N = 10), we averaged these differences to quantify how much the segmentation performance of each image deviated on average from the expected trend. The resulting values shown in Figure 3.4 indicate the influence of individual images on performance. A negative difference suggests that an image led to systematically lower performance, potentially due to a lower perceptual contrast making contours harder to perceive under noise masking. Conversely, a positive difference suggests that the image contains features that make contour tracing easier, regardless of noise or contrast.

The visualisation confirms that individual images influence performance scores and, consequently, contour perception. Examining the mean differences, we observe that some images deviate by more than 15% from the predicted performance, suggesting they were consistently harder or easier for participants to trace compared to others. This indicates that, despite RMS contrast normalisation, perceptual contrast and image-specific features affected visibility.

[1] These scores are unscaled, representing the originally calculated values. A scaled version of this graph, where the scores are adjusted relative to individual maximum performance, is provided in the Appendix (Figure A.2).

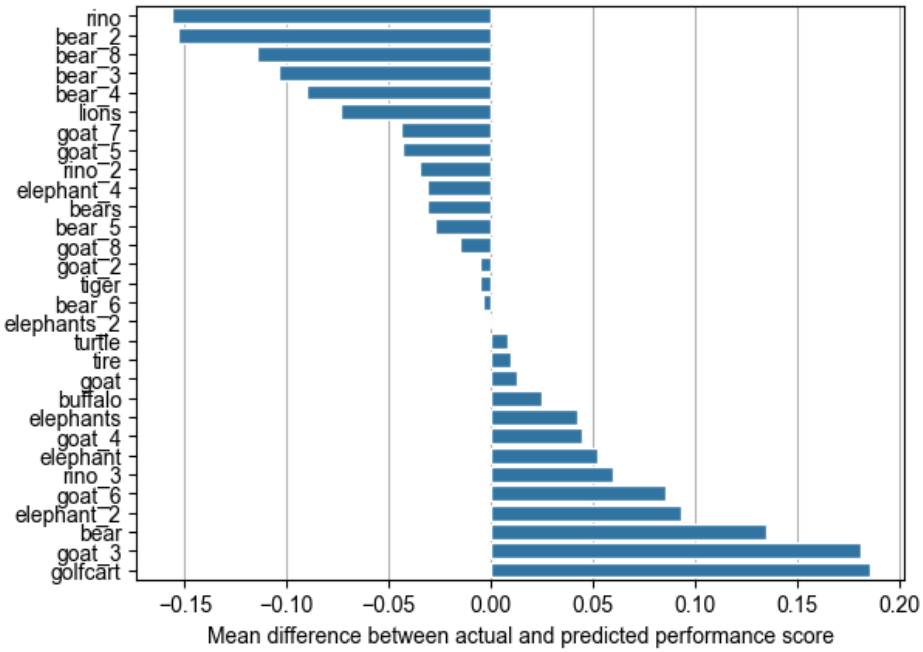


Figure 3.4.: Mean differences between actual performance scores and predicted values from the psychometric function at the same stimulus condition. Each bar represents a specific image, with negative values indicating worse-than-expected performance and positive values indicating better-than-expected performance.

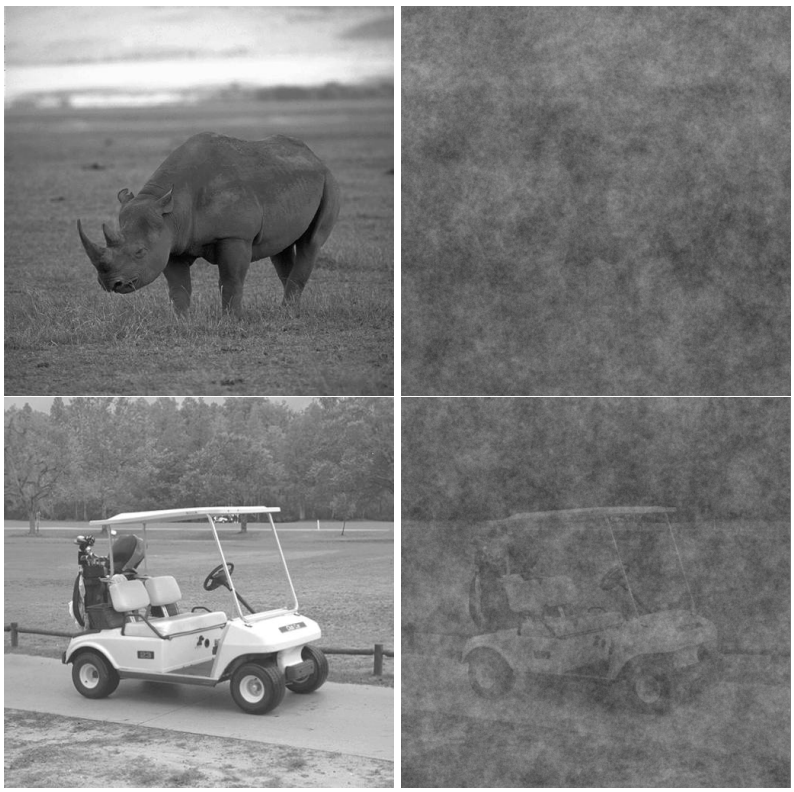


Figure 3.5.: Original *rino* and *golfcart*, and masked with pink noise and the same image contrast.

To explore this further, we analyse the images with the highest negative and positive differences. Participants' performance for the *rino* image was, on average, *-0.156* below the expected score for that stimulus condition, indicating it was more difficult to trace. In contrast, performance for the *golfcart* image was *0.186* above the expected score, suggesting it was easier to trace.
The differences in perceptual contrast are already noticeable when comparing the original versions of the images, shown in Figure 3.5. The *rino* appears to have a lower perceptual contrast, consisting mostly of similar greyscale tones. The only significant luminance difference is at the top of the image, but it lacks a well-defined contour separating it from the rest. In contrast, the *golfcart* has a bright main object with multiple well-defined edges that contrast sharply against darker components. The versions of these images masked with pink noise further highlight the impact of these features. While the *rino* becomes almost entirely obscured, the *golfcart* remains visible despite both images having the same RMS contrast and noise mask.
We conclude that perceptual contrast plays a key role for contour perception in natural scenes. Images with higher perceptual contrast were easier to segment, likely due to stronger luminance differences and well-defined contours that remained distinguishable even under noise masking. In contrast, images with lower perceptual contrast resulted in poorer performance scores. Their weaker luminance differences and less distinct contours made them more susceptible to noise, leading to greater contour loss and a more challenging segmentation task.

3.3. Inter-Observer Variability

To analyse how well participants perceived contours, we calculated performance scores that quantify the similarity between two segmentations of the same natural image – one drawn in the presence of noise and the other in its absence. In our experiment, we chose to compare segmentations of the same participant. Each participant completed two experiment sessions: in the first, they traced contours of natural images masked with noise, and in the second, they traced the same images without noise. While this approach led to a more complex and time-intensive data collection, it allowed us to compute performance scores that were independent of individual drawing styles or contour selection preferences.
This method ensured that participants' scores were not biased by individual tendencies, such as drawing more details than others. With no objectively correct set of contours, our main concern was whether participants remained consistent in which contours they traced across sessions. For each image, there was an individual choice as to which of the visible contours were to be traced. Exploring these selections provides insight into variability across individuals, allowing us to evaluate the reliability of contour-tracing tasks and identify potential improvements for future experiments.

To investigate this, we focused on the segmentations drawn in the absence of noise, referred to as 'ground truth' segmentations. For them, all participants traced each natural image once under identical viewing conditions, with a high RMS contrast of 0.16.

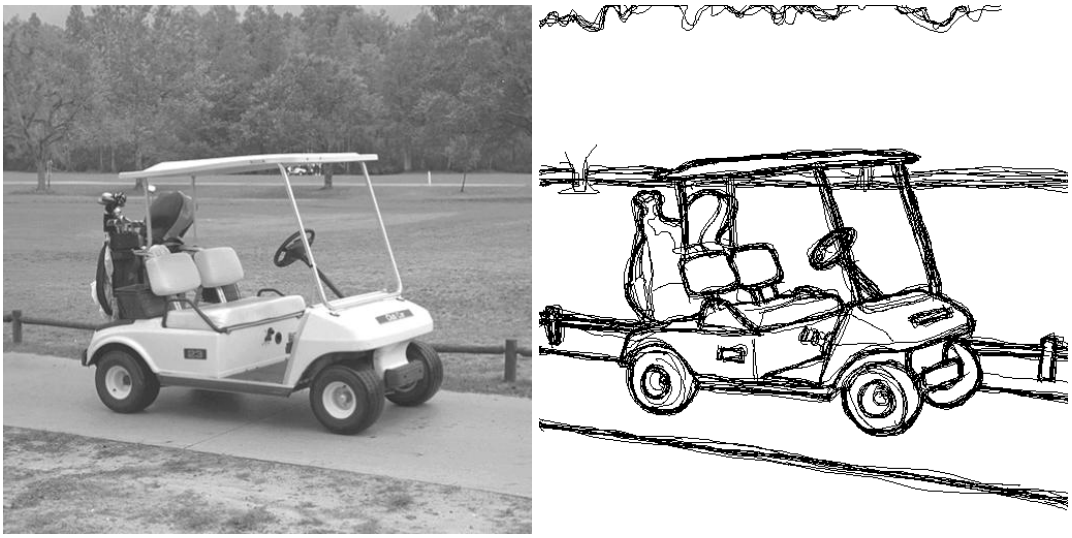


Figure 3.6.: Original image and the corresponding ground truth drawings from all ten participants, overlaid to visualise individual differences in contour selection.

The segmentation overlaps for this image, as shown in Figure 3.6, demonstrate a high degree of consistency across participants. While the individual lines do not perfectly align, they clearly indicate that all participants attempted to trace the same contours. This suggests a strong consensus regarding which edges were perceived as important in the original image. The varying width of the overlapping lines also supports the notion that humans are unable to produce pixel-perfect traces. This reinforces our reasoning for incorporating an error margin to account for this kind of variability.

However, there are also cases where participant agreement on which contours to draw was less consistent. This is illustrated in Figure 3.7. Here, the greater variability in contour selection indicates that participants can also differ in their interpretation of the most relevant edges in a natural scene.

The overlapping ground truth segmentation shows that most participants traced the same contours, indicated by the darker regions where multiple lines are on top of each other. The general shape of the goat was consistently outlined by all observers, whereas differences emerged in the background and in the details of the goat. Notably, only one participant traced the contours of the shrub, which we can see by the presence of a single, non-overlapping line.

Since we observe some variation in what participants consider relevant contours, we aim to quantify these differences. To do so, we require a baseline for each image – a reference against which we can compare the individual participants' ground truth segmentations.

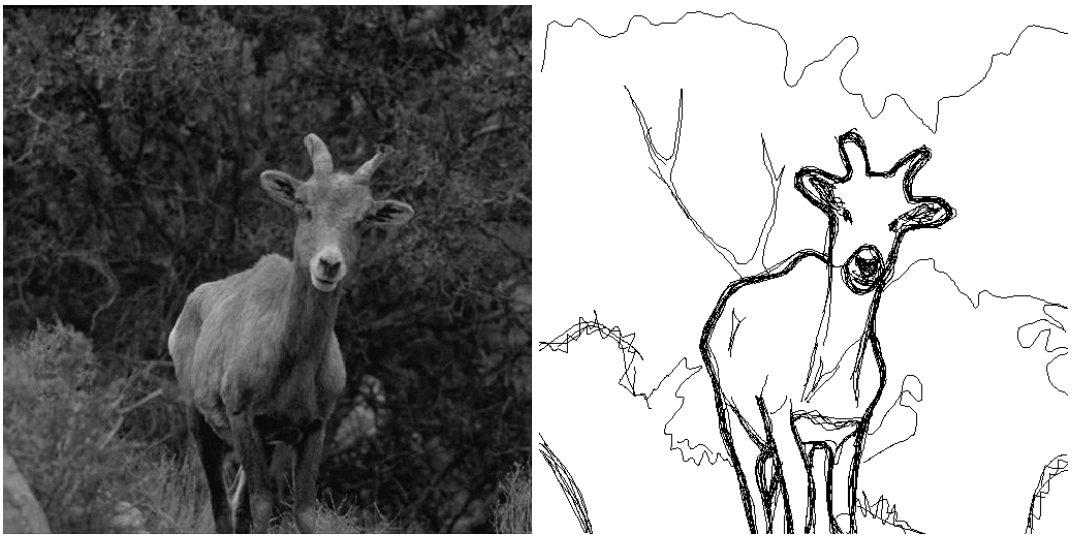


Figure 3.7.: Original image and the overlap of all participants' ground truths.

The images used in our experiment originate from the data set by Grigorescu et al. (2003), which, in addition to the 40 greyscale natural images, provides a corresponding ground truth contour map drawn by a human for each image. We chose to use these ground truth maps as our reference, as they offer a standardised example of human-drawn contours. To evaluate variability between participants, we computed performance scores by comparing each participant's segmentation (drawn in the absence of noise) to the corresponding ground truth contour map from Grigorescu et al. (2003).

This approach allows us to assess how closely each participant's segmentation aligns with a single reference observer, providing a consistent benchmark against which we can compare all participants. By measuring the similarity to this reference, we can evaluate the variation between participants and determine the overall consistency in their contour drawings. Figure 3.8 indicates that, relative to the contour maps by Grigorescu et al. (2003), participants generally produced similar ground truth segmentations. Most participants' distributions exhibit comparable shapes and fall within the same range of performance scores along the y-axis. The variation in the performance scores on the y-axis per participant can be attributed to varying level of detail present in the contour maps by Grigorescu et al. (2003) for the individual images.

Since these represent a single observer's interpretation of all relevant contours in an image, it is expected that some images align more closely with participant segmentations than others – similar to the examples shown in Figures 3.6 and 3.7. The distribution of variation appears to be similar across participants. When looking at the median performance scores across participants, we can see that they are also very similar – for seven out of ten participants they cluster between 0.6 and 0.7, with two participants deviating by approximately 0.1. Overall, the overlapping ground truths across all natural images indicate that participants mostly traced the same key contours and shared a similar judgment regarding the level of detail to include.

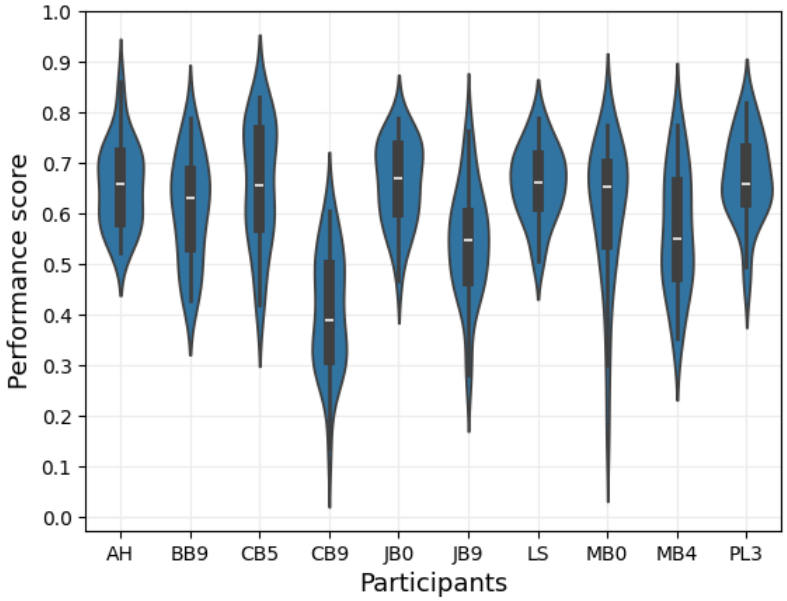


Figure 3.8.: Distribution of performance scores for individual ground truths per participant (the x-axis shows the anonymised participant codes). The width of each curve approximates how often each performance score was achieved by the participant. The interquartile range is indicated inside the curve, with the median performance score marked in white.

A notable deviation can be observed for participant *MB0*, where the curve extends below a score of 0.1. This can be considered an outlier. It was the only case in the experiment where a participant explicitly reported making an error and moved on to the next image before completing the contour tracing. Further, only one participant, *CB9*, consistently achieved lower performance scores compared to the ground truths by Grigorescu et al. (2003).

The reason for this becomes evident when examining this participant's ground truth, shown in Figure 3.9, for the image depicted in Figure 3.7. Compared to the general level of detail captured by other participants, this observer identified fewer contours as relevant.

Figure 3.9.: Exemplary ground truth by participant *CB9*.

Their ground truth segmentations consistently contained fewer contours than the average; however, when examining their contour traces from the first session (e.g., when noise was present at high image contrast), their selection remained consistent. Since the ground truths by Grigorescu et al. (2003) are highly detailed, this participant's simplified segmentations resulted in lower-than-average scores, see Figure 3.8.

However, when evaluating their performance in the presence of noise relative to their own ground truths, their scores were not consistently lower than those of other participants. This highlights the advantage of using participant-specific ground truths to calculate performance scores. This ensures that participants are not penalised for their contour selection choices, as long as they maintain a consistent approach. We will further analyse the importance of this advantage in the following Section 3.4.

Overall, these results suggest that our task design effectively led most participants to a common contour selection strategy without directly specifying which contours to draw for which image. Each image contained a primary focus, like the vehicle in Figure 3.6 or the animal in Figure 3.7, which all participants consistently outlined. The variation is in the extent to which finer details, such as an animal's eyes or background foliage, were traced. Even for these details, there was general agreement on what should and should not be included. Only a small subset of participants consistently drew either more or fewer contours than the average, and when they did, they seemingly maintained this approach consistently.

This consistency suggests that the contour-tracing task was suitable for measuring edge sensitivity in natural stimuli, with potential refinements to further reduce variation in detail selection. We can conclude that our initial concern of participants not focussing on the same contours while assessing contour perception, did not present a significant issue. This confirms the validity of our task design.

3.4. Suitability of Alternative Ground Truths

The previous section examined the variation between observers in contour selection, revealing that participants generally received similar performance scores when their noise-free segmentations were compared to those of Grigorescu et al. (2003). Although some participants deviated from this reference, the majority drew ground truths that resembled the example segmentation provided in the data set.

Given this finding, the question arises as to how effective it would be to use a single set of ground truths to calculate performance scores for all participants. The previous analysis demonstrated that individual ground truths best capture personal differences in contour selection and perceived importance of details. However, it also suggested that, overall, the participants tended to follow a similar pattern. The decision to collect individual ground truths required an additional experimental session. Since longer experiments are more expensive and time-consuming, this increased the data collection effort and potentially limited the number of participants. Using a single set of ground truths for all participants would reduce the experiment duration and resource requirements by half. Therefore, we want to assess the suitability of this approach. To do so, we evaluate how well a common set of ground truths, using the ones by Grigorescu et al. (2003) as an example, compares to using individual ground truths for each participant.

To quantify the differences between these two methods, we computed the performance score differences. Specifically, we calculated the difference in performance scores when a segmentation was evaluated either against the participant's own ground truth or against the ground truth from Grigorescu et al. (2003), which can be seen in Figure 3.10.

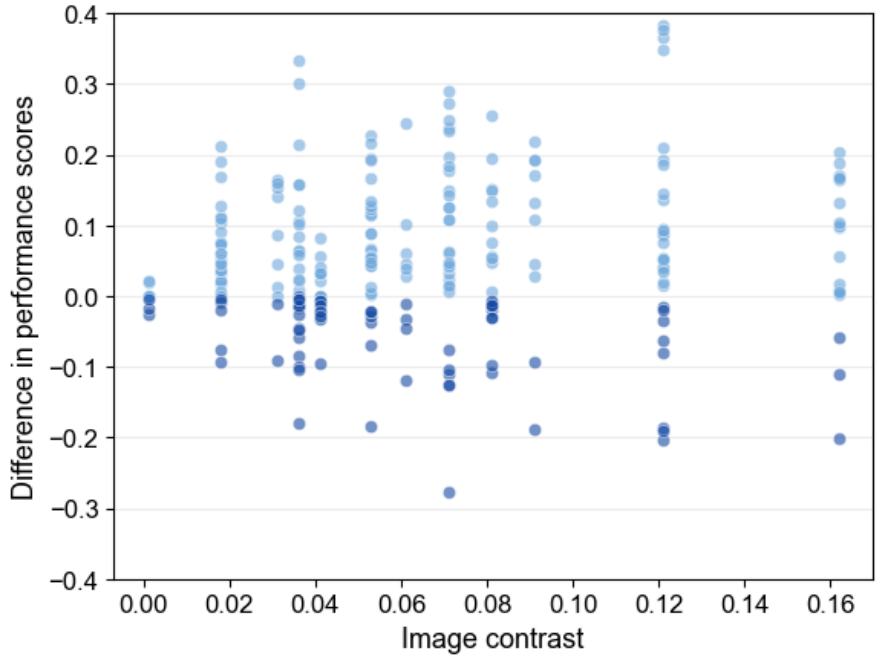


Figure 3.10.: Differences in performance scores – Each data point in the plot represents the difference for a single participant's segmentation of one stimulus condition, comparing the score obtained using their own noise-free segmentation versus the one from the external data set. A positive difference indicates a higher score with the individual ground truth, a negative a higher score with the ground truth by Grigorescu et al. (2003).

Examining the distribution of performance score differences, we observe that the majority of differences are positive. In 169 out of 300 segmentations, participants achieved a higher performance score when their individual ground truths were used as the reference. In 61 cases, there was no difference between the two methods, while in 70 segmentations, the performance score was higher when using the ground truths by Grigorescu et al. (2003).

Beyond the majority of positive differences, we also find that the range of positive differences is wider. Specifically, most performance scores were up to 0.2 higher when calculated using individual ground truths, with some differences reaching up to nearly 0.4. In contrast, performance scores calculated using the ground truths by Grigorescu et al. (2003) were mostly higher by around 0.1, with only one instance almost reaching 0.3. The mean difference across all 300 segmentations (30 per participant) is 0.04, indicating that participants received an average performance score that was 0.04 points higher when their individual ground truth was used as a reference.

To further analyse these differences, we grouped them by the image contrast for which the segmentations were drawn (Figure 3.10, x-axis). This reveals that the differences in performance scores vary systematically with image contrast. At the lowest contrast levels, differences are mostly zero, which aligns with the expectation that if no contours were drawn at all, the choice of reference ground truth becomes irrelevant. As image contrast increases, the variation in performance score differences becomes greater.
This is also expected: when contour perception is severely hindered by low image contrast and noise, participants are likely to draw only a few segmentation lines. In such cases, the choice of ground truth has a limited impact. The most prominent contours are likely to be included in both reference maps, with differences arising in the amount of details that were included. Conversely, when contour perception is less hindered, participants can draw segmentations that more closely resemble the reference ground truth, particularly their own. Under these conditions, the level of detail in the ground truth plays a larger role in determining performance scores.
From this, we can also infer that the specific natural image itself influences the results. Some images inherently contain fewer contours, reducing the number of decisions one must make regarding which contours to trace. In contrast, images with dense foliage or fine details offer a wider selection of possible contours, likely leading to greater variability in individual ground truths. To explore this further, we wanted to see whether we can find natural images that resulted in large differences in performance scores for multiple participants based on the choice of ground truth.
Since noise and contrast conditions were randomly assigned to images, we cannot determine whether specific images systematically led to larger or smaller differences in performance scores across all participants. For instance, if a particular image was randomly shown to all participants at a low contrast level, the resulting small differences in performance scores between the two ground truth methods could be attributed to the assigned contrast level. This means the differences would stem from the contrast rather than the properties of the image itself.
However, we found that for five out of ten participants, the same image (shown in Figure 3.11) had the highest positive difference between individual and common ground truth. While this could also be influenced by the contrast and noise conditions assigned to that image, its frequent occurrence was notable. The ground truth by Grigorescu et al. (2003) for this image is a good example of why we initially decided to have participants draw their own reference segmentations.
This ground truth contains a very high level of detail, which appears difficult to replicate even in the absence of noise, especially in the segmentation of the grass and the buffalo's fur. The rightmost image, displaying the overlap of all individual participant ground truths, clearly illustrates that none of the participants included this level of detail in their segmentations.

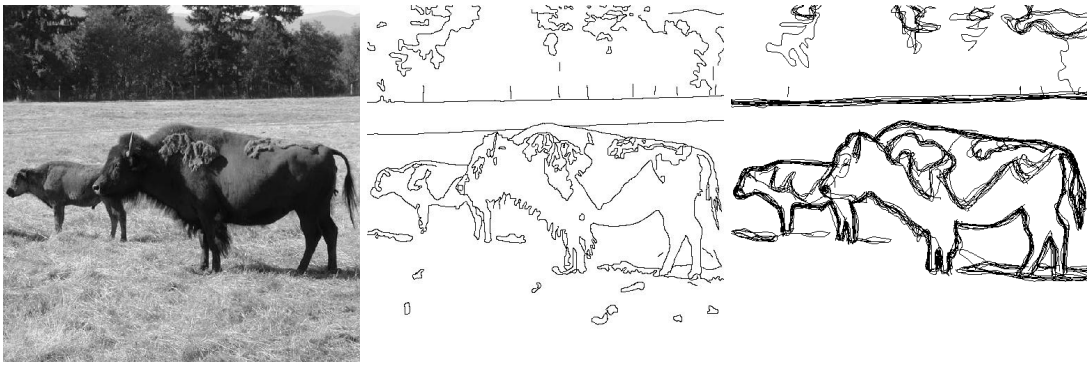


Figure 3.11.: From left to right: Original image, its segmentation map by Grigorescu et al. (2003) and the overlapping participants' ground truths.

Consequently, it is unsurprising that this image resulted in significantly higher performance scores when calculated using individual ground truths. While the buffalo image stands out as one where most participants achieved higher performance scores with their own ground truths, we did not identify a single image where the opposite was consistently true. Although there were some outliers where performance scores were up to 0.2 higher when using the ground truth by Grigorescu et al. (2003), these cases were scattered across different images rather than a single one.

Finally, we assessed whether the choice of ground truth influences the shape of the psychometric functions derived from the performance scores. Our analysis has demonstrated that using individual ground truths generally leads to higher performance scores, better reflecting the impact of noise on an individual's contour perception. However, the average differences remain relatively small. Since our research question investigates whether noise affects edge sensitivity in a consistent manner for controlled and for natural stimuli, we examined whether the psychometric functions fitted to performance scores calculated with the ground truths by Grigorescu et al. (2003) yield similar results. Figure 3.12 presents two exemplary psychometric functions, the full set of functions is provided in Figure A.3 in the Appendix.

The plots for all psychometric functions show that the patterns for both noise types remain similar regardless of the ground truth used, and their relative trends are consistent. However, using individual ground truths results in a curve that spans a broader performance range, with overall higher performance scores, particularly for NB noise of 3 cpd. This means that both the curve and the peak performance level are elevated compared to the other psychometric function.

Specifically, for NB noise of 3 cpd, performance scores calculated with individual ground truths are approximately ten percent higher than those derived from the ground truth set by Grigorescu et al. (2003). We also observe the previously discussed trend: at lower image contrasts, the choice of ground truth has less of an impact on the performance scores, as evidenced by the greater overlap of curves and data points in this range. As image contrast increases, the deviation between the curves becomes greater.

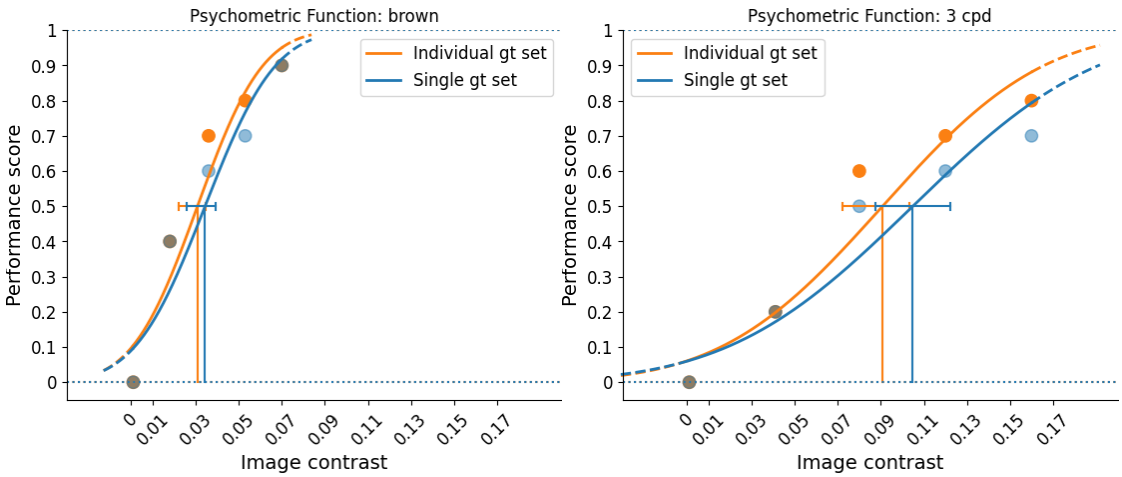


Figure 3.12.: Psychometric functions for brown noise and NB noise of 3 cpd. To remain consistent with our original analysis in Section 3.1, we used scaled performance scores for this comparison. The orange plot represents performance scores calculated with the individual participants' ground truths (*gt*), while the blue plot corresponds to scores derived from the ground truths by Grigorescu et al. (2003).

Overall, using the ground truth set by Grigorescu et al. (2003) appears suitable to analyse the general effects of noise on edge sensitivity in natural images. However, since this analysis was conducted using only one specific ground truth set, these findings cannot necessarily be generalised to other data sets. The suitability of a given ground truth set depends on factors such as the level of detail included and the specific task instructions given to participants. The ground truth set by Grigorescu et al. (2003) works relatively well in our case, likely because our contour definition and task design were influenced by the methodology presented in that study.

Nonetheless, using individual ground truths consistently leads to higher performance scores and reduces outliers in performance drops. This is particularly relevant for participants who systematically include more or fewer details than a single reference map specifies, ensuring that their performance is not unfairly penalised due to individual differences in contour selection.

4. Discussion

Edges are fundamental features of our visual environment, defining the outlines of objects we perceive. As an early stage in human visual processing (Hubel and Wiesel, 1962), edge detection has been extensively studied, particularly concerning the spatial frequencies (SF) most crucial for this process (Foster et al., 1985; Shapley and Tolhurst, 1973; Solomon and Pelli, 1994). However, much of this research has focused on well-controlled, isolated edges. While these studies have provided valuable insights into underlying mechanisms, their generalisability to natural visual behaviour remains uncertain (Olshausen and Field, 2005; Touryan and Dan, 2001).

In this thesis, we investigated how edge sensitivity for simple edges translates to natural stimuli. To achieve this, we conducted an experiment testing contour perception in natural images under different noise conditions and compared our findings to a study that examined edge sensitivity for simple edges. Specifically, we used the study by Schmittwilken et al. (2024) as a reference, which tested sensitivity in a 2-AFC task where observers indicated the location of Cornsweet edges.

While they examined edges at multiple SFs, we focused our comparison on the low SF edge (0.5 cpd) due to its similarity in power distribution to natural stimuli. The use of 2D noise to disrupt stimuli allowed us to probe the selectivity of SF-specific channels in the visual system. Schmittwilken et al. (2024) tested edge detection under three narrowband (NB) noise conditions (0.5, 3, 9 cpd) and three broadband noise conditions (brown, pink, white). We adapted this design for natural images, asking participants to trace contours with a drawing tablet while images were presented with and without noise, based on the experiment design by Sørensen (2023). To standardise responses and minimise different interpretations of what to trace, we provided participants with a contour definition to guide their tracing.

Participants completed two experimental sessions: one with noise-masked images and another with the same images without noise. We quantified performance by comparing the segmentation maps resulting from their drawings using the similarity heuristic by Grigorescu et al. (2003). This allowed us to calculate a performance score as a way to quantify the individual impact of the different noises. To compare the effects of noise between the experiments, we fitted psychometric functions for varying image contrasts, and compared those to the psychometric functions provided by Schmittwilken et al. (2024).

4.1. Edge Sensitivity in Natural Stimuli

Research has shown that cells in the early visual system respond to a limited range of SFs, with the frequency range around 3 cpd being most crucial for edge detection. Our experiment confirmed this, as NB noise of 3 cpd significantly hindered contour perception of our participants in natural images, compared to noise with low and high SFs. This suggests that image features around 3 cpd are critical for edge detection. Our results further aligned with Schmittwilken et al. (2024): apart from NB noise of 3 cpd, pink noise had the strongest effect on edge sensitivity, while NB noise of 0.5 cpd and brown noise had minimal effects.

The difference in effect between pink and brown noise is particularly interesting, as they have similar power distributions. They differ mainly in the amount of power at frequencies below 0.1 cpd, with brown noise having more power at those low frequencies. This supports the conclusion of Schmittwilken et al. (2024) that mostly SFs between 1–10 cpd are crucial for edge detection and that low SF noise does not significantly affect edge sensitivity. The latter being confirmed by the low impact on edge perception by NB noise of 0.5 cpd.

In addition, Schmittwilken et al. (2024) found that NB noise of 9 cpd primarily interfered with the visibility of the 9 cpd edge, but had little effect on the low SF edge on which we focused. We observed a similar trend, as 9 cpd noise did not significantly interfere with contour perception, suggesting that the natural images used lacked relevant image components at this frequency. White noise had an intermediate effect, although its effect on natural stimuli appeared to be relatively stronger than on simple stimuli.

These findings suggest that the presumed mechanisms underlying human edge sensitivity translate to natural images. Our results support the idea that simple stimuli can serve as a reasonable representation of edge perception in natural stimuli, reinforcing the relevance of previous research using controlled stimuli. Furthermore, they demonstrate the possibility of studying visual processes in ways that more closely resemble actual visual behaviour.

4.2. Feasibility of Contour Tracing

Our results show the feasibility of using contour tracing as a task for studying edge sensitivity in natural images. However, the experiment also revealed unique challenges associated with the task design, particularly in defining and analysing contours.

While edge detection for simple stimuli is a binary task, participants in our study had to interpret what qualified as a contour before tracing it. This introduced variability in contour perception. To mitigate this, we restricted the definition of contours, ensuring participants focused on prominent edges rather than fine-grained details and textures.

To investigate how much participants differed in their selection of contours, we analysed their segmentations in the absence of noise, which ensured that each participant was working under consistent conditions. We found that most people understood our contour definition in the same way and traced similar edges, although some people included finer details or omitted certain edges. This confirms that our task design was effective for measuring edge sensitivity in natural images.

Another consideration relates to the calculation of the performance scores. We calculated the performance score against the participant-specific segmentations for the analysis, but also explored an alternative scoring method using the ground truth set from Grigorescu et al. (2003). While participants generally received higher performance scores when using their own ground truths, the overall trends remained similar. This suggests that a standardised ground truth set could be a viable alternative for reducing the experiment duration. However, it also shows that participant-specific ground truths provide a more accurate measure of performance. As most of the participant-specific performance scores were higher, this tells us that participants seemed to be relatively consistent in what they drew across both experimental sessions. This is an important observation, as in the beginning of the experiment, it was not clear how consistent participants would be in their choice of contours. This consistency in contour selection reinforces that the task design worked well for investigating contour perception.

At the same time, the experimental setup presented some practical challenges. Participants used a drawing tablet while viewing a monitor, a method requiring hand-eye coordination that some found difficult. From the participants' feedback on the experience of the experiment, it appeared that those with previous experience of using tablets performed the task more comfortably and quickly, while those unfamiliar with the technology struggled with the fine contour tracing. This may have introduced variability in individual performance. While it seems improbable that providing even more pre-experiment training in future experiments would standardise motor skills across participants, one could consider including this in the demographic questionnaire to analyse a possible impact.

Another limitation was the inability to correct segmentation errors after continuing to the next stimulus. Participants had to press a button twice to proceed to the next stimulus, reducing accidental advancement, but one participant still prematurely skipped a stimulus. While this did not significantly impact results, introducing a way to revisit segmentations could further minimise errors. Additionally, including a mechanism to confirm when participants genuinely did not perceive a contour (rather than accidentally skipping it) would improve data reliability.

4.3. Using Natural Images as Stimuli

The natural images themselves posed challenges. Our findings highlight the significant impact of natural image properties on contour perception. During pilot testing, we observed high variability in performance scores that seemed independent of stimulus conditions (noise type or image contrast). We suspected that differences in perceptual contrast between images played a key role. To address this, we conducted a second pilot, selecting a subset of images with more similar perceptual contrast in the hope of reducing variability. As we did not select the subset at random, we are aware that we have screened out potentially interesting findings, as consistent outliers are also sources of knowledge.

Despite these efforts, there was still considerable variation in performance scores across the images in the experiment. Some images consistently led to higher or lower scores, across noise or contrast conditions. This shows that certain image features, beyond those directly manipulated in our experiment, influence edge perception. Differences in the perceptual contrast of the original images most likely influenced these deviations, resulting in some images being more affected by noise than others. This shows the importance of testing edge sensitivity with stimuli other than just controlled isolated edges. When using such stimuli, perceptual contrast is not an influencing factor as all stimuli have the same characteristics. It shows that there are other factors that affect our perception that go unnoticed when research focuses only on controlled edges.

Future research could further investigate the influence of natural stimuli and their specific characteristics. If there are quantifiable features in natural images that are consistent with the influence of perceptual contrast, what are the factors that lead to higher or lower perceptual contrast? Is it the distribution of light and dark areas, or are there certain shapes or similar, that help us to see the contours of images better, regardless of noise? Identifying quantifiable features in natural images that affect contrast perception could enhance our understanding of contour detection. As we have identified which images show the greatest variation in performance scores across participants and stimulus conditions, future research could also adjust to the contrast normalisation for these images to better test contour perception with less influence from the natural stimuli and more focus on the influence of the noise patterns.

Another interesting factor for further research may be the different SFs of contours in natural images. While our task focused on large edges, it remains uncertain how noise affects the perception of finer details such as foliage or textures. A greater emphasis on these elements might yield different results, although the current experimental setup would make this difficult due to practical constraints. In addition to motor skills, it takes a long time to draw more detailed contours. If one were interested in testing contour perception of finer details, the number of stimulus conditions and images to be drawn would need to be reduced.

4.4. Future Considerations

Looking ahead, certain aspects of the experimental design could be further optimised. For one, the selection of image contrasts per noise condition could be refined. While our initial pilot study was effective and enabled us to fit psychometric functions for all noise conditions, some adjustments could enhance accuracy. For example, in the case of NB noise of 0.5 cpd, the maximum contrast level could have been set lower, as performance already peaked after the third contrast level. A similar but less pronounced trend was observed for white noise. Our results now provide a more accurate baseline for contrast selection in future experiments. This will enable better tuning of contrast levels than our pilot study initially allowed. This refinement would help ensure that each noise condition is tested at optimal contrast levels, improving the accuracy and sensitivity of future measurements.

Finally, our sample size was relatively small. Small participant groups are common in vision research, as individual trials can provide extensive data (multiple trials per person per stimulus condition). However, our study differed in that each participant encountered each stimulus condition only once, resulting in only ten performance scores per stimulus condition. A larger sample size would allow us to explore aspects like inter-observer variability in greater detail, researching if differences in individual interpretation are isolated outliers, or a more common phenomenon. Future studies should consider expanding the participant pool to increase statistical confidence in the findings.

4.5. Conclusion

In summary, we investigated contour perception in natural images under different noise conditions and compared our results to previous research on edge sensitivity in simple stimuli. Our results revealed that pink noise and narrowband noise of 3 cpd had the strongest impact on contour perception, whereas low-frequency noise (0.5 cpd, brown) had minimal effects. Our findings indicate that the presumed mechanisms underlying human edge perception translate to natural images, with noise effects in natural stimuli mirroring those found for controlled stimuli experiments.

Furthermore, our experiment used contour tracing to measure edge sensitivity in natural stimuli, for the lack of an existing standard procedure. While it introduces interpretative challenges, our results suggest that participants generally prioritised similar contours, making the task a reliable measure of edge perception. This establishes contour tracing as a viable method for measuring edge sensitivity in natural stimuli.

Finally, our analysis underscores the importance of perceptual contrast in natural images, emphasising that factors beyond those traditionally considered in edge sensitivity research can influence perception. This highlights the need for further research using natural stimuli rather than relying solely on controlled edges.

References

Campbell, F. W., & Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *The Journal of Physiology*, *197*(3), 551–566. https://doi.org/10.1113/jphysiol.1968.sp008574

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394. https://doi.org/10.1364/JOSAA.4.002379

Foster, K. H., Gaska, J. P., Nagler, M., & Pollen, D. A. (1985). Spatial and temporal frequency selectivity of neurones in visual cortical areas v1 and v2 of the macaque monkey. *The Journal of Physiology*, *365*(1), 331–363. https://doi.org/10.1113/jphysiol.1985.sp015776

Graham, N. V. (2011). Beyond multiple pattern analyzers modeled as linear filters (as classical v1 simple cells): Useful additions of the last 25 years. *Vision Research*, *51*(13), 1397–1430. https://doi.org/https://doi.org/10.1016/j.visres.2011.02.007

Grigorescu, C., Petkov, N., & Westenberg, M. A. (2003). Contour detection based on nonclassical receptive field inhibition. *IEEE Transactions on Image Processing*, *12*(7), 729–739. https://doi.org/10.1109/tip.2003.814250

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding v1? *Neural Computation*, *17*(8), 1665–1699. https://doi.org/10.1162/0899766054026639

Schmittwilken, L., Wichmann, F. A., & Maertens, M. (2024). Standard models of spatial vision mispredict edge sensitivity at low spatial frequencies. *Vision Research*, *222*, 108450. https://doi.org/https://doi.org/10.1016/j.visres.2024.108450

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123. https://doi.org/https://doi.org/10.1016/j.visres.2016.02.002

Shapley, R. M., & Tolhurst, D. J. (1973). Edge detectors in human vision. *The Journal of Physiology*, *229*(1), 165–183. https://doi.org/10.1113/jphysiol.1973.sp010133

Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. *Nature*, *369*(6479), 395–397. https://doi.org/10.1038/369395a0

Sørensen, J.-S. (2023). *The influence of noise on human edge perception in natural images* [Bachelor's Thesis]. Technische Universität Berlin.

Touryan, J., & Dan, Y. (2001). Analysis of sensory coding with complex stimuli. *Current Opinion in Neurobiology*, *11*(4), 443–448. https://doi.org/https://doi.org/10.1016/S0959-4388(00)00232-4

A. Appendix

A.1. Additional Plots – Pilot

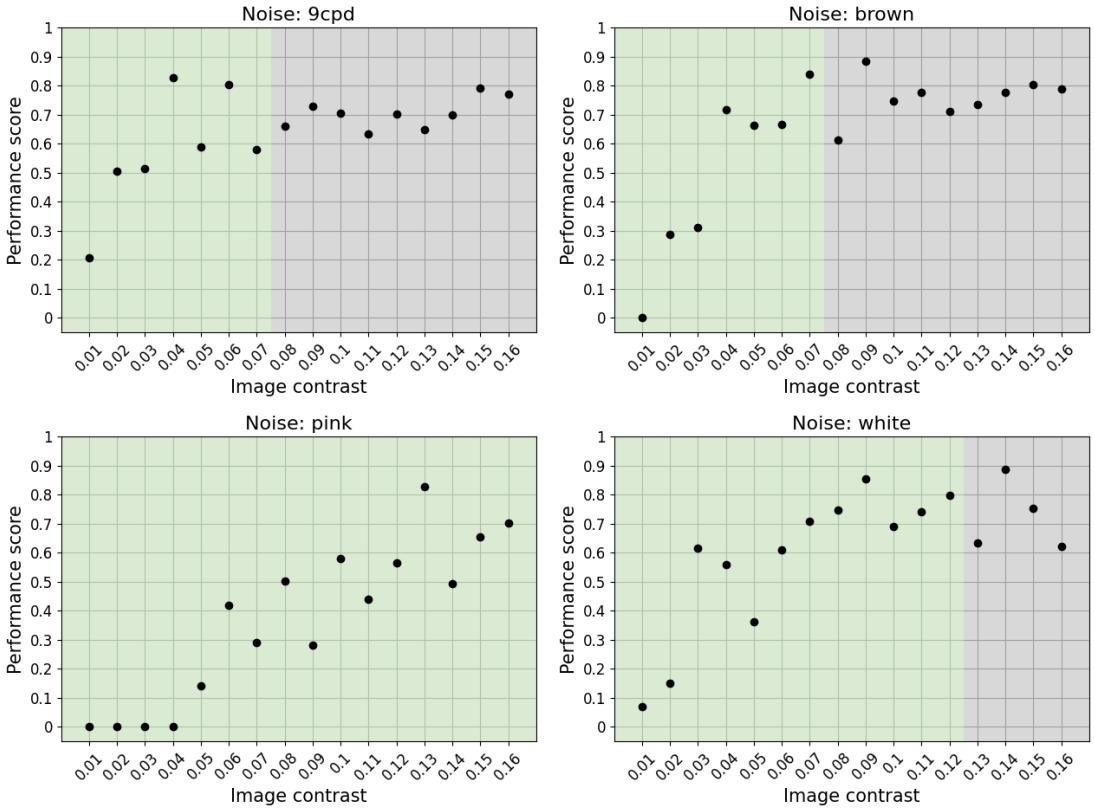


Figure A.1.: Overview of the results for the other noises of the first pilot.

A.2. Additional Plots – Experiment

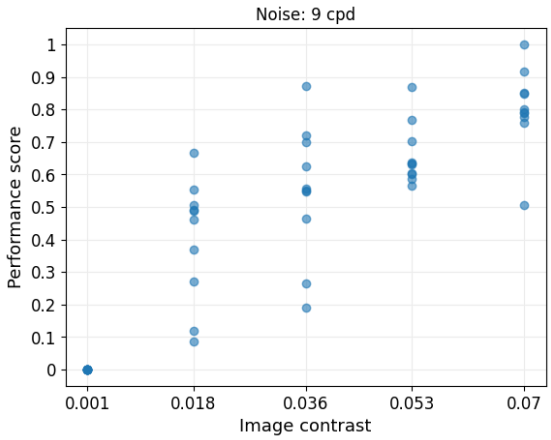


Figure A.2.: Individual performance scores of all participants for natural images with NB noise of 9 cpd, but the scores are scaled per participant.

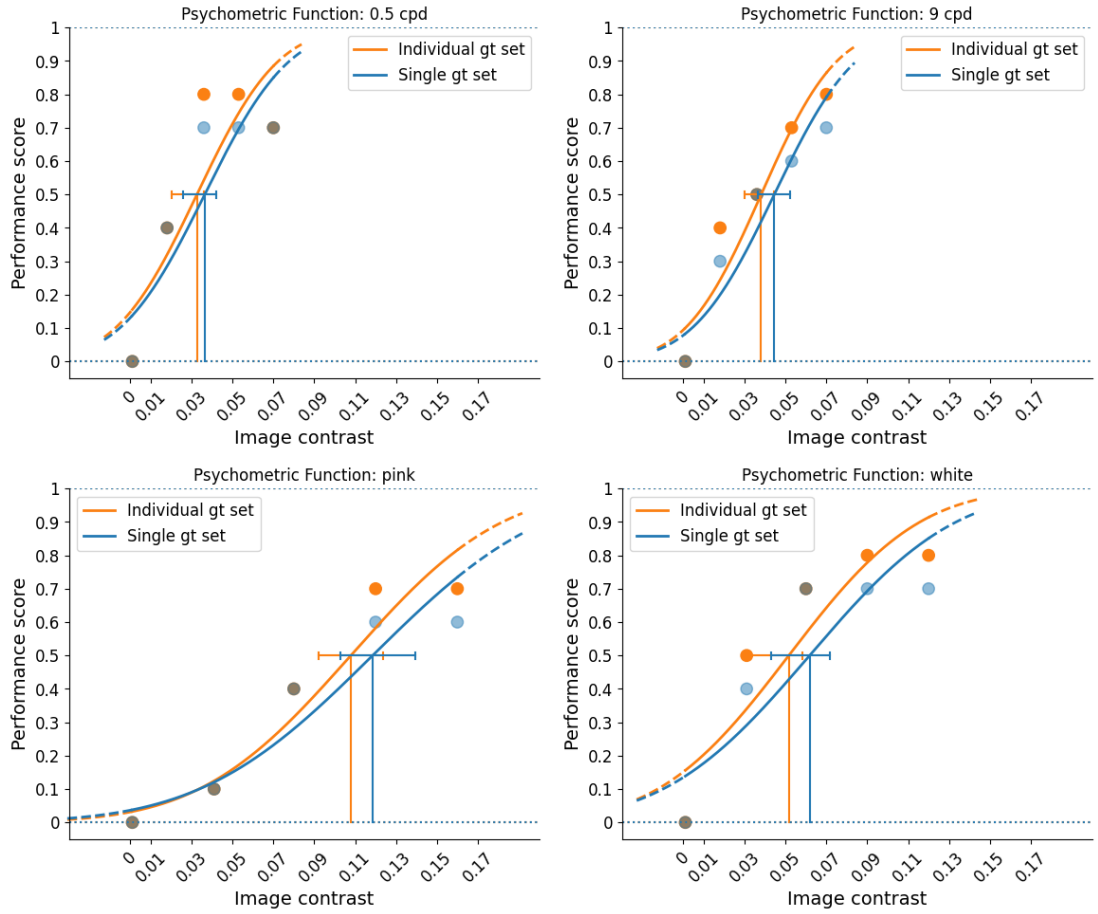


Figure A.3.: Remaining psychometric functions comparing using different of ground truths for the calculation of the performance, *gt = ground truth*.