

Horizon Estimation: Perceptual and Computational Experiments

Christian Herdtweck*

Max Planck Institute for Biological Cybernetics

Christian Wallraven†

Department of Brain and Cognitive Engineering, Korea University



Figure 1: Example images from each of the seven scene classes used in the experiments.

Abstract

The human visual system is able to quickly and robustly infer a wealth of scene information – the scene "gist" – already after 100 milliseconds of image presentation. Here, we investigated the ability to estimate the position of the horizon in briefly shown images. Being able to judge the horizon position quickly and accurately will help in inferring viewer orientation and scene structure in general and thus might be an important factor of scene gist. In the first, perceptual study, we investigated participants' horizon estimates after a 150 millisecond, masked presentation of typical outdoor scenes from different scene categories. All images were shown in upright, blurred, inverted, and cropped conditions to investigate the influence of different information types on the perceptual decision. We found that despite individual variations, horizon estimates were fairly consistent across participants and conformed well to annotated data. In addition, inversion resulted in significant differences in performance, whereas blurring did not yield any different results, highlighting the importance of global, low-frequency information for making judgments about horizon position. In the second, computational experiment, we then correlated the performance of several algorithms for horizon estimation with the human data – algorithms ranged from simple estimations of bright-dark-transitions to more sophisticated frequency spectrum analyses motivated by previous computational modeling of scene classification results. Surprisingly, the best fits to human data were obtained with one very simple gradient method and the most complex, trained method. Overall, global frequency spectrum analysis provided the best fit to human estimates, which together with the perceptual data suggests that the human visual system might use similar mechanisms to quickly judge horizon position as part of the scene gist.

1 Introduction

Recent advances in machine learning algorithms together with the ability of large amounts of labelled data have resulted in significant progress towards the "holy grail" of Computer Vision, that is, to understand the contents of an image. Nevertheless, despite impressive results in single application scenarios (face and person detection, face recognition), general image understanding is still beyond the reach of today's artificial recognition systems. One of the motivators for continued development along the lines of general image understanding is given by the ease and robustness with which we can grasp the contents of an image given only very little time. The first, quick impression of a scene — often referred to as the "gist" — is accessible already after brief presentation and well before even the first eye movements are made to scan further details in the image. Having quick access to such a coarse scene representation before

exploring it in detail is not only an important, beneficial feature of the human visual system, but it would also be adventitious for artificial vision systems, be it for the rough judgment whether an image is of interest in a search or recognition task, or even a loose alignment of images that one may want to stitch together to form a collage. Perhaps more importantly, "scene gist" may offer a solution to the common chicken-and-egg problems like segmentation and object detection, or determining surface color and lighting, by providing an approximate, first solution that can be subsequently improved by further processes.

1.1 Related Work

A large series of psychophysical experiments have determined that the scene gist contains both local and global information (see, for example [Greene and Oliva 2009]): we can see a few prominent objects set in the context of the larger environment, which can be quickly identified as a particular type of scene (outdoors, indoors, navigable, etc.). In addition, we are aware of the rough orientation of the viewer in the scene, that is, the position and angle relative to the pictured scene [Foulsham et al. 2008]. If only visual information is available, this position can only be inferred by indirect cues, such as the angle under which objects appear, or, more importantly, through the horizon position in the image i.e., the direction of "straight-ahead" which is perpendicular to gravity. Previous studies have shown, that despite additional cues to gravity in real life (derived from proprioceptive and vestibular information), the tilt of the horizon has a powerful influence on the direction of the perceived upright. The most dramatic demonstration of this effect is the tilted room experiment, in which a room that is viewed through a tilted mirror severely disrupts our judgment of the downwards/upwards direction. Despite several perceptual studies on the effects of horizon judgments and its conceptual importance for scene understanding, there has been surprisingly little work on how accurately and robustly the horizon *itself* can be estimated, and whether it might be part of the initial scene "gist". A closer understanding of our ability to estimate horizon position in addition to an investigation into potential image cues for this estimate thus forms the first part of the current study.

Given the perceptual studies on scene interpretation, recent work has shown that simple, global image statistics measured in frequency space might actually be able to explain parts of the scene "gist". [Oliva and Torralba 2001] have shown that such statistics are well suited to model coarse scene structure (open, wide landscapes versus closed, hemmed-in cityscapes, for example), that they conform well with human ratings of similar scene properties, and that, in addition, a descriptor based on such statistics can be used for coarse scene classification. This, and similar, later studies [Torralba and Oliva 2002; Oliva and Torralba 2006; Vogel et al. 2007] are remarkable in that they have identified rather simple image measurements that might be computed very quickly and globally by the

*e-mail: christian.herdtweck@tuebingen.mpg.de

†e-mail: wallraven@korea.ac.kr

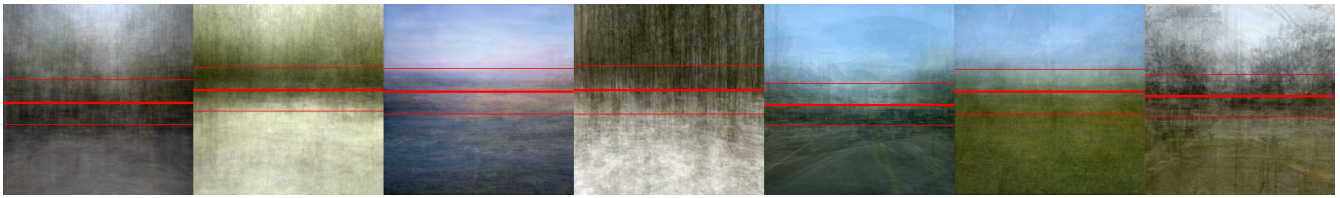


Figure 2: Averages over all images per scene class (from left to right: city, closed nature, coast, forest, non-urban street scenes, open countryside, and other). Each image also shows the mean horizon estimate from all participants across all conditions (thick red line) plus and minus one standard deviation (thin red lines).

visual system over the whole image that could then be used to infer important properties of scene "gist". The second part of the present study therefore asks, whether we can find simple image measurements (or even more complex ones such as [Oliva and Torralba 2001]) to explain the results found in the perceptual experiments.

Despite its conceptual importance for scene understanding and other perceptual tasks (heading, orientation, navigation, etc.), there is surprisingly little work that directly deals with estimation of horizon position in the perceptual literature. There is a large body of work, however, that implicitly uses horizon estimates for investigating, for example, self orientation judgments.

Already in the 19th century, Aubert ([Aubert 1861]) described how a horizontal bar of light seemed to change its tilt when the observer tilted, the first of a series of studies on human self orientation when observer and/or visual stimulus are tilted. This line of investigations also includes the now-famous tilted room experiment by Wertheimer [Wertheimer 1912]. Newer studies like [Dyde et al. 2006; Barnett-Cowan and Harris 2008] have tried to model effects in these estimates for different conditions of tilt and visual stimulation using factors like the perceptual upright and subjective visual vertical. The latter study has performed experiments with and without visual stimuli which could also be tilted. The study found a relative contribution of 14% for vision compared to body alignment and true gravity for estimating the true vertical, and an influence of 3% on estimation of the own body orientation. Interestingly, although often assumed, perceptual estimates of horizontal and vertical directions in the world seem actually not to be orthogonal [Betts and Curthoys 1998].

The interest of the present study, however, does not lie in (body) tilt (i.e., rotations sideways in the fronto-parallel plane) but rather in (body) pitch (i.e., backwards/forwards rotations in the saggital plane), which seems to be a less studied topic. Tilt estimation as a separate task was used in two studies by Tremblay *et al.* [Tremblay et al. 2004; Tremblay and Elliott 2007]. In both studies participants were fixed in a supine position on a bed that could be rotated around its transverse axis, thereby pitching the participants. The task was to estimate the morphological horizon i.e., the direction straight ahead with respect to the body, by manually moving a laser pointer along a fixed arc in the saggital plane, centered at eye level. Other than the laser point the room was dark. The first study [Tremblay et al. 2004] investigated how different instructions influenced participants' estimate of the morphological horizon when oriented upright and pitched backward by 45° . The second study [Tremblay and Elliott 2007] also used a backward pitch of 135° , which placed the head slightly below the feet, and investigated the effect of time after the rotation on the estimates. Both studies showed accurate estimations overall, although there was a slight footward bias of the horizon of $\approx 2^\circ$ which also increased with pitch angle.

In two experiments reported in [Bringoux et al. 2000], participants were pitched (1) from the vertical until they became aware of the pitch direction; and (2) from arbitrary angles until they reported that

they had reached the vertical. The study compared performance of sports experts versus novices with normal and impeded use of somatosensory cues. Errors again were small (from 1° - 16° for (1) and from 2° - 8° in (2)) and depended on expertise and availability of somatosensation. As tilting and pitching happens naturally in aeronautics, several studies were concerned with motion sickness as a factor of participant tilt. In [Klosterhalfen et al. 2008], for example, authors found a vection drum to be quicker to elicit motion sickness in supine (90° tilt) position than upright, however with no or marginal effects of visual patterns on the drum. In conclusion, estimations of one's own body orientation in pitch based on visual and body cues seem to be reasonably precise. However, this does not tell us whether the body orientation can also be perceived from a picture alone, and how the visual horizon can be estimated from that visual data.

Another interesting study, this time *using* the horizon for explanatory purposes, is [Foulsham et al. 2008], in which saccade directions were measured in rotated and upright images. The authors found most saccades near and along the horizon. This study makes explicit what many newer studies implicitly assume: "Coarse information gathered from the first glimpse might also include simple knowledge about the location of the horizon or the overall structure" ([Foulsham et al. 2008]). Such knowledge is assumed in many studies ranging from the famous ecological perception work of Gibson [Gibson 1979] and the work of Sedgwick [Sedgwick 1973; Sedgwick 1980] to newer perceptual studies of [Warren and Whang 1987] and [Rogers 1996]. In addition, many recent developments in scene understanding in computer vision such as [Hoiem et al. 2006; Hoiem 2007; Sivic et al. 2008] assume some notion of horizon.

Finally, it should be mentioned that the horizon is a central topic in literature on photography — especially in recommendations for an optimal, aesthetically pleasing picture layout: among others, the aforementioned study [Foulsham et al. 2008] states that "beginners' photography heuristic suggests that the horizon should be around two-thirds of the way up the picture."

2 Psychophysical Experiment

In our first experiment we addressed two questions: (1) How well can people estimate the horizon in an image if they see it very briefly? and (2) How important are visual cues such as high-frequency information, the global image composition, or data from the upper and lower region of the image for such estimates?

2.1 Stimuli

Our goal was to test horizon estimation in "typical" images covering a wide range of different scene types. We therefore selected 287 color images from a well-known public database (LabelMe [Russell et al. 2008]) and a set of annotated images of natural scenes [Tanner 2009]. Images were chosen such as to be free of (viewer) tilt and to provide a roughly homogeneous distribution of horizon positions (or rather, a homogeneous distribution as estimated by the experi-



Figure 3: Examples of one image in all six manipulations (from left to right: normal, inverted, blurred, lower-, middle-, and upper-subwindow) with expert horizon.

menters) across the middle third of the image. Furthermore, we selected images that provided a sufficient amount of cues for horizon estimation and made sure to include samples from different scene categories. In addition to the 287 color image stimuli, we also selected several further images from the same two sources as training and example stimuli. These images were not shown during the main experiment and were only used during the pre-experiment briefing and the first, few training trials. Each image was hand-annotated for the horizon position with several passes and cross-checks. While these annotations of course do not reflect the real ground-truth for the most difficult scenes, they nevertheless provide a reasonable estimate in almost all cases and allowed us to select images with a reasonable variation of horizon position. The annotations were used in the second part of the analysis of the horizon estimates.

As the LabelMe database provides a rough categorization of images into different scene types, we chose 7 scene categories closely related to that database, with each scene category containing between 18 and 70 images. Each scene category depicted different scene volumes or scales depending on the field of view offered by the surrounding environment. The scene categories were: city (70 images, depicting parts of a street lined with several houses, mid-scale), closed nature (70 images, depicting parts of a meadow in a forest, but always surrounded by forest or other scenery, mid-scale), coast (48 images, typical larger scale images of the ocean, large-scale), dense forest (27 images, depicting the immediate surround consisting of a few trees well within the forest, mid-scale), non-urban street scenes (24 images, cross-country roads with a wide panorama, large scale), open countryside (43 images, panorama shots of landscapes, large-scale), and other (18 images, mostly by the river-side, mid-scale). Examples for each scene can be found in Figure 1.

Given that most images were taken from normal eye-height and some of them (mostly from the LabelMe database) also with the typical photographer’s eye, we wanted to get an idea of how similar images in these categories were. For this, we simply averaged all images used in the experiment for each category — the results are shown in Figure 2 (which already shows the mean and standard deviation of participants’ horizon estimates). In almost all cases, the averaged images seem to provide a good impression of the overall scene category as well as the rough scene layout and scale. In addition, the position of the horizon seems to be well-defined even in these images, which lack high-frequency content. This simple averaging thus already provides an intuition that analysis of (at least some properties of) the scene gist might rely much more on low-frequency, global image cues than on high-frequency, detailed ones. This observation is another confirmation of the results obtained by [Oliva and Torralba 2001] on the same image material, who were able to *synthesize* plausible images from global frequency statistics.

2.2 Stimulus manipulations

In order to specifically test for image cues that might help to make a fast decision about the horizon position, we created six stimulus manipulations for each image (see Figure 3 for one image shown in all six conditions). In the experiment, every image was shown

once in each of these six conditions, randomized over the experiment. The conditions were designed to disrupt different sources of information within the image. An inverted condition was included, as inversion is known to affect holistic processing of scenes [Kelley et al. 2003] which might be also important for scene gist. Local information, such as edges, however, are kept intact by inversion. As an earlier study has shown that scene categorization depends on both local (high-frequency) and global (low-frequency) information [Vogel et al. 2007], we also included a blurred condition. In this condition, images were convolved with a gaussian filter of width $\sigma \approx 10 \times 10$ pixels (depending on the image resolution) which removed fine, high-frequency information from the image but kept the global image layout unaffected.

Finally, fast perceptual access to the horizon presumably depends on the information around the horizon position itself (such as the vanishing point of perspective lines, a light-dark transition, etc.). In three cropped conditions, only the upper, lower or middle two thirds of the image were kept, thus restricting the use of information above or below the horizon or both. To avoid recognition of these conditions by the changed image aspect ratio, a random number of pixels between 0 and 1/3 image width was removed from the left image border, and the amount remaining to one third image width from the right image border, resulting in an image cropped to two thirds of its height and width. This subimage was then resized to fill the same amount of space on the screen as the original using bicubic interpolation¹. Care was taken that in all cases the horizon position (as indicated by the annotation) would be well visible in the picture.

2.3 Training and horizon definition

Since some participants in an earlier pilot experiment expressed concern about the lack of a clear definition for horizon estimation, every participant of the present experiment was first informed about the formal definition of the mathematical/astronomical horizon as the intersection of a perfectly horizontal plane (i.e., perpendicular to gravity) at eye-height with a huge sphere centered around the participant (c.f. Figure 4). It was made clear to participants that the horizon in an image can be sometimes fully occluded (Figure 4a or hard to be inferred precisely in case of tilted ground surfaces, yet still can be estimated e.g., by finding points of the same height above a flat ground (like other peoples’ eyes) or by estimating the vanishing point which lies on the horizon. A set of scene images not used in the later experiment was used as explanation material. Again, this briefing was only done to give participants more explicit background knowledge about the definition of the horizon and to make them more comfortable about the later experiment. In the experiment itself, we were aiming at very brief presentation times together with short response times, which should minimize any high-level cognitive influences as much as possible. Finally, before entering the experiment, participants were familiarized with

¹We are well aware of the fact, that this changes some of the local scale information in the image. However, there were no other straightforward options that could be used which would not immediately give away the different conditions just by the size of the image on the monitor

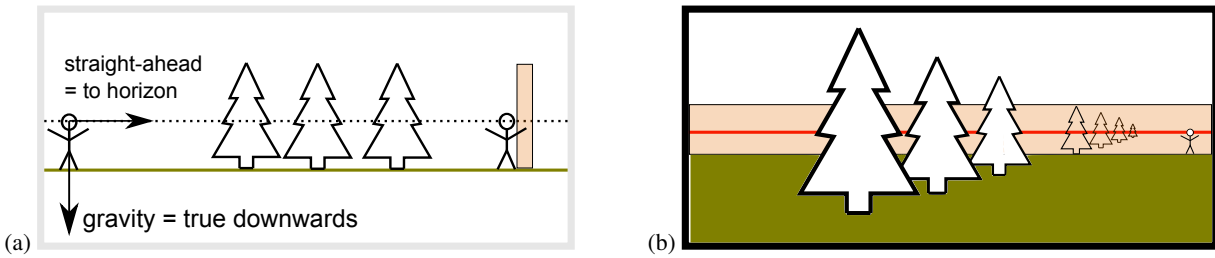


Figure 4: Drawings used before the experiment to (a) explain problems when estimating the horizon (it is usually occluded) and (b) available cues (on flat ground, the horizon is at people’s eye height; trees would be infinitely small there).

the experiment interface in a last short training block and saw at least one image in the blurred and inverted condition to avoid confusion during the experiment. For this training block, we also used different images than during the main experiment.

2.4 Experiment setup and task

Participants were seated in a dark room in front of an ordinary computer with a CRT monitor (distance: 63cm, viewing angle: 34°). The experiment consisted of four blocks of equal length. For each participant a pseudo-random order of images and conditions was created such that between two presentations of the same image there were at least four other images. Matlab (The Mathworks, Inc., Natick, USA) together with PsychToolbox 3.0 [Brainard 1997; Kleiner et al. 2007] were used to ensure precise timing. Images were presented centrally, with the image height filling half of the screen height. Before presentation, a fixation cross was shown for 400ms in the middle of the screen. The rest of the screen showed a 50%-gray. After 150ms the image was masked by a pixel-scrambled versions of the same image, which remained on the screen for 350ms. Finally, the mask was replaced by a dark grey outline of the image border and a mouse ‘cursor’ in the shape of a blue horizontal line spanning the whole screen was displayed.

Participants were asked to move the line to the position where they estimated the horizon and click the left mouse button as quickly and accurately as possible. They were also instructed that they should use the right mouse button if they were not able to make an estimate, and that they were allowed to click outside of the grey image outline if they estimated the horizon to be outside the image. After each response, the grey image outline disappeared and was replaced by a horizontal bar of 16 pixels width into which participants had to move the blue line with their mouse and click in order to start the next trial. This bar’s vertical position was randomized with the exception that positions too close to the estimated horizon position were avoided. This was done to ensure that participants would start the mouse movement towards the horizon position from different, randomized vertical positions, forcing participants to make a mouse movement for every trial. Since participants might mistake the grey bar for some kind of feed-back on their response, the experimenter emphasized that this bar was no feed-back and, indeed, was positioned at random for every trial. Participants were also informed that their reaction time, vertical click position and used mouse button were registered and saved together with a participant id. Since each of the four experiment blocks took on average 31 minutes, short breaks for relaxation were explicitly allowed before starting a new trial but not when an estimate should be given.

In total, twenty paid participants (12 male, 8 female) took part in this experiment. They all had normal or corrected to normal color vision and a mean age of 28.05 years (std: 8.0). All participants received standard rates of 8 Euros per hour for their participation.

2.5 Results

Data from two participants was removed because they misunderstood the instructions in the inverted condition – both verbal accounts and results show that they did not click at the position on the screen where they had seen the horizon in the inverted image, but instead had tried to mentally invert the image and then clicked at the position where the horizon would have been had the image been shown upright.

For the analysis, position data was first normalized to the height of the image with a value of 0.0% indicating the upper image border, and a value of 100.0% indicating the lower image border. Furthermore, all values were rectified such that they refer to the upright, original image size (for example, a value of 60% in the inverted condition refers to an original response of the participant of 40%).

In addition, all right-click answers (indicating that the participant was not able to decide on the position of the horizon) were removed from the position estimation data. Overall, these accounted for $2.8\% \pm 0.06\%$ of all trials – a fairly low number indicating that participants were overall capable of doing the task even given the short presentation time of the stimuli.

2.5.1 Position estimates

On average, and over all conditions, participants chose a position close to the middle of the screen with a value of $48\% \pm 12\%$. We conducted a one-way ANOVA with factor ‘condition’ on the position estimates². The effect of the different stimulus manipulations was found to be highly significant: $F(5, 85) = 15.25, p < 0.001, \eta^2 = 0.49$. The data for this is plotted in Figure 5a. Post-hoc comparisons of the different conditions using the Scheffé criterion show that the mean estimates for all conditions are significantly different, *except* for the normal and blurred conditions.

As can be expected from looking at Figure 2, the average horizon position is put somewhere at the middle of the image. Interestingly, we found that all stimulus manipulations resulted in a change in horizon estimate, except for the normal and blurred conditions, which gave similar results. This already hints at the importance of low-frequency information for making the horizon judgment. In addition, the inverted condition affected estimates, showing that turning the image upside-down changes the processing capabilities of the visual system.

The three subwindow conditions performed as follows: the middle subwindow condition gave very similar (although slightly lower) values than the normal and blurred conditions, whereas the lower subwindow condition resulted in a downward shift, and the upper subwindow condition in an upward shift, respectively. There

²Note, that the factor ‘scene type’ cannot be used here, since we were not able to guarantee an equal spread of horizon estimates for all 7 scene types.

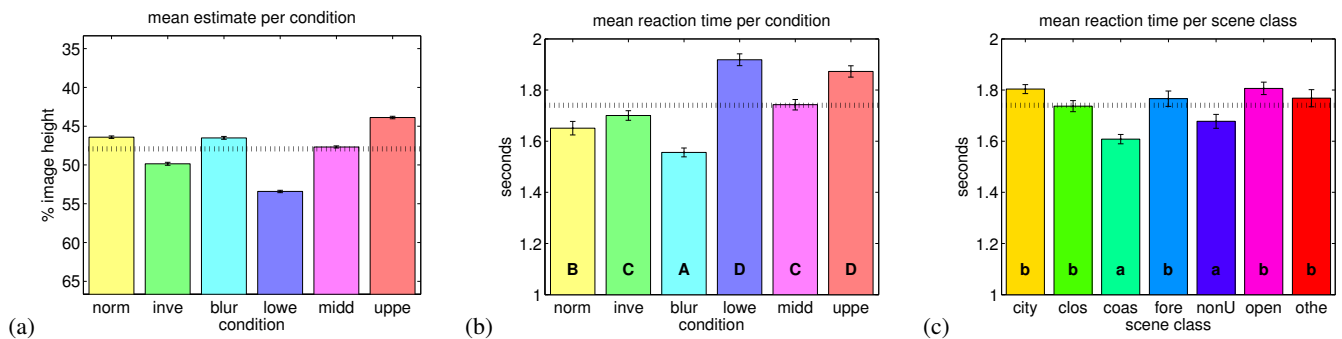


Figure 5: (a) Position estimates across stimulus manipulations. Reaction times across (b) stimulus manipulations and (c) scene category. Letters in bars indicate subgroups as identified by a post-hoc test.

are two possible reasons for this effect: first, the horizon estimate in most cases will depend on image information from below the horizon line, as well as information from above the horizon line — decreasing either information by cutting away image cues will likely impede estimates and result in a higher weight given to the present information and a subsequent adjustment towards the preserved information. Secondly, despite the fact that we made sure that the horizon position was set in the displayed subimage, it could happen that its (annotated) position was close to the image border. Although we told participants that they were allowed to use the full image frame for their responses (and even could click outside, if they wished), and no participant indicated to have problems with doing so in the debriefing, perhaps they were still “shying away” from those extreme values at the image border. Further experiments are needed to disentangle these two effects in more detail.

Finally, we looked at the distribution of horizon positions across scene categories — the average estimates together with their standard deviation are shown in Figure 2 on the averaged scene images. As expected, there is little variation in both overall position of the horizon and its variation.

Whereas this analysis shows that participants answered differently depending on condition and thus demonstrates that certain image cues have different effects on the estimate, it does not necessarily tell us how consistent they were in their answers. That is, the analysis of the raw position estimate data does not yield insights into how well participants were able to do the task. In the following, we therefore look at reaction time and the consistency with which participants performed.

2.5.2 Reaction times

On average, participants took $1.74s \pm 0.74s$ to make a judgment. Given that the reaction time was measured from the onset of the picture frame to the mouse click and therefore included the time of the mouse movement itself, this movement might be a factor in the reaction time. A correlation of the amount of screen space that participants traversed with the reaction time did not reveal a large effect, however, indicating the participants had made their decision well before the initiation of the mouse movement and were quickly moving to their intended horizon location.

Similarly to the position estimates, we analyzed the reaction times using a two-way ANOVA with factors ‘condition’ and ‘scene type’. Both main effects were significant (condition: $F(5, 85) = 11.18, p < 0.001, \eta^2 = 0.39$; scene type: $F(6, 102) = 3.49, p < 0.01, \eta^2 = 0.17$). In addition, we found an interaction of condition and scene type ($F(30, 510) = 2.23, p < 0.001, \eta^2 = 0.11$). The data for the two main effects is plotted in Figure 5(b),(c). As could be expected, participants also got faster during the experi-

ment, however, a separate analysis showed that there was no interaction of the effect of block with any other factor. The post-hoc Scheffé criterion identified four clear subgroups for the stimulus manipulations: the blurred condition (A), the normal condition (B), the inverted and middle subwindow condition (C), and finally the upper and lower subwindow conditions (D). For the scene categories, the two most pronounced subgroups were the non-urban street and coast scenes (a), versus the remaining scene types (b).

Reaction time results show clearly that participants were fastest in the blurred condition, which — taken together with the previous results on position estimates — speaks very much in favor of global processing strategies for horizon estimation given the briefly presented information. With a little added reaction time, the normal condition came next. Interestingly, all subwindow conditions took more time, with the upper and lower subwindow conditions in most cases incurring the longest response times. This finding corresponds very well to the first hypothesis outlined above, namely that participants had to invest more time to extract horizon positions from the reduced information available in the subwindows. This effect was even visible for the middle subwindow condition in which (in most cases) information was available at both sides of the horizon — by a reduced amount. Finally, the inverted condition — while definitely slower to respond to — fared better than all of the subwindow conditions showing that scale changes might have a more detrimental effect even than inversion.

In terms of scene types, clear effects were found for a separation of large-scale, “easy” scenes such as coasts and non-urban, panoramic street scenes. In both cases, the horizon was well-defined and clearly visible due to strong perspective cues and little occlusion. The more objects the scenes contained, the longer participants took to parse all available cues resulting in higher response times even for larger-scale scenes such as the open country scenes. The observed interaction was mainly due to the fact that the upper and lower subwindow conditions (with the highest average reaction times) had different effects on each scene type: for coast and non-urban street scene they mattered less than, for example, for city scenes.

2.5.3 Correlation

So far, we have treated the different participants as “random factors” in the analysis. Given the fact that each participant actually made an informed judgment about the horizon position, we might ask how consistent they were in their judgments. One of the easiest measures that can be used to check this is the correlation coefficient. We therefore calculated the Pearson correlation coefficient between all single trials for all pairs of participants. The average correlation value obtained in this manner is $r = 0.42$, a reasonably high correlation given that this is a *trial-by-trial* correlation.

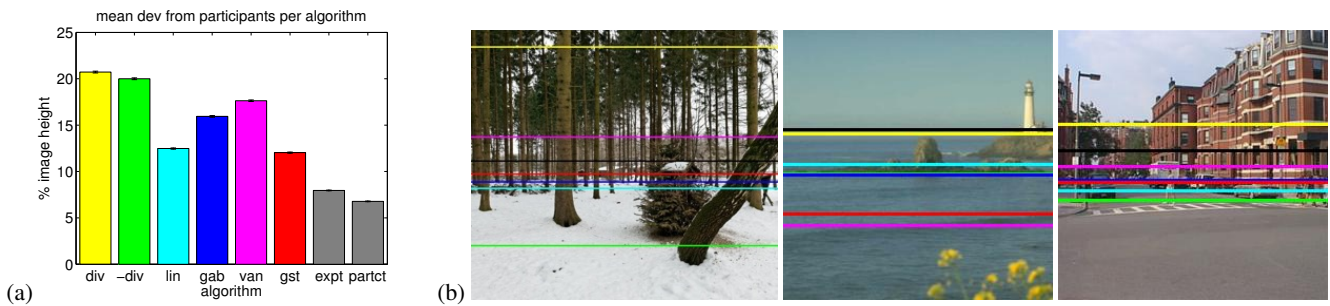


Figure 6: (a) Mean deviations of computational estimates from participant estimates, shown across algorithm, including mean deviation from expert and mean subject rating for comparison; (b) Three example images comparing computational and participant estimates (black line: human data, colored lines: computer data following the color scheme of (a))

Another, related, measure is the intraclass correlation coefficient (ICC). Following the taxonomy introduced by [Shrout and Fleiss 1979], we used the so-called ICC(2,k) coefficient, which measures interrater agreement of several raters in terms of both consistency and absolute agreement. This analysis yielded a value of ICC(2,k)=0.91, which is a high amount of agreement even given the larger number of raters (n=18) that entered the analysis. As it is well known that the number of raters will increase the ICC coefficient (similarly to the effect that averaging data across conditions will increase measured correlations as individual variations are averaged out), we also performed the analysis for all single rater pairs and obtained an average value of ICC(2,k)=0.56. Overall, agreement between participants about the position of the horizon in the images therefore seems to be high. Interestingly, when splitting the analysis across different conditions, the lowest agreement was found for the inverted condition with the remaining conditions being roughly equal.

2.5.4 Comparison with annotations

Since our stimulus images were taken without measuring the camera tilt, we do not have information about the true horizon in the stimulus images. However, we have tried to carefully annotate the horizon position for all images used in the experiment. To get some absolute measure of estimate error, we therefore subtracted participants' results from this "expert" rating and took the absolute value, referring to this difference as the "error" in the following. Overall, the average error was $8\% \pm 6\%$.

Similarly to the previous analyses, we performed a two-way ANOVA with factors 'condition' and 'scene type' on the error in position. The overall error was significantly different from zero ($F(1, 17) = 340.84, p < 0.001$). In addition, both main effects were highly significant (condition: $F(5, 85) = 6.77, p < 0.001, \eta^2 = 0.29$; scene type: $F(6, 102) = 8.66, p < 0.001, \eta^2 = 0.34$). In addition, we found an interaction of condition and scene type ($F(30, 510) = 5.70, p < 0.001, \eta^2 = 0.25$). Post-hoc Scheffé criteria indicated highly significant differences between all conditions, except for the upper subwindow and blurred condition. Stable subgroups for scene type consisted of coast scenes, followed by non-urban street scenes, open country scenes and others, followed by closed outdoor scenes, city scenes, and, finally, forest scenes.

First of all, we found a significant error of 8% on average indicating that participants were rather close to our annotations with absolute errors up to roughly a tenth of the image height. The data for the conditions shows small deviations for the middle, lower, and normal conditions with the highest deviation for the inverted condition — again, together with the previous reaction time and correlation results, indicating that inversion does, indeed, have an effect on scene processing.

The data for scene types follows a similar pattern as for the reaction time data: we found the least differences for the larger-scale scene types with progressively more deviation the more closed the scene type became. Forest scenes, which offer little cues for horizon estimation, resulted in the largest errors (and the longest reaction times) — presumably because these scenes require more processing to make a robust judgment about the horizon position. Finally, the interaction seems to be mostly due to the fact that errors for all conditions were equally low for coast scenes, with other scene types incurring more changes across conditions.

3 Computational Experiment

In the previous, perceptual experiments, we found that — despite changes in image presentation and scene content — human observers seemed quite capable of producing a robust estimate of the horizon position given only a short glimpse at an image. In order to further understand what types of cues might be used in these judgments, we implemented several horizon estimation algorithms and investigated how well participant responses and computational estimates agree. A high agreement in this experiment would not only be beneficial for Computer Vision algorithms, but might also be indicative of similar mechanisms occurring in the human visual system.

3.1 Algorithms and methods

We used the same images as in the perceptual experiment with the same stimulus manipulations. For the computational analysis, we first decomposed all images into the CIE $L^*a^*b^*$ color space [Wyszecki and Stiles 2000]. Due to space constraints, in the following analysis, we focus only on the results of the luminance channel, disregarding color effects.

Several methods for horizon estimation were implemented using a range of complexity from very simple approaches such as finding the line with maximum vertical gradient to more complex estimates using previously trained frequency spectra of different scene classes. More specifically, we implemented the following algorithms:

- [lin]:** computes the vertical gradient of the image, smoothed with a gaussian of 10 pixels width, which roughly corresponds to finding horizontal lines. The algorithm then sums up the absolute gradient strength for each image row and determines the maximum row
- [gab]:** determines the horizon as the line corresponding to the maximum response to a horizontal gabor filter with wavelength $\sigma = 10$ pixels, and a bandwidth and orientation selectivity of 0.5
- [div]:** for every image row, calculates how well it divides the image into an upper region of high values and a lower region of low values. It also takes into account how high the (signed) vertical gradient is at that position to find the optimal division between a bright region (sky) and a dark region (ground)
- [-div]:** similar to [div], except that it looks for an upper region of

low values and a lower region of high values with the gradient at that partition to be as negative as possible;

[van]: tries to find the vanishing point by applying the canny edge detector and calculating the Hough transform of the resulting edge image, removing near-to-vertical and horizontal lines and adding up lines to determine the point where most lines converge. Again, values are added up over all image rows. This approach was expected to work well in human-structured environments like street scenes, where many clear perspective cues from lines are available.

[gst]: uses the Horizon detector contained in the Matlab LabelMe toolbox [Russell et al. 2008] which was employed by [Hoiem et al. 2006; Sivic et al. 2008]. This employs the "gist descriptor" developed by [Torralba and Oliva 2002] as a feature for a mixture of linear regressors, trained on a set of approx. 1500 images with annotated horizons of different scene types that were not used in the experiment. The gist descriptor creates a low-dimensional representation of the scene structure, based on the output of filters tuned to different orientations and scales. This is achieved through a wavelet image decomposition with a steerable pyramid tuned to several orientations and scales. The descriptor has been shown to model some aspects of scene perception [Oliva and Torralba 2006], but also has its limitations for more complex processing [Vogel et al. 2007].

3.2 Results

To compare computational with perceptual data we calculated the absolute deviation of each computational estimate with each participant's analogue, arriving at (number of algorithms) \times (number of participants) results for every image and condition. These deviations were averaged over participants, conditions and images to compare overall algorithm agreement (Figure 6a), as well as over participants and images to compare deviations per algorithm and condition (Figure 7a), and finally over participants, conditions and images per scene type to give one deviation measure per algorithm and scene (Figure 7b). For comparison we added the corresponding deviations of participants from their mean rating as well as their deviations from the expert rating.

Overall computational deviation from participant estimates shows (again) that Computer Vision has much to learn from human data — the deviation within participants as well as from the expert rating is much smaller on average than that of the computational estimates. Among these, however, there are also clear differences with the very simple division of the image into a bright and dark region above/below the horizon (**[div]**) showing the greatest deviation with more than 20% of the image height on average. Surprisingly, disregarding light and dark and simply using absolute brightness change as a horizon cue (**[lin]**) works much better and reaches deviations of $\approx 12\%$, nearly as low as the best algorithm (**[gst]**) and much better than the more complicated filtering with a gabor filter (**[gab]**) or estimation of the vanishing point (**[van]**). As can be expected, since it works on trained horizon data, the best agreement is achieved by the regressor on the spatial envelope (**[gst]**), indicating that a holistic image interpretation better explains participant behaviour.

Overall, the blurred condition showed the highest deviation from human data, followed by inverted and normal conditions. The subwindow conditions showed much better agreement with middle being the best. However, this could also be due to the scaling present in the subwindow conditions: algorithms will always estimate the horizon inside the image and therefore had a smaller chance of being wrong when presented with a smaller image part. The pattern differed significantly by algorithm, however: whereas the **[lin]** and **[gst]** algorithms show comparatively little variation across conditions, deviations for the vanishing point algorithm (**[van]**), for example, are much lower in the subwindow conditions. This might indicate that the information needs to be pooled over a smaller region in order to result in more stable estimates. The Gabor-filter

estimate (**[gab]**) has quite low deviations overall, but it cannot handle the blurred condition. It remains to be tested, whether different filter sizes could lead to better prediction of human estimates. Interestingly, the gist algorithm (**[gst]**) also fares worst for the blurred condition indicating similar problems with its main frequency band — in addition, the training data seems to have restricted the algorithm very much, leading also to less accurate predictions in the inverted condition.

For scene category, we found less within-algorithm variation than for conditions. Non-urban roads were most similar to participant estimates, followed by the 'other' class and one homogeneous subset consisting of the classes city, coast, open country and closed nature, with the highest deviation found for the forest scene type. This pattern follows that which one would expect from absolute algorithm performance given the availability of perspective features and a clear horizon in images of these scene classes. For non-urban roads, for example, both are easily available, for all other classes except forest either perspective cues or a clear wide horizon can usually be found. This can also be seen in Figure 7b: the vanishing point estimation (**[van]**), for example, did best agree for the non-urban roads and city class; algorithms **[div]**, **[lin]** and **[gab]** which benefit from clearly visible horizontal lines showed most similarity for the non-urban road, coast and open country classes.

This pattern can also be observed in the three examples comparing computational (colored) and human (black) estimates for the normal condition, which are depicted in Figure 6b. Although the average human response does deviate from the true horizon in the right-most image of a city scene, overall it provides a reasonable estimate of horizon position. Computational estimates show larger variability, especially in the forest example. Furthermore, although a few errors are 'understandable' such as estimating the horizon at the bottom of the cliff in the middle, coast image, sometimes locations are determined that humans would dismiss at once.

4 Discussion

Humans are surprisingly robust at inferring important scene properties given only a briefly shown image. Among these properties are not only the type of scene that is depicted (indoor, outdoor, mountain, coast, etc.) and salient object categories, but also information about the general scene layout. One important cue that belongs to the scene layout is the horizon position and orientation in the image as this relates to the viewer orientation and also to the scene structure itself. So far, however, little is known about whether we are able to estimate the horizon reliably and quickly as part of the scene gist.

In our first experiment, we have shown that humans can estimate the horizon position given an image presentation of only 150 milliseconds. Agreement among participants in this experiment was high showing the robustness with which such estimates can be made. In addition, we showed that blurring an image did not affect the estimates, rather to the contrary, the blurred condition also had the fastest reaction time showing that global, low-frequency image cues might be enough to make the decision about the horizon position. This result is supported by the fact that restricting the amount of information in the subwindow conditions not only resulted in different estimates (compared to the normal position) but also in increased reaction times. Our findings are consistent with previous results from other studies e.g., [Oliva and Torralba 2001; Vogel et al. 2007] which indicate the importance of global processing for scene categorization. In addition, we found that image inversion resulted in longer reaction times, less agreement among estimates, and increased error with respect to our annotations. Following a previous study on object detection in scenes, where inversion was found to affect performance [Kelley et al. 2003], our result shows

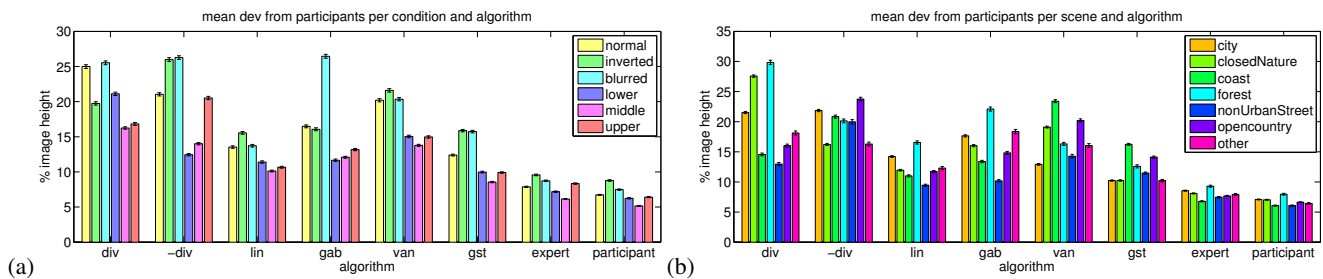


Figure 7: Mean deviations of computational estimates from participant estimates, shown across (a) condition, and (b) scene type; error bars denote standard error.

the importance of global scene consistency also for horizon processing. Further work will investigate whether and how the effects described here change with shorter or longer image presentation times. For example, one might assume that local features might play a more important role once participants have more time to explore the image more closely.

Our computational experiments showed that a surprisingly simple feature based on the most salient vertical gradient (**lin**) was already a good predictor of human horizon estimates. Given that the best feature was much more sophisticated and, in addition, trained in a supervised fashion, this result may suggest that the visual system might actually use very simple heuristics to estimate horizon position. In both cases, however, the best performance was provided by algorithms based on a global analysis of low-frequency statistics (**gst**), which underlines the importance of quick, global processing in making judgments about scene properties. Nevertheless, it should be noted that no computational algorithm was able to fully match human performance — algorithms were off by at least 4–5% of image height with their estimates compared to human annotations. It would therefore be interesting to combine the output of the different feature types — perhaps weighted additionally by their reliability. Such a combination of simple features could eventually approach human estimation quality.

References

- AUBERT, H. 1861. Eine scheinbare bedeutende Drehung von Objecten bei Neigung des Kopfes nach rechts oder links. *Virchows Archiv* 20, 3-4 (Mai), 381–393.
- BARNETT-COWAN, M., AND HARRIS, L. R. 2008. Perceived self-orientation in allocentric and egocentric space: effects of visual and physical tilt on saccadic and tactile measures. *Brain Research* 1242 (11), 231–243.
- BETTS, G. A., AND CURTHOYS, I. S. 1998. Visually perceived vertical and visually perceived horizontal are not orthogonal. *Vision Research* 38, 13, 1989 – 1999.
- BRAINARD, D. H. 1997. The Psychophysics Toolbox. *Spatial Vision* 10, 433–436.
- BRINGOUX, L., MARIN, L., NOUGIER, V., BARRAUD, P.-A., AND RAPHEL, C. 2000. Effects of gymnastics expertise on the perception of body orientation in the pitch dimension. *Journal of Vestibular Research* 10, 6, 251–258.
- DYDE, R. T., JENKIN, M. R., AND HARRIS, L. R. 2006. The subjective visual vertical and the perceptual upright. *Experimental Brain Research* 173, 4 (September), 612–622.
- FOULSHAM, T., KINGSTONE, A., AND UNDERWOOD, G. 2008. Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research* 48, 17, 1777 – 1790.
- GIBSON, J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston.
- GREENE, M., AND OLIVA, A. 2009. The briefest of glances: The time course of natural scene understanding. *Psychological Science* 20, 4, 464–472.
- HOIEM, D., EFROS, A., AND HEBERT, M. 2006. Putting objects in perspective. In *Proceedings IEEE Computer Vision and Pattern Recognition (CVPR)*.
- HOIEM, D. 2007. *Seeing the World Behind the Image - Spatial Layout for 3D Scene Understanding*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213.
- KELLEY, T., CHUN, M., AND CHUA, K. 2003. Effects of scene inversion on change detection of targets matched for visual salience. *Journal of Vision* 3, 1, 1–5.
- KLEINER, M., BRAINARD, D., PELLI, D., INGLING, A., MURRAY, R., AND BROUSSARD, C. 2007. What’s new in Psychtoolbox-3? *Perception (ECVP Abstract Supplement)* 14.
- KLOSTERHALFEN, S., MUTH, E. R., KELLERMANN, S., MEISSNER, K., AND ENCK, P. 2008. Nausea induced by vection drum: Contributions of body position, visual pattern, and gender. *Aviation Space and Environmental Medicine* 79, 4 (APR), 384–389.
- OLIVA, A., AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3, 145–175.
- OLIVA, A., AND TORRALBA, A. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research* 155, 23–39.
- ROGERS, S. 1996. The horizon-ratio relation as information for relative size in pictures. *Perception and Psychophysics* 58, 1, 142–152.
- RUSSELL, B., TORRALBA, A., MURPHY, K., AND FREEMAN, W. 2008. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision* 77, 1-3 (May), 157–173.
- SEDGWICK, H., 1973. The visible horizon: A potential source of visual information for the perception of size and distance.
- SEDGWICK, H. 1980. *The geometry of spatial layout in pictorial representation*, vol. 1. Academic Press New York, 33–90.
- SHROUT, P., AND FLEISS, J. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 86, 2, 420–428.
- SIVIC, J., KANEVA, B., TORRALBA, A., AVIDAN, S., AND FREEMAN, W. T. 2008. Creating and exploring a large photorealistic virtual space. In *First IEEE Workshop on Internet Vision*. associated with CVPR 2008.
- TANNER, T., 2009. A database of 3500 natural images, provided by the Max-Planck Institute for Biological Cybernetics in Tuebingen, Germany; <http://images.kyb.tuebingen.mpg.de>.
- TORRALBA, A., AND OLIVA, A. 2002. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 9, 1226–1238.
- TREMBLAY, L., AND ELLIOTT, D. 2007. Sex differences in judging self-orientation: The morphological horizon and body pitch. *BMC Neuroscience* 8, 1, 6.
- TREMBLAY, L., ELLIOTT, D., AND STARKES, J. 2004. Gender differences in perception of self-orientation: Software or hardware? *Perception* 33, 3, 329–337.
- VOGEL, J., SCHWANINGER, A., WALLRAVEN, C., AND BÜLTHOFF, H. 2007. Categorization of natural scenes: Local versus global information and the role of color. *ACM Transactions on Applied Perception* 4, 3, 19–39.
- WARREN, W. H., AND WHANG, S. 1987. Visual guidance of walking through apertures: Body-scaled information for affordances. *Journal of Experimental Psychology: Human Perception and Performance* 13, 3, 371–383.
- WERTHEIMER, M. 1912. Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie* 61.
- WYSZECKI, G., AND STILES, W. 2000. *Color science: Concepts and Methods, Quantitative data, and Formulae*. Wiley-Interscience.