i-PERCEPTION

a Pion publication

i-Perception (2013) volume 4, pages 122–136

dx.doi.org/10.1068/i0587 ISSN 2041-6695 perceptionweb.com/i-perception

The role of binocular disparity in rapid scene and pattern recognition

Matteo Valsecchi
Abteilung Allgemeine Psychologie, Justus-Liebig-Universität, Otto-Behaghel-Str. 10F, D-35394 Giessen, Germany; e-mail: matteo.valsecchi@psychol.uni-giessen.de

Baptiste Caziot
Graduate Center for Vision Research, SUNY College of Optometry, 33 W 42nd St., New York, NY 10036, USA; SUNY Eye Institute, 550 Harrison Street, Suite 340, Syracuse, NY 13202, USA; e-mail: bcaziot@sunyopt.edu

Benjamin T. Backus
Graduate Center for Vision Research, SUNY College of Optometry, 33 W 42nd St., New York, NY 10036, USA; SUNY Eye Institute, 550 Harrison Street, Suite 340, Syracuse, NY 13202, USA; e-mail: bbackus@sunyopt.edu

Karl R. Gegenfurtner
Abteilung Allgemeine Psychologie, Justus-Liebig-Universität, Otto-Behaghel-Str. 10F, D-35394 Giessen, Germany; e-mail: Karl.R.Gegenfurtner@psychol.uni-giessen.de

Received 25 January 2013, in revised form 22 March 2013; published online 16 April 2013.

Abstract. We investigated the contribution of binocular disparity to the rapid recognition of scenes and simpler spatial patterns using a paradigm combining backward masked stimulus presentation and short-term match-to-sample recognition. First, we showed that binocular disparity did not contribute significantly to the recognition of briefly presented natural and artificial scenes, even when the availability of monocular cues was reduced. Subsequently, using dense random dot stereograms as stimuli, we showed that observers were in principle able to extract spatial patterns defined only by disparity under brief, masked presentations. Comparing our results with the predictions from a cue-summation model, we showed that combining disparity with luminance did not per se disrupt the processing of disparity. Our results suggest that the rapid recognition of scenes is mediated mostly by a monocular comparison of the images, although we can rely on stereo in fast pattern recognition.

Keywords: stereo vision, natural image, scene recognition, random-dot stereogram, visual masking.

1 Introduction

Binocular disparity, along with luminance, color and their changes over space and time, is one of the elements that define our visual world (Adelson & Bergen, 1991). Yet the purpose, extent, and conditions under which our visual systems use disparity are not fully understood.

Valsecchi and Gegenfurtner (2012) demonstrated that binocular stereo can improve the long-term memory for natural scenes, although this is the case only under very specific conditions, i.e. only for scenes where a semantic encoding could not support recognition (forest scenes). Another situation in which binocular disparity could be of advantage is when scenes have to be encoded quickly for immediate recognition. Binocular disparity could contribute to the fast encoding of the scene by enhancing the segmentation of the single objects (Caziot, Valsecchi, Gegenfurtner, & Backus, 2011). Moreover, it cannot be excluded that the disparity of at least some of the objects and surfaces in the scene is directly encoded and contributes to recognition.

It has been shown that disparity can enhance the recognition of objects (Edelman & Bülthoff, 1992) particularly if the viewpoint changes between the encoding and recognition phases (Burke, 2005; Burke, Taubert, & Higman, 2007; Lee & Saunders, 2011), that it has the potential to guide movements within the peri-personal space (McKee, Levi, & Bowne, 1990; Sheedy, Bailey, Buri, & Bass, 1986) and to break camouflage (Julesz, 1971; McKee, Watamaniuk, Harris, Smallman, & Taylor, 1997; Wardle, Cass, Brooks, & Alais, 2010). More recent evidence indicates that stereopsis can be helpful in determining the depth arrangement of objects embedded in scenes and beyond peri-personal space (Allison, Gillam, & Vecellio, 2009; McKee & Taylor, 2010).

Although disparity might contribute to the visual processing of scenes, it might do so with a slower time course as compared with luminance-based shape extraction (e.g. McKee et al., 1990; also see Valsecchi & Gegenfurtner, 2012; Westheimer, 2011). The first question we asked in the present study is thus whether the visual system uses disparity in order to encode scenes for short-term recognition. The second question we asked is whether the visual system relies on binocular disparity when other visual dimension(s) defining the visual scene are greatly impoverished. After finding no evidence for a stereo enhancement of recognition with images of natural and simulated scenes we asked a third question: is disparity actually available to the visual system, but simply not used for encoding scenes when other cues are available?

Researchers have dealt extensively with the combination of disparity and other visual cues in the estimation of depth (e.g. see Landy, Maloney, Johnston, & Young, 1995), in the estimation of slant (e.g. Backus & Banks, 1999; Backus, Banks, van Ee, Crowell, & Crowell, 1999; Banks & Backus, 1998; Girshick & Banks, 2009; Hillis, Ernst, Banks, & Landy, 2002; Hillis, Watt, Landy, & Banks, 2004; Knill & Saunders, 2003), and in the perception of shape (e.g. Adams & Mamassian, 2004; Doorschot, Kappers, & Koenderink, 2001; Liu, Collin, & Chaudhuri, 2000). These studies were typically conducted by having participants judge stimuli along one dimension (depth, slant, shape), often building psychometric curves and estimating points of subjective equality. Tasks requiring a yes/no response, such as detection at threshold were less frequent (but see Ichikawa, Saida, Osa, & Munechika, 2003; Meese & Holmes, 2004). The role of binocular disparity in visual recognition has been largely neglected, except for the study by Liu and colleagues (2000) who found a relatively weak enhancement in recognition rate when faces were defined by both stereo and luminance as compared to luminance alone.

We first tested the relative contribution of binocular disparity to scene recognition by using natural and artificial scenes and manipulating the presence of monocularly available cues, such as the chromaticity of object surfaces. Then, using random dot stereograms (RDS) we investigated the recognition of patterns defined solely in stereo and luminance.

2 Experiment 1: natural scenes

The first question we addressed in the present study is thus whether and how fast disparity contributes to the visual encoding of scenes. To this aim, we applied a modified version of the paradigm which Gegenfurtner and Rieger (2000) used in order to demonstrate that color plays a role in the early processing of visual scenes. In this paradigm, the processing of the scene image was interrupted by the use of a pattern mask. So far, dynamics of disparity processing have been studied with short presentations (e.g. Foley & Tyler, 1976; Harwerth, Fredenburg, & Smith, 2003; see Westheimer, 2011), but asynchronous masking has rarely been used (but see Lehmkuhle & Fox, 1980 for a metacontrast masking example), and we know of only one study that used a post-mask whose 2D location coincided with the one of the targets (Ritter, 1980). Little direct evidence was thus available guiding our choice of the mask to use. Nonetheless, the patterns of interference that have been observed are compatible with the notion that the visual system processes stereo-defined depth effectively as a third spatial dimension (e.g. Butler & Westheimer, 1978; Long & Over, 1974; Tyler & Kontsevich, 2005). We thus decided to use masks overlapping with the targets also in depth, i.e. nonstereo masks for nonstereo targets and stereo masks for stereo targets. Furthermore, we constructed our mask images with a stereo structure that had depth relations consistent with the overlap (occlusions) in the 2D pattern. As targets we used forest images, a category of images for which Valsecchi and Gegenfurtner (2012) showed that binocular disparity enhances long-term memorization.

2.1 Methods

2.1.1 Observers

One group of 13 students from the Justus-Liebig University of Giessen (11 females, mean age: 25.1 years) participated in the study in exchange for payment. Participants in this and the following experiments provided written informed consent in agreement with the Declaration of Helsinki. Methods and procedures were approved by the local ethics committee LEK FB06 at Giessen University (proposal number 2009-0008). All observers were naive as to the aim of the study.

2.1.2 Stimuli

Stimuli were 216 color pictures of forest scenes. The pictures were taken in the Schiffenberger Wald, in the vicinity of Giessen, using a Fujifilm Finepix W1 3D digital camera (Fujifilm Holdings Corporation, Tokyo, Japan). A subset of the images was used by Valsecchi and Gegenfurtner (2012). In the present and in the following experiments, the images were displayed in a mirror stereoscope. The pictures were rescaled to 656 × 492 pixels and shown on a pair of identical 19-inch Dell UltraSharp 1907FP LCD monitors (Dell, Inc., Round Rock, TX) with a 75-Hz frame rate. The monitors were viewed through two orthogonal first surface mirrors (169 × 194 mm). From the effective viewing distance of 55.5 cm, the pictures subtended 19.9° × 14.8°. The vergence demand of the fixation mark specified a distance equal to the viewing distance.

For each trial, a picture was randomly chosen from the database to be used as a target, and another to be used as the distractor. A mask, composed of a superimposition of 50 irregular polygons whose colors were randomly sampled from the target and distractor images, was created for each trial. From direct inspection of a subset of images, we estimated that the largest (uncrossed) disparities were 4.5% of the horizontal size of the image for the farthermost elements that could be individuated, whereas generally no element had crossed disparity relative to the image frame. We thus created our masks by distributing the polygons evenly from the maximum disparity (54.3 arcmin, i.e. 4.5% of 19.9°) to 0 disparity. The furthest polygons were occluded by the nearest ones and their projected area varied along the disparity gradient (from around 64 deg^2 for the furthest polygons to around 2.6 deg^2 for the nearest ones). Mask image generation and the stimulus presentation were carried out using Matlab (MathWorks, Inc., Natick, MA) and the PsychToolbox (Brainard, 1997; Pelli, 1997).

2.1.3 Procedure

The trial event sequence is depicted in Figure 1. First, the fixation square alone was displayed for 1 s in order to obtain proper binocular alignment. Then, the target image was shown for a variable time interval (13, 27, 53, or 80 ms). Subsequently, the mask image was shown for 500 ms. The target and distractor pictures were shown in two successive 500-ms intervals. Participants indicated the target interval with a key press. After the key press, if the answer was incorrect, the fixation point turned red for 1 s.

The images could be shown in one of two modalities: stereo and nonstereo (randomly choosing the left- or right-eye view and displaying it to both eyes). In each trial, the same modality of presentation was applied coherently to the target, mask, and choice images. Each combination of modality of presentation and image duration was tested in 48 trials. Twenty extra pictures were used for practice trials at the beginning of the experiment. The whole experimental session lasted approximately 1 hr.

2.1.4 Data analysis

Throughout the current paper, the proportions of correct responses were transformed to *z*-scores before being analyzed with ANOVAS and *t*-tests. The same transformation was applied to the data before aggregating over observers and calculating confidence intervals. Confidence intervals were calculated as percentiles after bootstrapping the original sample 10,000 times. The aggregated values and confidence intervals were transformed back to proportion correct for plotting.

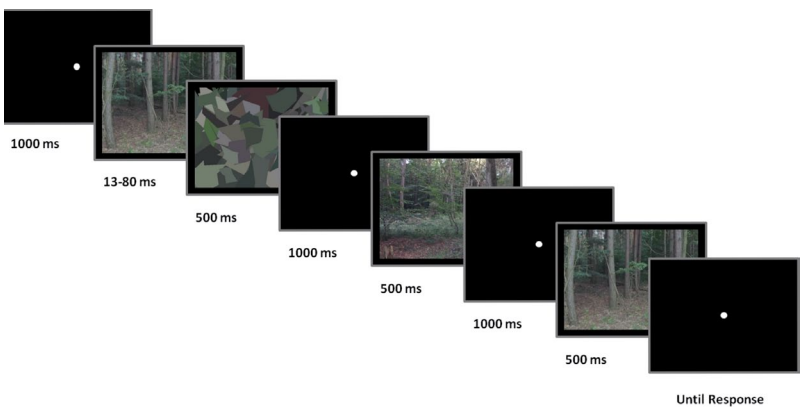


Figure 1. Trial event sequence in Experiment 1 (the left-eye views of the images are shown). Observers indicated which of the scenes in the last two intervals had been presented at the beginning of the trial.

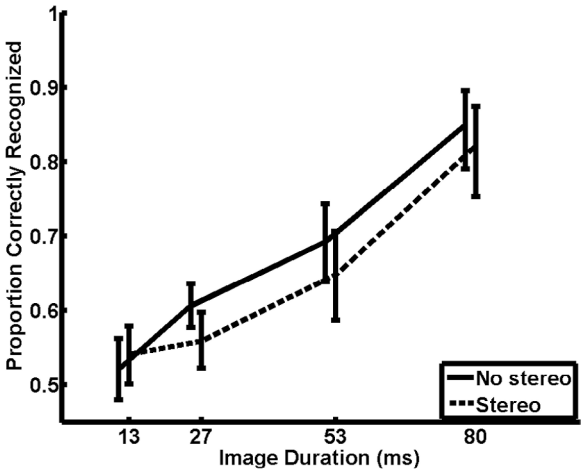


Figure 2. Recognition accuracy in Experiment 1 (forest scenes) as a function of image duration and modality of presentation. No stereo indicates binocular presentation without disparity. Error bars are between-subject 95% confidence intervals of the mean. The points have been displaced horizontally to increase visibility. The presence of binocular disparity did not enhance recognition performance.

2.2 Results

Accuracy in Experiment 1 (forest scenes) is depicted in Figure 2. Once again, performance increased as a function of exposure time from chance level at 13 ms (stereoscope setup) to quite good recognition at 80-ms exposure time. It is quite evident that stereo presentation does not constitute an advantage compared to nonstereo presentation. In order to confirm this finding, we performed a repeated-measure ANOVA with Image Duration and Presentation Modality as factors on the accuracy.

The effect of Image Duration was significant ($F(3, 36) = 42.700$, $p < 0.001$, $\eta_p^2 = 0.78$), whereas the effect of Presentation Modality ($F(1, 12) = 4.042$, $p = 0.067$, $\eta_p^2 = 0.25$) and the two-way interaction ($F(3, 36) = 1.453$, $p = 0.244$, $\eta_p^2 = 0.11$) were not significant.

2.3 Discussion

The results of Experiment 1 seem to indicate that binocular disparity does not contribute to the fast recognition of visual scenes, if anything, a nonsignificant trend for worse recognition of stereo pictures emerged, a result which is strikingly different from the observation of Liu and colleagues (2000), who found that stereo produced a small (2.5%) enhancement in the recognition rate of faces with a viewing time of 1.5 s and the observation of Valsecchi and Gegenfurtner (2012), who found a significant improvement in the long-term memorization rate when observers viewed the same forest pictures for 7 s.

For the sake of brevity, we only report here one experiment conducted using forest images, where, based on the finding by Valsecchi and Gegenfurtner (2012), we expected a stereo advantage for scene recognition. Notice that in multiple additional experiments we replicated the current findings with urban scenes, indoor scenes, and when the target and the distractor scenes were presented simultaneously.

In the following two experiments, we test two possible accounts for this finding. In Experiment 2, we test whether observers relied only on objects defined by contours and chromaticity in order to recognize the whole scene. In Experiment 3, we test whether the access to binocular disparity was limited by the fast presentation time.

3 Experiment 2: artificial scenes

The aim of Experiment 2 was to test whether impoverishing the monocular cues to scene identity could increase reliance on binocular disparity. For this purpose, we turned to artificial scenes, for which scene parameters could be better controlled. We created scenes exclusively populated by cubes suspended in space, while manipulating the presence of color on the surface of the cubes. In this context, once surface information is removed, the only aspect distinguishing scenes is their spatial layout, in principle maximizing the relevance of binocular disparity. Furthermore, we chose to simulate small scenes, entirely contained within the peri-personal space, where the relevance of disparity is supposed

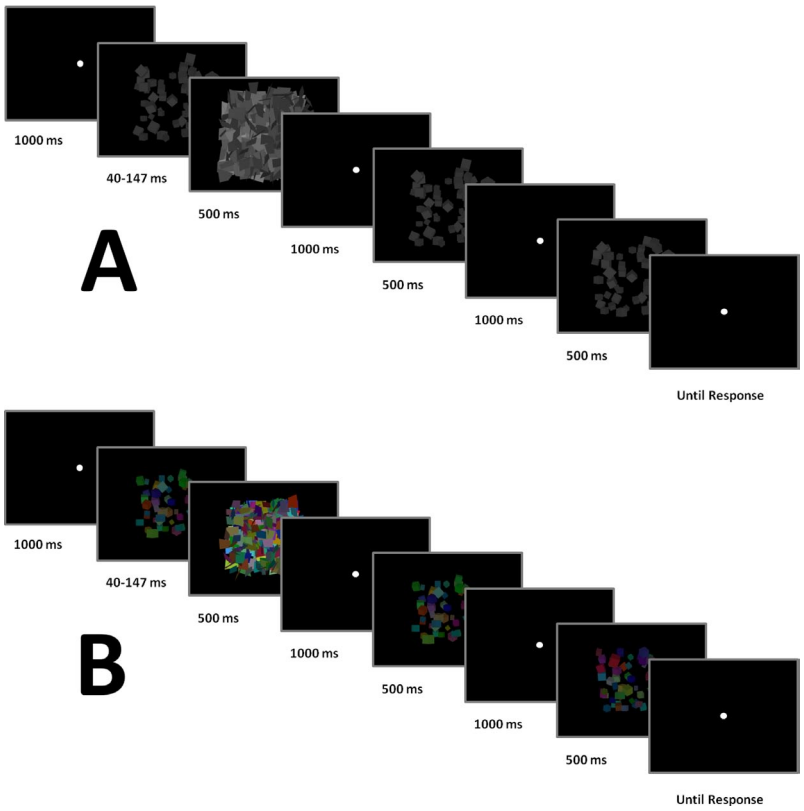


Figure 3. Trial event sequence in Experiment 2. Examples of grayscale (A) and colored (B) artificial scene pictures are shown. Observers indicated which of the two intervals contained the target scene which had been presented at the beginning of the trial. Cyclopean viewpoint versions of the images (as used for nonstereo presentation) are shown. Notice that the 2D layouts of the choice stimuli differ due to their perspective projection.

to be maximal (McKee et al., 1990; Sheedy et al., 1986). Finally, in order to increase the opportunity for our observers to adjust to the presence of disparity, we tested stereo and 2D performance in separate sessions. In this and the next experiments, we use rendered random arrangements of surfaces in depth as masks. These mimicked the properties of the scene images, including, besides disparity, residual monocular cues to 3D structure such as illumination.

3.1 Methods

3.1.1 Observers

Sixteen observers (13 females, mean age: 24.9 years) participated in Experiment 2 (artificial scenes). All observers were naive as to the aim of the study.

3.1.2 Stimuli

Stimuli were scenes containing 50 cubes rendered with OpenGL. The same stereoscope setup was used as in Experiment 1 (forest scenes), with a 55.5-cm viewing distance. In the simulated environment the side of each cube ranged randomly in length between 20.5 and 40.1 mm (2.1° and 4.1° of visual angle at fixation distance). Each cube was randomly rotated around its center. The center of one cube was always placed directly behind the fixation point (its center was displaced beyond the fixation point by half the size of the cube, so that participants would be fixating near the surface of the central cube during stimulus presentation). The centers of the cubes were distributed within a cube centered on the fixation point and whose side was 153.7 mm (15.7°), with the constraint that the minimum distance between the centers of two cubes should be at least 25.6 mm (2.64°). The centers of the cubes were not allowed to be nearer than 25.6 mm (2.64°) to the cyclopean line of sight in order to avoid the occlusion of the central cube.

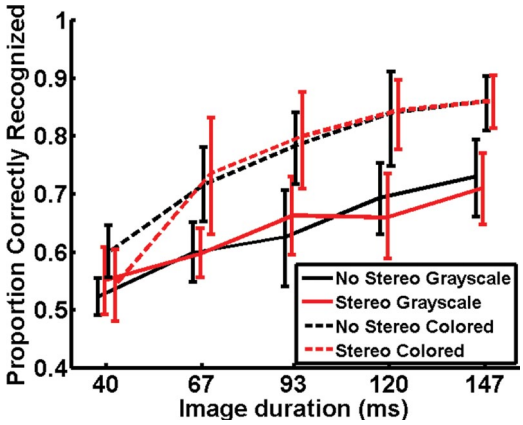


Figure 4. Proportion of correctly recognized pictures in Experiment 2 (artificial scenes) as a function of image duration. Stereo and No Stereo indicate binocular viewing with and without disparity, respectively. Error bars are between-subject 95% confidence intervals of the mean. The points have been displaced horizontally to increase visibility. Color greatly increased recognition performance but performance did not improve when binocular disparity was present.

For each scene, we created two versions by re-seeding only the *z*-coordinate of the cube centers (meaning the position along a line perpendicular to the display screen) and keeping the orientation of each cube constant, one to be used as target and the other to be used as the corresponding distractor. Moreover, for each scene we rendered three views, one from the left eye, one from the right eye and a cyclopean view (the latter to be used in 2D presentation).

We created 220 scenes with homogeneous gray cubes (as depicted in Figure 3) and 220 scenes where each cube was colored (the colors were randomly drawn from the whole RGB 32-bit color space). The colors of the cubes (except for the central one) differed between the two versions of each scene, and thus constituted a cue for the recognition of the target. The scenes were rendered with both ambient light and a point-light positioned 85.4 mm below the line of sight and 256.2 mm in front of the screen.

We also created two sets of mask images (grayscale and colored). The mask images were rendered scenes containing 1,500 irregular quadrilateral surfaces. The surfaces were located within a rectangular parallelepiped whose sides measured 174.2 mm (17.8°) on the *x*- and *y*-axes and 153.7 mm on the *z*-axis. The surfaces nearer to the observer had smaller size and a higher probability of being rotated from the fronto-parallel plane. This ensured that the mask completely occluded the space behind the configuration and allowed for surfaces with intermediate disparities to be visible from the observer's point of view. As we did with the scene images, we created left-eye, right-eye, and cyclopean views for masks.

3.1.3 Procedure

The trial event sequence is depicted in Figure 3. The test image was shown for a variable interval (40, 67, 93, 120, or 147 ms). Subsequently, the mask was shown for 500 ms. The two choice images were shown in two successive intervals in order to maximize display size while keeping size constant from target to choice presentation. Participants indicated the target interval with a key press. After the key press, if the answer was incorrect, the fixation point turned red for 1 s.

The images could be shown in two modalities (stereo or nonstereo). In each trial, the same modality of presentation was applied coherently to the target, mask and choice images. Each combination of Color and Presentation Modality was tested in a separate session including 220 trials. There were 44 trials for each combination of Color, Presentation Modality and Image Duration. Ten practice trials using extra scenes were run at the beginning of each session. Each session lasted approximately 45 min.

3.2 Results

Accuracy in Experiment 2 (artificial scenes) is depicted in Figure 4. Recognition performance was close to chance level with 40-ms exposure time and did not reach 90% even when the scenes were exposed for 147 ms. In most of the observations, removing the color information from the surface of

the cubes reduced the correct recognition rate by about half (discounting the 50% chance rate). Once again, there is no evident advantage for stereo viewing of the scene images over nonstereo viewing. A repeated-measure ANOVA with Image Duration (40, 67, 93, 120, or 147 ms), Presentation Modality (stereo vs. no stereo) and Color (vs. grayscale) as factors was performed. The main effect of Color ($F(1, 15) = 33.998$, $p < 0.001$, $\eta_p^2 = 0.693$), the main effect of Image Duration ($F(4, 60) = 48.732$, $p < 0.001$, $\eta_p^2 = 0.764$) and the Image Duration × Color interaction ($F(4, 60) = 9.296$, $p < 0.001$, $\eta_p^2 = 0.38$ were significant. All remaining effects and interactions were not significant (main effect of presentation modality: $F(1, 15) < 0.001$, $p = 0.989$, $\eta_p^2 < 0.001$. Presentation Modality × Color interaction: $F(1, 15) = 0.003$, $p = 0.959$, $\eta_p^2 < 0.001$. Presentation Modality × Image Duration interaction: $F(4, 60) = 0.859$, $p = 0.494$, $\eta_p^2 = 0.054$. Three-way interaction: $F(4, 60) = 1.567$, $p = 0.195$, $\eta_p^2 = 0.094$).

3.3 Discussion

Similar to what we observed when observers were faced with real-world scenes in Experiment 1 (forest scenes), and despite the limited availability of monocular cues, our observers still did not benefit from disparity while recognized artificial scenes in Experiment 2. Contrary to Experiment 1 (forest scenes), in this case the lack of advantage given by the addition of binocular disparity to the display cannot be due to the fact that the images depicted objects outside of the peri-personal space. This finding is all the more striking because the stereo vs. no-stereo factor was fixed within a given session in Experiment 2.

Our synthetic scenes were deprived of a number of monocular cues to scene identity, i.e. the presence of recognizable objects in the scene, their number and, depending on the condition, color. Evidently, the scene images still contained a number of monocular cues which the observers could use in order to recover the spatial arrangement of the objects, including occlusions and illumination and objects' size gradients. Moreover, due to perspective, the different position of the cubes along the depth axis produced different 2D projections between the target and distractor images. We think, however, that any display lacking these residual monocular cues would not qualify as a scene, so we feel confident in stating that binocular stereo does not produce appreciable improvements in the recognition of scenes when fast encoding is required.

As stated above, the second possible explanation for our results could be that the fast presentation prevented the extraction of disparity structures from the display. In order to tackle this question in the next experiment, we used dense RDS images, where disparity can be completely decoupled from any monocular cue (Julesz, 1971).

4 Experiment 3: RDS patterns

The rationale for Experiment 3 was to determine whether binocular disparity was available to observers in the "recognition after masked presentation" paradigm of Experiments 1 and 2. In other words, did observers fail to use disparity because it was not available to them, or because this signal was measured by the visual system but did not participate when they did the task? As a proxy for scene recognition, which does not allow for the decoupling of stereo from monocular cues, we used an RDS pattern-recognition task. As a benchmark, we also investigated the effect of masked presentation on the recognition of the corresponding luminance patterns of matched difficulty.

Within the same experiment, we also tested the performance on the recognition of patterns defined by both disparity and luminance. This approach has not yet been applied to the domain of binocular disparity, given that the only study directly dealing with the effect of stereo on scene recognition/object recognition (as opposed to spatial layout) lacked a stereo-only condition (Liu et al., 2000). However, researchers did investigate the combination of other visual dimensions in recognition tasks, such as spatial frequency (Olds & Engel, 1998), color and luminance (Syrkin & Gur, 1997), orientation, contrast, and spatial frequency (Thomas & Olzak, 1990). Furthermore, the question of how disparity and luminance contribute to pattern recognition is formally related to the question of how different sources of information contribute to the detection of threshold stimuli, and specifically of whether subthreshold summation takes place (e.g. Meinhardt, Persike, Mesenholl, & Hagemann, 2006), and to the question of how changes in individual dimensions contribute to the appearance of complex stimuli (Landy et al., 1995; To, Baddeley, Troscianko, & Tolhurst, 2011; To, Lovell, Troscianko, & Tolhurst, 2008). In all of these contexts, it is possible to compare the performance which is observed when combined stimuli are presented with the performance which can be predicted based on the results of the single-modality stimuli assuming independent processing of the two dimensions and linear combination of the responses. Following Dosher, Sperling, and Wurst

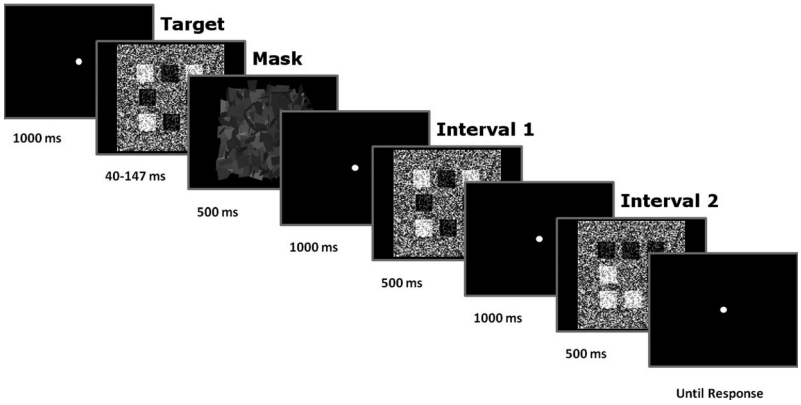


Figure 5. Trial event sequence in the Luminance condition of Experiments 3 (RDS patterns). Observers indicated which of the two intervals contained the target pattern which had been presented at the beginning of the trial.

(1986) and Backus (2009), we decided to implement a probit model in which the strength of each signal (Stereo and Luminance) is derived from the recognition accuracy for the stimuli in the single modality conditions. The situation in which both signals are present can be modeled assuming that each signal (cue) is analyzed by an independent Bernoulli Expert (Backus, 2009) that contributes a subjective reliability to the observer's overall belief about a binary property of the world. Within this context, the predicted combined strength of the two signals is simply given by the sum of the individual strengths.

4.1 Methods

4.1.1 Observers

Twelve observers (nine females, mean age: 25.5 years) participated in Experiment 3 (RDS patterns). One observer had already participated in Experiment 2 (artificial scenes). All observers were naive as to the aim of the study.

4.1.2 Stimuli

The stimuli were configurations of eight squares defined by disparity, luminance or both (Figure 5). Both the squares and the background were textured by a grayscale white noise pattern composed of large "display pixels" that ranged in luminance from 0 to 96 cd/m^2. The background had a size of 100 $\times$ 100 display pixels and had zero disparity. Each of the eight squares had a size of 16 $\times$ 16 display pixels (area 256 display pixels) and subtended 1.96°. The centers of the squares on the cardinal axes were located 2.81° from fixation.

In each configuration, when the pattern was defined in luminance and/or binocular disparity, the contrast polarity of the squares and the sign of the disparity of the squares were evenly distributed, i.e. four squares were darker than the background and four were brighter, four squares had crossed disparity and four had uncrossed disparity, i.e. they appeared in front of and behind the background, respectively. When present, the disparity was 7.3 arcmin. The visibility of the configurations was manipulated by varying the contrast or the disparity coherence (from 0 to 256 pixels) of the squares. The same mask images as in the Grayscale condition of Experiment 2 (artificial scenes) were used.

4.1.3 Procedure

The trial event sequence is depicted in Figure 5. The test image was shown for a variable interval (40, 67, 93, 120, or 147 ms). Subsequently, the mask was shown for 500 ms. The two choice images were shown in two successive intervals. The target image and the distractor image differed only in the polarity of four out of the eight squares; the reason for this was to push observers to process the configurations in a global fashion. Participants indicated the target interval with a key press. Between trials, a 1-s interval where the fixation point turned red was inserted if the answer was incorrect, otherwise the next trial started immediately after key press.

The experiment was composed of four sessions. The first was a pre-test session whose aim was to titrate the visibility of the stimuli for the subsequent three experimental sessions. The pre-test ses-

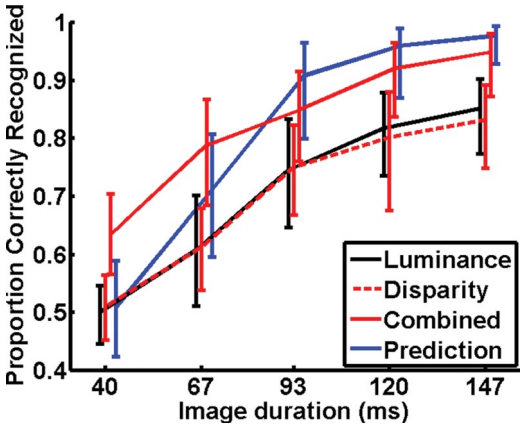


Figure 6. Observed and predicted proportion of correctly recognized trials in Experiment 3 (RDS patterns). Error bars are between-subject 95% confidence intervals of the mean. The points have been displaced horizontally to increase visibility. Recognition performance was equally affected by brief image presentation for binocular disparity and luminance patterns. Performance increased in combined trials to the level predicted by the strength summation model for longer image durations and beyond the predicted level for the shorter image durations.

sion was preceded by a training session of 35 trials in which the image exposure duration decreased linearly from 2 s to 120 ms. During the training and pre-test sessions only single-modality trials were presented, i.e. the stimuli were defined either by disparity or by luminance, and the modalities were randomly interleaved. The luminance contrast (i.e. the difference between the average luminance of each square and the average luminance of the background) and the disparity coherence (i.e. the proportion of pixels within each square displaced between the left and right eye images) were chosen randomly from a wide range of values in the training trials. In the subsequent pre-test session their values were chosen in order to target the 75% visibility range for each participant and stimulus modality. The pre-test session consisted of 250 trials in which the stimulus exposure time was fixed at 120 ms. In the first 30 trials of each modality and in 25% of the subsequent trials, the luminance contrast and disparity coherence values were drawn from two Gaussian distributions ($\mu = 0$ cd/m^2, $\sigma = 37.2$ cd/m^2 for luminance difference; $\mu = 0\%$, $\sigma = 39\%$ for disparity coherence) derived from pilot data, trimming values below 0 for both modalities and values above 100% for coherence. After the 30th trial, in 75% of the cases the participants' responses from the current modality were fitted with a cumulative Gaussian model using the psignifit toolbox version 2.5.41 for Matlab (see http://bootstrap-software.org/psignifit/), which implements the maximum likelihood method described by Wichmann and Hill (2001). The value for the subsequent trial of the same modality was chosen from the Gaussian distribution with the parameter values obtained from the fit (trimming σ values below 1.42 cd/m^2 and 1.9%), so as to adapt to the observer's performance. The same fitting procedure was used at the end of the training session in order to calculate the 75% accuracy thresholds for each participant and modality, which were then used for the subsequent sessions. In the remaining 25% of the trials, the stimulus values were drawn from the original broad distributions so as to allow further exploration of the stimulus space. Three participants failed to demonstrate 75% accuracy in at least one modality and did not continue with the experiment. The first session lasted approximately 75 min.

Each of the three subsequent experimental sessions consisted of 270 trials and lasted approximately 75 min. Three modality conditions were tested: Disparity, Luminance, and Combined. In the Disparity and Luminance conditions, the stimuli were defined in only one modality, as in the pre-test session. In the Combined condition, both Disparity and Luminance defined the difference between the target and the distractor. The polarity of the squares in the two modalities was matched coherently through each session. Pooling all three sessions, observers were exposed to 54 trials in each combination of Modality and Image Duration.

4.2 Results

Recognition accuracy is reported in Figure 6. The rate of correct recognition increases with a remarkably similar slope as a function of exposure time for both stereo- and luminance-defined patterns. In both cases, performance is at chance level with 40-ms exposure and increases to around 80% correct

recognition if the image is presented for 147 ms. Combining the two dimensions increases the correct recognition rate by around 15%, independent of exposure time.

Accuracy results were analyzed with a two-way repeated-measure ANOVA with Image Duration (40, 67, 93, 120, or 147 ms) and Modality (Luminance, Disparity, or Combined) as factors. This yielded a significant effect of both factors (Image Duration: $F(4, 32) = 40.913$, $p < 0.001$, $\eta_p^2 = 0.84$; Modality: $F(2, 16) = 19.606$, $p < 0.001$, $\eta_p^2 = 0.71$) and no significant interaction ($F(8, 64) = 0.985$, $p = 0.456$, $\eta_p^2 = 0.071$). Planned comparisons indicate that the Combined trials yielded higher performance than either Luminance ($F(1, 8) = 18.594$, $p < 0.003$, $\eta_p^2 = 0.70$) or Disparity ($F(1, 8) = 46.097$, $p < 0.001$, $\eta_p^2 = 0.85$) trials.

This result indicates that the information from both modalities is pooled when participants recognize patterns which differ in both luminance and disparity. Nonetheless, one can still ask whether the performance in combined trials is higher or lower than the prediction from linear summation. One convenient way of handling 2AFC accuracies in order to compute predictions is by converting them to signal strength (see Backus, 2009; Dosher et al., 1986; Meinhardt et al., 2006):

$$S_i = \Phi^{-1}(A_i), \tag{1}$$

where A_i is the accuracy in trials where a given modality is presented in isolation and Φ^{-1} is the inverse of the standard normal cumulative probability density function. Similarly, Equation (2) can be used to convert strength scores to accuracy:[1]

$$A_i = \Phi(S_i). \tag{2}$$

Equation (1) maps 100% correct performance to infinite strength. This inflates the variability of strength estimates when accuracy values are high. We thus postulated an error rate of $1/2n$, where n is the number of trials, in all design cells where all answers were correct.

When both single-modality recognition rates are above chance, the predicted sensitivity for the combined trials, assuming linear summation is simply given by

$$S_C = S_D + S_L, \tag{3}$$

where S_D is the strength of the disparity signal and S_L is the strength of the Luminance signal. We perform our calculations assuming that the two modalities contribute the same strength to the recognizability of the combined stimulus as they did when tested in isolation.

The prediction is plotted along with the results in Figure 6. Notice that our model handles negative sensitivities correctly and is unbiased. Indeed, at a 40-ms exposure, the prediction overlaps with the two single-modality performances which are recognized at chance level. Models which cannot handle negative sensitivities such as the one used by Meinhardt and colleagues (2006) would produce a positively biased prediction in this situation. We performed a repeated-measure ANOVA with Prediction (Observed vs. Predicted) and Image Duration (40, 67, 93, 120, or 147 ms) as factors comparing the observed performance with the one predicted by the cue strength summation model. The difference between predicted and observed performance was not significant ($F(1, 8) = 0.929$, $p = 0.363$, $\eta_p^2 = 0.10$), but the interaction between Prediction and Image Duration was ($F(4, 32) = 7.455$, $p < 0.001$, $\eta_p^2 = 0.48$). Obviously, the analysis also yielded a significant main effect of Image Duration ($F(4, 32) = 40.913$, $p < 0.001$, $\eta_p^2 = 0.84$). The interaction seems to be driven by the fact that the observed performance was higher than the predicted one at short image durations, although post-hoc paired t-tests with Bonferroni correction failed to show a significant difference between observed and predicted performance at any image duration (40 ms: $t(8) = 2.740$, $p = 0.127$; 67 ms: $t(8) = 2.797$, $p = 0.116$; 93 ms: $t(8) = 1.889$, $p = 0.477$; 120 ms: $t(8) = 1.849$, $p = 0.507$; 147 ms: $t(8) = 2.334$, $p = 0.239$).

[1] Notice that the strength formulas differ from the d′ calculation in 2AFC tasks (as used in Meinhardt et al., 2006) only by the $\sqrt{2}$ factor. The d′ and strength values are thus linearly related and our prediction model could be applied to either score yielding the same expected accuracies.

Furthermore, the linear summation model we apply corresponds to a cue summation with a Minkowski exponent equal to 1 (also see Macmillan & Creelman, 2005; e.g. To et al., 2011). Other models such as Euclidean summation (Minkowski exponent equal to 2) or MAX rule summation (infinite Minkowski exponent) predict a lesser increase of performance when both stimuli are presented at the same time. Our model thus constitutes a conservative test for superadditivity.

4.3 Discussion

A few results emerged from Experiment 3 (RDS patterns). First, the fact that the recognition of disparity and luminance patterns was affected in a remarkably similar fashion by our masking procedure indicates that the lack of any performance improvement in the presence of binocular disparity with real-world and artificial scenes in Experiments 1 and 2 was not due to a failure in ability to measure disparity, such as might have been caused by general "slowness" of stereo. Indeed, partially degraded stereo patterns could be recognized with above chance performance at 67-ms exposure, and with an accuracy superior to 75% at 93-ms exposure, whereas in Experiment 2 we failed to observe any improvement of artificial scene recognition even with 147-ms presentation. Second, observers could recognize combined patterns better than they could recognize either single-modality pattern.

The comparison between the observers' responses in combined trials and the predictions by our probability summation model showed superadditivity at short presentation times and subadditivity at longer presentation times. One possible interpretation of the superadditivity could be that limited changes in luminance within a patch can enhance its segmentation and its binocular fusion.

5 General discussion

In three experiments, we investigated how human observers make use of binocular disparity when they had to recognize briefly presented images of natural scenes (Experiment 1), images of artificial scenes (Experiment 2), and patterns within random-dot stereograms (Experiment 3).

The results of Experiments 1 and 2 indicate that binocular disparity does not contribute in any significant way to the recognition of briefly presented scenes, even when monocular cues are artificially reduced and recognition is largely impaired. The results of Experiments 3 show that, in principle, if observers are forced to process binocular disparity by the use of displays completely deprived of any monocular cues, they can recognize patterns defined by disparity presented briefly.

Overall, we suggest that human observers strategically ignore the information provided by binocular disparity as soon as monocular cues provide enough information to support even a poor recognition of rapidly encoded scenes. This is true regardless of whether the scenes represent objects outside of or within the peri-personal space and even when the amount of monocular cues is reduced to a minimum consistent with still being a "scene."

We believe that two factors contribute to the primacy of luminance in scene recognition tasks with brief presentation. First, an encoding capacity limit is consistent with this result. When faced with relatively simple patterns in Experiment 3 (RDS patterns), our observers combined the information coming from luminance and disparity, but when faced with the complex patterns typical of scenes, such as in Experiments 1 (forest scenes) and 2 (artificial scenes), they did not. Using disparity to recognize a complex scene may require cognitive resources (attention and/or memory) that were allocated instead to the processing of luminance and color. Notice that the fact that Valsecchi and Gegenfurtner (2012) found improved recognition from long-term memory for stereo pictures of the same forest scenes rules out the hypothesis that a general storage capacity limits the using of disparity in the encoding phase. The factor limiting the contribution of binocular disparity to scene recognition must be related to the rate at which visual information can be processed before encoding is terminated by the presentation of the mask.

The results from Experiments 1 and 2, using real-world and artificial scenes, have implications for our understanding of the scene recognition process itself. The first straightforward conclusion is that even if human observers use binocular disparity to segment objects in briefly presented scenes, the results from that process are not used for scene recognition. Stereo differs from chromaticity in this regard. Gegenfurtner and Rieger (2000) found enhanced fast encoding for colored natural images and attributed this advantage to the use of color-defined edges, which occur frequently in natural scenes (Hansen & Gegenfurtner, 2009), for segmentation. Within this framework, the present results indicate that, despite our previous report that binocular disparity can speed the detection of isolated targets against a zero-disparity background (Caziot et al., 2011), disparity-based segmentation does not contribute when the visual system must recognize a briefly presented complex visual scene. Consider again our forest scene experiment. Forest scenes lack the sharp luminance boundaries that characterize man-made objects and our artificial scene objects. Any factor enhancing the segmentation of the single elements should have an immediate impact on performance in this case, but stereo did not have that effect.

Second, our results indicate that human observers do not base the recognition of briefly presented scenes on the binocular disparity associated with elements or surfaces. The presence of salient objects

marked by a strong disparity discontinuity, such as a near tree in front of further-off trees, should be able to support the recognition of the target. Although the encoding of salient objects could contribute to the advantage of color in previous work, such as that of Gegenfurtner and Rieger (2000), we found no evidence that stereo labels objects as salient for the purpose of recognizing whole scenes.

Third, given that binocular disparity can enhance the perception of the depth arrangement of objects in scenes (McKee & Taylor, 2010), our results suggest that humans do not encode a detailed model of the spatial structure of rapidly presented scenes for their short-term recognition. Rapid short-term scene recognition has not been investigated extensively, since most of the studies on rapid scene perception have dealt with scene categorization (e.g. Greene & Oliva, 2009) or with object/animal detection within scenes (e.g. Drewes, Trommershauser, & Gegenfurtner, 2011; Rousselet, Fabre-Thorpe, & Thorpe, 2002). For these tasks, observers use low-level properties of the image (e.g. Crouzet & Serre, 2011) not a detailed representation of the 3D structure. It is thus quite possible that the representation used for the fast encoding of images in our experiments does not go beyond their 2D properties.

Our finding that the immediate recognition of briefly presented scenes is not enhanced by binocular disparity is in contrast with the finding by Valsecchi and Gegenfurtner (2012) that the same stereo pictures could be better recognized from long-term memory. Evidently, the tasks induced two qualitatively different encoding strategies. In the current study, where performance was largely determined by the encoding speed, observers relied on the straightforward matching of the 2D layout of the image, or possibly of a portion of it. Conversely, in the study by Valsecchi and Gegenfurtner (2012), where the main factor limiting long-term memory for scenes was the interference from other memorized items, observers did rely on binocular disparity in order to maximize performance. Anyway, it was quite evident from the result of Valsecchi and Gegenfurtner (2012) that reliance on binocular stereo was limited to the case where scenes had to be encoded visually. Binocular stereo produced no advantage when scenes could be encoded semantically, for instance based on the artifacts they contained.

Our findings with RDS displays are broadly consistent with studies in which binocular disparity was processed for presentation times as fast as 150 ms. Our observers recognized patterns defined by disparity (7.3 arcmin crossed or uncrossed) with 120-ms masked exposure. This is compatible in particular with the better-than-chance discrimination of 6-arcmin disparity between adjacent segments reported by Foley and Tyler (1976), and with the lack of any considerable increase in stereo thresholds with presentation times as short as 100 ms reported by Harwerth and colleagues (2003). Ritter (1980) found that both simple and relative disparity judgments were above chance with 120-ms masked presentations. Our findings extend the previous results by showing that even masks containing similar disparity as the target do not completely disrupt the extraction of stereo patterns. Our experimental question originated in the study of natural images. Object boundaries in natural images are marked both by luminance and stereo, so we used superimposed stereo and luminance patterns in our experiments. Future research could address how human observers combine binocular disparity and luminance when the patterns are not co-localized, a manipulation that influences how signals from different visual dimensions are combined (e.g. Krummenacher, Muller, & Heller, 2002).

We found that observers do not take into account stereo when they recognize briefly presented pictures of scenes if they have to choose between an identical presentation of the same picture and a distractor. This finding does not necessarily extend to other tasks. The RDS recognition task results show that stereo can in fact be processed. We chose RDS recognition when we tested stereo reliance in a different task because it cannot be performed monocularly. Other tasks might also benefit from stereo. For example, viewpoint-independent recognition of objects is facilitated by binocular stereo (Lee & Saunders, 2011). It is possible that this facilitation intervenes also in the case of the more complex pictures depicting scenes, even if little time is available to encode them.

In conclusion, even though binocular disparity is an omnipresent quality of our visual experience, and even though it can be extracted quickly from brief visual displays, human observers do not make use of it to rapidly encode the various parts of a scene, possibly because it requires additional resources to do so. We suggest that the recognition of briefly presented scene images is largely mediated by the comparison of their 2D structure.

References

Adams, W. J., & Mamassian, P. (2004). Bayesian combination of ambiguous shape cues. *Journal of Vision, 4*, 921–929. doi:10.1167/4.10.7

Adelson, E. H., & Bergen, J. R. (1991). The plenoptic function and the elements of early vision. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 3–20). Cambridge, MA: MIT Press.

Allison, R. S., Gillam, B. J., & Vecellio, E. (2009). Binocular depth discrimination and estimation beyond interaction space. *Journal of Vision, 9*(1), 10.1–14. doi:10.1167/9.1.10

Backus, B. T. (2009). The mixture of Bernoulli Experts: A theory to quantify reliance on cues in dichotomous perceptual decisions. *Journal of Vision, 9*(1), 6.1–19. doi:10.1167/9.1.6

Backus, B. T., & Banks, M. S. (1999). Estimator reliability and distance scaling in stereoscopic slant perception. *Perception, 28*, 217–242. doi:10.1068/p2753

Backus, B. T., Banks, M. S., van Ee, R., Crowell, J. A., & Crowell, D. (1999). Horizontal and vertical disparity, eye position, and stereoscopic slant perception. *Vision Research, 39*, 1143–1170. doi:10.1016/S0042-6989(98)00139-4

Banks, M. S., & Backus, B. T. (1998). Extra-retinal and perspective cues cause the small range of the induced effect. *Vision Research, 38*, 187–194. doi:10.1016/S0042-6989(97)00179-X

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436. doi:10.1163/156856897X00357

Burke, D. (2005). Combining disparate views of objects: Viewpoint costs are reduced by stereopsis. *Visual Cognition, 12*, 705–719. doi:10.1080/13506280444000463

Burke, D., Taubert, J., & Higman, T. (2007). Are face representations viewpoint dependent? A stereo advantage for generalising across different views of faces. *Vision Research, 47*, 2164–2169. doi:10.1016/j.visres.2007.04.018

Butler, T. W., & Westheimer, G. (1978). Interference with stereoscopic acuity: Spatial, temporal, and disparity tuning. *Vision Research, 18*, 1387–1392. doi:10.1016/0042-6989(78)90231-6

Caziot, B., Valsecchi, M., Gegenfurtner, K. R., & Backus, B. T. (2011). Role of binocular vision during image encoding. *Perception, 40 (ECVP Abstract Supplement)*, 145. doi:10.1068/v110529

Crouzet, S. M., & Serre, T. (2011). What are the visual features underlying rapid object recognition? *Frontiers in Psychology, 2*, 326. doi:10.3389/fpsyg.2011.00326

Doorschot, P. C. A., Kappers, A. M. L., & Koenderink, J. J. (2001). The combined influence of binocular disparity and shading on pictorial shape. *Perception & Psychophysics, 63*, 1038–1047. doi:10.3758/BF03194522

Dosher, B. A., Sperling, G., & Wurst, S. A. (1986). Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3d structure. *Vision Research, 26*, 973–990. doi:10.1016/0042-6989(86)90154-9

Drewes, J., Trommershauser, J., & Gegenfurtner, K. R. (2011). Parallel visual search and rapid animal detection in natural scenes. *Journal of Vision, 11,* art. 20. doi:10.1167/11.2.20

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research, 32*, 2385–2400. doi:10.1016/0042-6989(92)90102-O

Foley, J. M., & Tyler, C. W. (1976). Effect of stimulus-duration on stereo and vernier displacement thresholds. *Perception & Psychophysics, 20*, 125–128. doi:10.3758/BF03199443

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology, 10*, 805–808. doi:10.1016/S0960-9822(00)00563-7

Girshick, A. R., & Banks, M. S. (2009). Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. *Journal of Vision, 9,* 8.1–20. doi:10.1167/9.9.8

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58*, 137–176. doi:10.1016/j.cogpsych.2008.06.001

Hansen, T., & Gegenfurtner, K. F. (2009). Independence of color and luminance edges in natural scenes. *Visual Neuroscience, 26*, 35–49. doi:10.1017/S0952523808080796

Harwerth, R. S., Fredenburg, P. M., & Smith, E. L. (2003). Temporal integration for stereoscopic vision. *Vision Research, 43*, 505–517. doi:10.1016/S0042-6989(02)00653-3

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298*, 1627–1630. doi:10.1126/science.1075396

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision, 4*, 967–992. doi:10.1167/4.12.1

Ichikawa, M., Saida, S., Osa, A., & Munechika, K. (2003). Integration of binocular disparity and monocular cues at near threshold level. *Vision Research, 43*, 2439–2449. doi:10.1016/S0042-6989(03)00432-2

Julesz, B. (1971). *Foundations of cyclopean perception*. Chicago, IL: University of Chicago Press.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research, 43*, 2539–2558. doi:10.1016/S0042-6989(03)00458-9

Krummenacher, J., Muller, H. J., & Heller, D. (2002). Visual search for dimensionally redundant pop-out targets: Parallel-coactive processing of dimensions is location specific. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 1303–1322.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination - in defense of weak fusion. *Vision Research, 35*, 389–412. doi:10.1167/12.4.6

Lee, Y. L., & Saunders, J. A. (2011). Stereo improves 3D shape discrimination even when rich monocular shape cues are available. *Journal of Vision, 11*, art. 6. doi:10.1167/11.9.6

Lehmkuhle, S., & Fox, R. (1980). Effect of depth separation on metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 605–621.

Liu, C. H., Collin, C. A., & Chaudhuri, A. (2000). Does face recognition rely on encoding of 3-D surface? Examining the role of shape-from-shading and shape-from-stereo. *Perception, 29*, 729–743. doi:10.1068/p3065

Long, N., & Over, R. (1974). Stereospatial masking and aftereffect with normal and transformed random-dot patterns. *Perception & Psychophysics, 15*, 243–248. doi:10.3758/BF03213940

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

McKee, S. P., Levi, D. M., & Bowne, S. F. (1990). The imprecision of stereopsis. *Vision Research, 30*, 1763–1779. doi:10.1016/0042-6989(90)90158-H

McKee, S. P., & Taylor, D. G. (2010). The precision of binocular and monocular depth judgments in natural settings. *Journal of Vision, 10*, art. 5. doi:10.1167/10.10.5

McKee, S. P., Watamaniuk, S. N. J., Harris, J. M., Smallman, H. S., & Taylor, D. G. (1997). Is stereopsis effective in breaking camouflage for moving targets? *Vision Research, 37*, 2047–2055. doi:10.1016/S0042-6989(96)00330-6

Meese, T. S., & Holmes, D. J. (2004). Performance data indicate summation for pictorial depth-cues in slanted surfaces. *Spatial Vision, 17*, 127–151. doi:10.1163/156856804322778305

Meinhardt, G., Persike, M., Mesenholl, B., & Hagemann, C. (2006). Cue combination in a combined feature contrast detection and figure identification task. *Vision Research, 46*, 3977–3993. doi:10.1016/j.visres.2006.07.009

Olds, E. S., & Engel, S. A. (1998). Linearity across spatial frequency in object recognition. *Vision Research, 38*, 2109–2118. doi:10.1016/S0042-6989(97)00393-3

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442. doi:10.1163/156856897x00366

Ritter, M. (1980). Perception of depth: Different processing times for simple and relative positional disparity. *Psychological Research (Psychologische Forschung), 41*, 285–295. doi:10.1007/BF00308874

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*, 629–630. doi:10.1038/nn866

Sheedy, J. E., Bailey, I. L., Buri, M., & Bass, E. (1986). Binocular Vs monocular task-performance. *American Journal of Optometry and Physiological Optics, 63*, 839–846.

Syrkin, G., & Gur, M. (1997). Colour and luminance interact to improve pattern recognition. *Perception, 26*, 127–140.

Thomas, J. P., & Olzak, L. A. (1990). Cue summation in spatial discriminations. *Vision Research, 30*, 1865–1875.

To, M. P. S., Baddeley, R. J., Troscianko, T., & Tolhurst, D. J. (2011). A general rule for sensory cue summation: evidence from photographic, musical, phonetic and cross-modal stimuli. *Proceedings of the Royal Society B-Biological Sciences, 278*, 1365–1372. doi:10.1098/rspb.2010.1888

To, M. P. S., Lovell, P. G., Troscianko, T., & Tolhurst, D. J. (2008). Summation of perceptual cues in natural visual scenes. *Proceedings of the Royal Society B-Biological Sciences, 275*, 2299–2308. doi:10.1098/rspb.2008.0692

Tyler, C. W., & Kontsevich, L. L. (2005). The structure of stereoscopic masking: Position, disparity, and size tuning. *Vision Research, 45*, 3096–3108. doi:10.1016/j.visres.2005.07.034

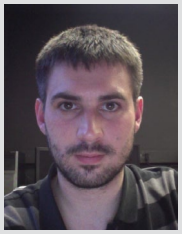
Valsecchi, M., & Gegenfurtner, K. R. (2012). On the contribution of binocular disparity to the long-term memory for natural scenes. *Plos One, 7,* e49947. doi:10.1371/journal.pone.0049947

Wardle, S. G., Cass, J., Brooks, K. R., & Alais, D. (2010). Breaking camouflage: Binocular disparity reduces contrast masking in natural images. *Journal of Vision, 10*. doi:10.1167/10.14.38

Westheimer, G. (2011). Three-dimensional displays and stereo vision. *Proceedings of the Royal Society B-Biological Sciences, 278*, 2241–2248. doi:10.1111/j.1475-1313.1993.tb00419.x

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63*, 1293–1313.

Matteo Valsecchi studied Psychology at the Vita-Salute San Raffaele University in Milan (Italy). He subsequently got a PhD in Cognitive Sciences and Education from the University of Trento (Italy). He is currently a Humboldt postdoctoral fellow at the Department of General Psychology of the Justus-Liebig University of Giessen (Germany).



Baptiste Caziot studied psychology at the Université Paris Descartes and electronics at the CNED. He then received a research master's degree in cognitive science jointly delivered by the ENS, EHESS, and Université Paris Descartes. He is now a PhD student at the Graduate Center for Vision Research, SUNY College of Optometry.

Benjamin T. Backus studied mathematics at Swarthmore College (BA), vision science at UC Berkeley (PhD), and neuroscience at Stanford University (postdoctoral). He is currently Empire Innovation Associate Professor in the Graduate Center for Vision Research at SUNY College of Optometry. His interests include binocular vision and stereopsis, perceptual learning, neural plasticity, amblyopia, and strabismus.

Karl R. Gegenfurtner studied psychology in Regensburg (Germany) and then did a PhD in Experimental Psychology at New York University. After spending time as a PostDoc in New York and Tübingen, he became Professor of Psychology in Magdeburg. Since 2001, he has been at Giessen University (http://www.allpsych.uni-giessen.de/karl/).