

1 **Explicit and implicit depth-cue integration: evidence of systematic biases with real objects**

2

3 **Authors**

4 Carlo Campagnoli¹, Bethany Hung², and Fulvio Domini³

5

6 **Affiliation**

7 ¹Department of Psychology, Princeton University, Princeton, NJ, USA; ² The Johns Hopkins University School of
8 Medicine, Baltimore, MD, USA; ³Department of Cognitive, Linguistic and Psychological Science, Brown University,
9 Providence, RI, USA

10

11 **Running head**

12 Depth-cue integration with real objects

13

14 **Corresponding author**

15 Carlo Campagnoli

16 Email: carlo.campagnoli@princeton.edu

17 Address: Department of Psychology, Princeton University, Peretsman-Scully Hall, South Drive, Princeton, NJ, USA

18

19

20

21

22

23

24

25

26

27

28

29

1 **Abstract**

2 In a previous series of experiments using virtual stimuli, we found evidence that 3D shape estimation agrees to a
3 superadditivity rule of depth-cue combination. According to this rule, adding depth cues leads to greater perceived depth
4 magnitudes and, in principle, to depth overestimation. The mechanism underlying the superadditivity effect can be fully
5 accounted for by a normative theory of cue integration, through the adaptation of a model of cue integration termed the
6 Intrinsic Constraint (IC) model. As for its nature, it remains unclear whether superadditivity is a byproduct of the artificial
7 nature of virtual environments, causing explicit reasoning to infiltrate behavior and inflate the depth judgments when a
8 scene is richer in depth cues, or the genuine output of the process of depth-cue integration. In the present study, we addressed
9 this question by testing whether the IC model's prediction of superadditivity generalizes beyond VR environments to real
10 world situations. We asked participants to judge the perceived 3D shape of cardboard prisms through a matching task. To
11 assay the potential influence of explicit control over those perceptual estimates, we also asked participants to reach and hold
12 the same objects with their fingertips and we analyzed the in-flight grip size during the reaching. Using physical objects
13 ensured that all visual information was fully consistent with the stimuli's 3D structure without computer-generated artifacts.
14 We designed a novel technique to carefully control binocular and monocular 3D cues independently from one another,
15 allowing to add or remove depth information from the scene seamlessly. Even with real objects, participants exhibited a
16 clear superadditivity effect in both explicit and implicit tasks. Furthermore, the magnitude of this effect was accurately
17 predicted by the IC model. These results confirm that superadditivity is an inherent feature of depth estimation.

18

19 Keywords: grasping; 3D vision; depth perception; virtual reality.

20

21

22

23

24

25

26

27

28

29

30

1 Introduction

2

3 In previous studies we showed that, when sources of depth information (cues) are added to the virtual rendering of a
4 3D shape, the perceived magnitude of a given 3D property of that shape (e.g., its depth, slant or curvature) increases. For
5 instance, when viewing a cylindrical surface from a close distance (50 cm), its overall depth is consistently underestimated
6 when the shape is only specified by binocular disparity, compared to when specified by a mixture of disparity signals and
7 texture gradient (Campagnoli & Domini, 2019).

8 This pervasive phenomenon, which we refer to here as superadditivity effect (the estimated depth from a combination
9 of depth cues is greater than the depth estimates from the individual cues) has been previously observed in the context of a
10 series of perceptual studies investigating a model of depth cue combination termed Intrinsic Constraint (IC) model (Di Luca
11 et al., 2010; Domini & Caudek, 2003, 2009, 2010, 2011, 2013; Domini et al., 2006; Domini et al., 2011; Kemp et al., 2018).
12 The IC model can be adapted so to successfully predict these kinds of shifts in 3D shape estimation, by pooling the retinal
13 depth cues through a combination rule that prioritizes maximizing sensitivity to depth information over accuracy (Domini
14 & Vishwanath, 2020).

15 While the model provides a description of the mechanism behind the behavioral effect, it is unclear whether the
16 presence of superadditivity in the first place is the symptom of an explicit overcompensation introduced by the observers,
17 or the signature of a characteristic feature of depth estimation. On the one hand, when witnessing more depth signals in the
18 visual field, superadditivity may be the result of participants trying to qualitatively express the greater richness of their
19 visual experience by consciously inflating their judgments. If, on the other hand, superadditivity relates directly to the
20 readout of the internal depth estimates, we should expect it to equally affect non-explicit tasks such as reaching towards a
21 3D object with the hand to pick it up. Converging evidence in support of the latter possibility has been documented in
22 previous VR investigations (Foster et al., 2011; Campagnoli & Domini, 2016, 2019). However, the artificial setups used in
23 those studies may have introduced yet another potential source of explicit bias in the responses, because participants
24 immersed in a virtual scene tend to supervise their behavior and exert top-down control over their own actions (Chessa et
25 al., 2019).

26 To investigate the nature of the phenomenon, we sought to test the predictions of the IC model for the superadditivity
27 effect in the context of a real-world scene, by asking participants to judge as well as to interact with physical objects. To
28 have the same visual conditions of the previous VR experiments, we designed a novel technique to manipulate the amount
29 of disparity, motion and focus cues specifying the shape of a cardboard-made prism, and we examined how participants
30 judged the stimuli's shape under the various visual conditions. In a within-subjects design, we compared explicit reports of
31 perceived depth with an implicit measure of depth provided by the in-flight grip aperture when participants reached with
32 the hand towards the same object to hold it. As a generalization test of the predictions, the stimuli were purposely different
33 in shape from those used before. The results were qualitatively identical to what previously obtained in VR, supporting the

1 hypothesis that the visuomotor system integrates depth cues following a combination rule that is not triggered by computer-
2 generated environments, but describes a more general principle of how 3D shapes are represented internally by the brain.
3

4 *The mechanism of superadditivity predicted by the IC theory of cue integration*

5 A traditional assumption in modeling cue integration is that separate visual modules analyze independent sources of
6 information to derive magnitudes of distal 3D properties (e.g., the overall depth of an object, the slant of a plane, the
7 curvedness of a smooth surface) in an accurate fashion. Although the assumption of veridicality is lawful both theoretically
8 and computationally, it is often challenged by empirical findings showing that 3D shape estimation is affected by systematic
9 distortions under a variety of visual conditions (Glennerster et al., 1996; Hecht et al., 1999; Tittle et al., 1995; Tassinari et
10 al., 2008; Todd & Norman, 2003). One way to reconcile this evidence is through the post-hoc introduction of additional
11 free parameters to the model – for example prior terms accounting for specific biases in the data – with the downside that
12 they may potentially lead to overfitting. Another way is by seeking a combination rule that aims to maximize something
13 other than accuracy, without increasing its complexity.

14 The IC theory of cue integration follows the latter approach and posits that the goal of the visual system is to maximize
15 the sensitivity to the available 3D information, while minimizing the undesired influence of other scene parameters.
16 According to this framework, independent visual modules provide 3D estimates that are only linearly related to the
17 magnitude of physical 3D properties, not veridical to begin with. Veridicality is still allowed under ideal conditions, but as
18 a special case of the theory. In general, a given 3D estimate \hat{z} is assumed to be proportional to the corresponding distal 3D
19 property z , through a proportionality factor that expresses the strength of the visual information at any given time.

20 To illustrate the concept of information strength in a visual scene, let's consider how depth is estimated from two
21 fundamental cues: binocular disparity and motion. At a distance within the peripersonal space of an agent, depth judgements
22 from relative disparity are usually quite accurate. Formally, we can re-write this statement as $\hat{z}_D = k_D z$, with $k_D = 1$. When
23 objects are farther away from the observer, certain distance information that is necessary to scale binocular disparities
24 becomes weaker ($k_D < 1$). Eventually, at very large distances k_D approaches zero. In a similar fashion, depth from motion
25 information depends on the relative motion between the observer and the distal object. For ideal conditions, $\hat{z}_M = k_M z$ with
26 $k_M = 1$. However, for a static observer the proportionality factor k_M is nil.

27 In other words, separate 3D modules analyze distinct depth cues (e.g., disparity and motion) which are affected by
28 independent confounding variables associated with viewing conditions (e.g., fixation distance and observer's motion,
29 respectively). For the i -th depth cue, the parameter k_i embodies the effect of those confounding variables on the mapping
30 between the physical depth z and its estimate \hat{z}_i . In the framework of the IC theory, we will refer to k_i as cue strength.

31 Goal of the IC model is *to maximize sensitivity to depth while keeping the influence of confounding variables to the*
32 *minimum possible*. In the Appendix (Proof 1) we show that this goal is achieved computationally by a model that combines
33 single-cue estimates through a vector sum. For the general case of a vector space with n image signals, the combined estimate
34 \hat{z}_C is the resultant vector given by the equation:

1
$$\hat{z}_C = \sqrt{(\hat{z}_1)^2 + (\hat{z}_2)^2 + \cdots (\hat{z}_n)^2} = \sqrt{(k_1 z)^2 + (k_2 z)^2 + \cdots (k_n z)^2} \quad [1]$$

2 Note that this model is additive, that is, increasing the number of cues yields larger depth estimates. This should not
3 be misunderstood as the combination rule producing systematic overestimations of depth as more cues are added to a
4 stimulus. On the contrary, the underlying logic is that removing cues brings the model outside of its optimal operating
5 conditions, which results in the underestimation of single-cue stimuli (namely, stimuli isolating only one source of
6 information).

7 The goal of this study was twofold. First, it aimed to investigate whether the phenomenon of superadditivity
8 observed in VR generalizes to the physical world and is therefore inherent to the process of depth-cue combination. To this
9 end, we used real objects and measured their estimated depth through both an explicit (perceptual) and an implicit (motor)
10 task. Second, the study also served as a quantitative test of the IC model predictions. This second aim requires a careful
11 assessment of all individual depth signals that are at play in a given visual scene. In principle, to isolate a specific depth cue
12 and be able to measure its estimate \hat{z}_j , one must reduce the strength k_i of all the other cues to zero, such that the combined
13 estimate $\hat{z}_C = \hat{z}_j$ as per equation [1]. This logic applies to both VR and real-world environments. For instance, a static
14 object does not provide any depth from motion information – which in IC terms means that the strength of the motion cue
15 k_M is equal to zero. Certain types of depth signals, however, cannot be eliminated, and that is because they do not depend
16 on the visual input itself but on the limits of the human optical apparatus. We will use the term “focus cues” to refer to all
17 these signals at once, of which the most relevant one is the blurring gradient. In a VR setting, focus cues specify the flat
18 surface of the monitor screen ($z = 0$), in contrast with the three-dimensional quality of the simulated environment. With
19 real objects, focus cues add 3D shape information on top of, and consistent with, what already specified by stronger depth
20 cues such as disparity or motion.

21 Since in the present study we investigated the contribution of stereo and motion cues to the estimation of depth,
22 equation [1] becomes:

23
$$\hat{z}_C = \sqrt{(\hat{z}_D)^2 + (\hat{z}_M)^2 + (\hat{z}_F)^2} = \sqrt{(k_D z)^2 + (k_M z)^2 + (k_F z)^2} \quad [2]$$

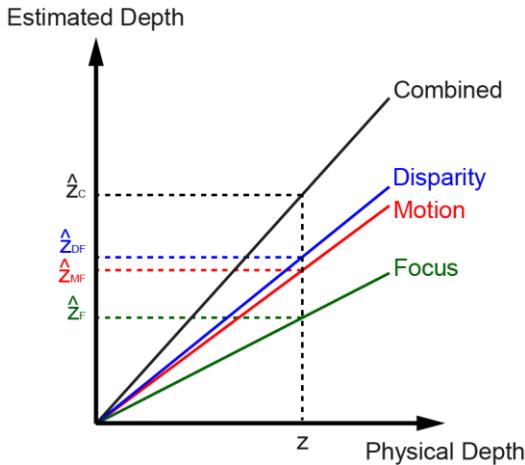
24 where the depth estimates \hat{z}_i are all related to the same physical object’s depth z , but each is scaled by the strength
25 term appropriate for the signal in question (k_D , k_M and k_F for disparity, motion and focus cues, respectively). Importantly,
26 focus cues are dramatically different in computer-generated versus real-world scenes. VR environments require the presence
27 of a flat screen, therefore the depth associated with the focus cues is zero and the last term of equation [2] is dropped. Under
28 these visual conditions, it is possible to study a true disparity-only (that is, motion has strength $k_M = 0$) condition and test
29 whether adding motion to the display ($k_M \neq 0$) yields larger magnitudes of perceived depth. This was done in our past study
30 (Campagnoli and Domini, 2019). Consider now a real object viewed at a distance yielding sizable focus cues, like a
31 reachable distance. Unlike VR, this viewing condition does not enable to isolate disparity and motion cues, because the
32 focus cues are also attuned to the 3D structure of the same object, given that the scene is no longer simulated on a flat screen.

1 As a result, the “disparity-only” stimulus will in fact be a combination of disparity and focus signals (yielding a depth
 2 estimate \hat{z}_{DF}), and the same will be true for the “motion-only” stimulus (yielding a depth estimate \hat{z}_{MF}). This will impact
 3 how the combined estimate of depth \hat{z}_C is derived from the three cues. It can be demonstrated (see Appendix, Proof 2) that
 4 combined cue estimates from “single-cue” estimates is given by:

5

$$\hat{z}_C = \sqrt{(\hat{z}_{DF})^2 + (\hat{z}_{MF})^2 - (\hat{z}_F)^2} = \sqrt{k_{DF}^2 + k_{MF}^2 - k_F^2} z \quad [3]$$

6 This operational definition allows to make specific quantitative predictions about superadditivity, since equation [3]
 7 is effectively expressed as a function of “single-cue” estimates. In this investigation, we tested this prediction of the IC
 8 model with stimuli that allowed to equalize the strength of the disparity and the motion cues ($k_S \approx k_M$). This way, we
 9 expected that the psychometric functions relating disparity- and motion-derived depth estimates to the physical depth of the
 10 stimuli would have approximately the same slope (Fig. 1, red and blue lines). Secondly, we also expected a significant effect
 11 of the focus cues ($k_F \neq 0$, green line in fig. 1). Finally, equation [3] allowed to predict the slope of the combined-cue
 12 psychometric function (Fig. 1, black line) without introducing additional free parameters, but only from the slopes (strength
 13 terms) of the single-cue functions.



14
 15 Figure 1. Superadditivity prediction: estimates of depth (values on the ordinate) were expected to be greatest for the visual
 16 condition with the maximum amount of 3D sources (Combined, in black), intermediate under Disparity and Motion (blue and red
 17 respectively), and smallest for the condition with the scarcest amount of depth information (Focus, in green). The increase in slope of
 18 the Combined estimate was predicted on the basis of the slope of the other conditions according to equation [3].
 19

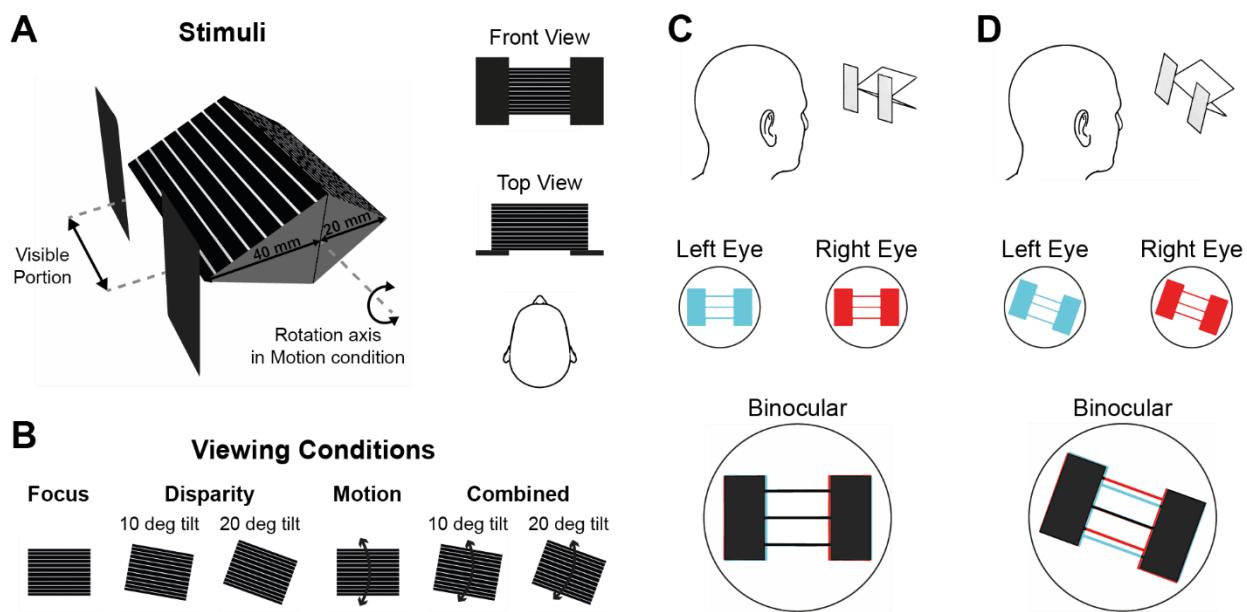
20 *Manipulating depth information with real objects*

21 In a natural environment, binocular vision and motion parallax are among the most important depth cues available in
 22 the visual array. While it is relatively easy to balance depth cues in VR, it is much harder to do so in a natural environment.
 23 The biggest challenge is represented by stereovision, since objects seen within reaching distance inevitably produce a strong

1 stereo signal which prevails over the other signals to depth. On the other hand, the pattern of retinal velocities resulting from
2 the relative motion between object and observer can be easily modulated by rotating the object at different speeds.

3 To overcome this problem, we designed a novel technique to allow binocular disparities to have similar strength to
4 that of retinal velocities for the judgment of depth, without using optical devices or introducing other artificial surrogates.

5 We constructed prism-shaped objects out of cardstock, to be used as stimuli (Fig. 2A). When oriented horizontally,
6 the visible portion of the stimuli consisted of 60 mm wide, 40 mm tall and 20 or 40 mm deep wedge. The surface of the
7 objects was matte and painted with white stripes that ran parallel to the long side. Subjects saw the visible portion of the
8 prisms in six configurations (Fig. 2B): static horizontal (Focus condition), static tilted (Disparity condition; 10 and 20 deg
9 tilt), horizontal with motion (Motion condition), and tilted with motion (Combined condition; 10 and 20 deg tilt).



10
11 Figure 2. (A) Schematic of the stimuli used in the experiment (see also the pictures in figure 3). Two prisms (20 and 40 mm in depth)
12 were attached back-to-back and displayed one at the time by rotating them 180 degrees around the axis passing through the common
13 face (the same axis for the rocking of the Motion condition). Two rectangular occluders prevented the vision of the edges of the stimulus.
14 (B) Viewing conditions of the experiment. Binocular disparity was gradually added by tilting the stimulus relative to the frontoparallel
15 plane (Disparity). Motion was added by rocking the stimulus around its main axis at constant angular velocity. (C) When the stimulus
16 is horizontal, the left and right retinal projections (in cyan and red, respectively) align such that after fusion ("Binocular" inset) the
17 binocular disparities from the prism's surface are unspecified, because the correspondence problem cannot be solved. (D) Tilting the
18 stimulus relative to the frontoparallel plane disambiguates the binocular disparities: note the greater binocular disparity (misalignment
19 of the cyan and blue projections) at the top and bottom edges of the prism relative to the tip (assumed to be at fixation). The misalignment
20 of the retinal projections increases up to a maximum tilt angle of 90 degrees (vertical stimulus – see Van Ee & Schor, 2000).

21 During the Focus condition, the prism was static and oriented horizontally along its long side, with the tip facing the
22 observer. When in this position, the horizontal disparities were unspecified, since the uniformity of the horizontal lines on

1 its surface made the correspondence problem unsolvable (figure 2C). The overall angular size of the visible stimulus (about
2 8 degrees) was such that vertical disparities were negligible. Moreover, the presence of cardboard occluders eliminated
3 contour information from the stimulus edges. In this visual condition, subjects could rely mainly on focus cues of extraretinal
4 origin, such as ocular vergence and lens accommodation.

5 In the Disparity condition, the object was static and also tilted around the axis perpendicular to the frontoparallel
6 plane, pivoting around the cyclopean eye. Rotating the surface in a direction orthogonal to the orientation of the stripes
7 provided an efficient way to selectively manipulate the strength of stereo information (for a detailed analysis of the
8 mechanisms involved see Van Ee & Schor, 2000). As the tilt increased, the binocular disparities blended less and less into
9 the white stripes, becoming more resolvable (figure 2D). The limit case would be represented by a tilt of 90 degrees,
10 corresponding to a perfectly vertical prism, yielding maximum binocular disparity. For this study, we selected two small tilt
11 angles, 10 and 20 degrees. This fulfilled two goals: 1) to test if the strength of stereo did increase with a greater tilt angle –
12 as a general test of the technique itself; 2) most importantly, to limit the overall strength of stereo information relative to
13 motion (when present), so that the two signals could contribute to depth perception equally. Overall, the total depth
14 information in the Disparity condition was the sum of the focus cues plus the disparity gradient.

15 When the stimulus was horizontal and rocking around the horizontal axis (Motion condition), a pattern of relative
16 velocities of the white stripes allowed the observer to judge the prism's 3D layout from the retinal motion gradient. Note
17 that this rotation did not interact with the gradient of binocular disparities, which remained unspecified. Therefore, the total
18 depth information in this condition resulted from the combination of focus cues plus the motion gradient.

19 Finally, in the Combined condition, the prism was tilted relative to the frontoparallel plane as in the Disparity
20 condition, and it also rocked along its tilt axis, so that the movement did not change the orientation of the stripes, as in the
21 Motion condition. During this presentation, the binocular disparities generated from tilting the object were combined with
22 the motion gradient produced by the object's oscillation. This visual condition provided the maximum amount of depth
23 signals specifying the stimulus geometry: focus cues plus disparity and motion gradients.

24

25 Materials and Methods

26

27 Participants

28 Forty-three students gave written informed consent to participate in the experiment (29 females, age ranging between
29 18 and 24). All participants self-reported normal or corrected-to-normal vision. The participants received either a
30 reimbursement of \$8/hour or coursework credit for their effort. The experiment was approved by the Institutional Review
31 Board at Brown University. Each participant gave informed consent prior to the experiment.

32

33

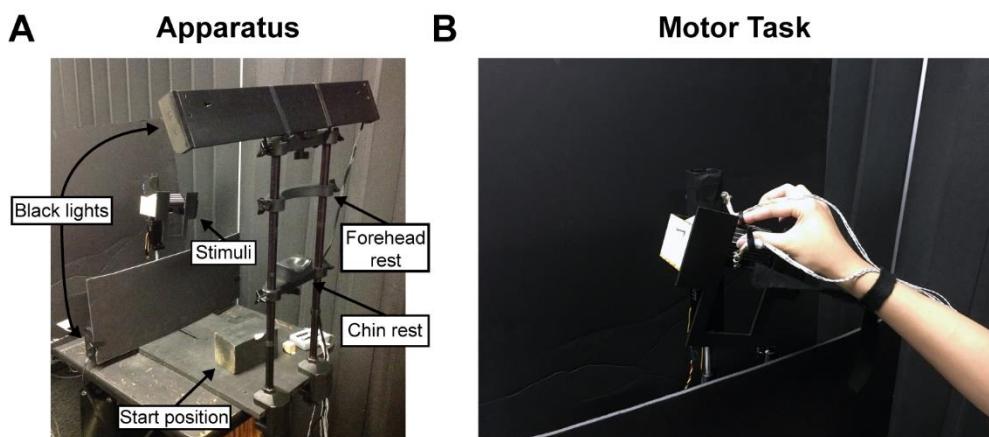
34

1 **Apparatus**

2 Subjects sat with their head locked on an optometric chin and forehead rest, installed along one of the short sides of
3 a rectangular table, directly facing the physical stimulus (figure 3A). The stimulus consisted of an isosceles triangular prism
4 made from cardstock, oriented with the tip facing the participant. We built two of these prisms, one with a depth (tip-to-
5 base extent) of 20 mm (small) and the other of 40 mm (large). The prisms were attached base-to-base such that only one
6 was visible per trial (figure 2A). Printed onto the prism surface were pure white pinstripes on a matte black ink base, spaced
7 6.5 mm apart.

8 During the experiment, the testing room was dark and two black lights guaranteed uniform illumination of the
9 presented stimulus; the net effect was a series of bright white stripes on a dark surface. The two prisms were mounted on a
10 custom, laser-cut wooden base, which was then installed on a vertical metal stand. The position and height of the stand was
11 adjusted such that the tip of the prism was 42 cm above the tabletop, matching the average eye height of the subjects, and
12 either 44 or 42 cm away from the observer, for the small and the large object respectively. Rectangular wooden occluders,
13 also painted matte black, covered the far left and right sides of the stimulus to prevent participants from detecting the edges
14 of the prisms. The visible portion of the stimulus thus was 60 mm wide by 40 mm tall.

15 Two stepper motors (Phidgets Inc.) were used to switch between stimuli as well as to manipulate their orientation and
16 motion. The first motor rotated the prisms around the horizontal axis centered on their plane of contact (the prisms' bases),
17 producing an up-and-down rocking motion of the prism's tip at 60 deg/sec that swept a total angle of 15 degrees (± 7.5
18 degrees, figure 2A and 2B). The second motor rotated the stimulus and occluders around the axis perpendicular to the
19 subject's frontoparallel plane and centered on the cyclopean eye, setting a total of three angles: 0 degrees (Baseline and
20 Motion conditions), ± 10 and ± 20 degrees (Disparity and Combined conditions). The values of the angles for the Disparity
21 condition and of the angular velocity for the Motion condition were chosen based on pilot data in order to equalize the
22 strength of stereo and motion information.



23
24 Figure 3. Apparatus (A) and grip gesture in the motor task (B), pictures taken from the actual experimental setup.

25 Although the integration of disparity and motion cues may lead to two additional sources of binocular information,
26 namely changing disparity over time (CDOT) and interocular velocity difference (IOVD – Harris et al., 2008), these were

1 prevented in the Combined condition. CDOT and IOVD are informative about motion-in-depth, that is when an object is
2 approaching the observer along the line of sight. Our setup ensured that motion-in-depth was negligible. Given the small
3 rotation of the prism about an axis parallel to the white stripes, the motion of each stripe was almost entirely in the front
4 parallel plane, with a negligible component along the line of sight. Consequently, binocular disparities did not change as a
5 result of the rotation, effectively stripping away this potential source of information.

6 Prior to each trial, the black lights turned off. The first motor rotated to an intermediate position between the prisms,
7 so that subjects could not anticipate which prism was being prepared based on auditory cues, and then positioned the next
8 stimulus. The scene was then illuminated again, via a custom *USB* relay that synced the lights' power source to the motion
9 of the motors, allowing participants to see the stimulus (and their hand, in the case of the implicit task).

10 To record the location of the subject's fingers, we used an *NDI* Optotrak Certus motion capture system, with the
11 position sensor facing the workspace at a distance of 224 cm from the subject. Subjects wore index and thumb markers with
12 protruding flags that held three *NDI* infrared *LEDs* each, oriented such that the infrared light would directly hit the Optotrak
13 cameras. Optotrak recordings, the motion of both Phidgets motors and the *USB* relay were all controlled by custom C++
14 programs using specific *API* routines provided by each vendor.
15

16 **Procedure**

17 Each subject underwent thorough calibration upon putting on the infrared *LEDs*. First, the experimenter calibrated
18 the XYZ axes of the Optotrak position sensor so that the origin was aligned with the observer's cyclopean eye and the
19 forward direction coincided with the line of sight. Second, the experimenter recorded the position of the subjects' fingers
20 on a designated home position, marked by a wooden block positioned on the tabletop about 35 cm away from the stimulus
21 (figure 3A). Participants performed an explicit (perceptual) task and an implicit (grasping) task in 4 separate blocks,
22 following an ABBA order.

23 In the explicit task, the stimulus was illuminated at the beginning of each trial, and subjects performed a Manual
24 Estimation (*ME*), matching the perceived tip-to-upper-edge distance to the proprioception of the distance between thumb
25 and index fingers. The contact surface had length of 28.3 and 44.7 mm for the small and large prism respectively. Subjects
26 were not allowed to look at the hand while performing the task. To confirm their estimate, they pressed a button with the
27 other hand. At the button press, the lights turned off, and the next trial would load as the subjects returned to home position.

28 In the implicit task, once the lights turned on, subjects reached out and physically contacted the tip and upper rear
29 edge of the stimulus with their thumb and index fingers, respectively (fig. 3B). Subjects could see their hand and feel the
30 prism. If the prism was moving, it stopped once the index markers passed 40 cm in the z direction, to make certain that
31 subjects could firmly contact it. To ensure a comfortable grasping posture, we did not constrain the thumb and index final
32 positions to specific locations along the edges of the prism. On average, subjects approached the object by gradually rotating
33 the wrist, such that the index's contact position was 8-15 mm farther to the right than the thumb's contact position. This

1 behavior produced final grip apertures that were systematically larger than the actual stimulus's side. When subjects returned
2 their fingers to the home position, the lights went off, and the motors prepared the next presentation.

3 Each task was comprised of 144 trials, divided in two blocks: in each block, a total of six viewing conditions (Baseline,
4 Disparity 10 deg tilt, Disparity 20 deg tilt, Motion, Combined 10 deg tilt, Combined 20 deg tilt) were presented six times
5 for each depth (total of 72 trials). The trials were pseudo-randomized within each block: participants performed six adjacent
6 strings of 12 unique trials (6 viewing conditions times 2 depths) randomized within repetitions. Including both explicit and
7 implicit task, participants performed a total of 288 trials.

8

9 *Analysis of kinematic data*

10 Offline analysis of kinematic data was performed using R (R Core Team, 2020). In every trial of every subject, the
11 frame-by-frame 3D positions of index and thumb markers were checked for missing frames (replaced through spline
12 interpolation), smoothed with a fourth-order Savitzky-Golay filter, and used to compute the instantaneous Euclidean
13 distance between them, or Grip Aperture (GA). The thumb path was then segmented into regular intervals (or bins) of 2 mm
14 each, and the mean GA was calculated in each bin of each trial of each subject. The GA's grand mean and variance were
15 then calculated by pooling across the trials of each viewing condition and object side (bin-by-bin, subject-by-subject).

16 Further dependent variables extracted from the trajectory data (i.e., Maximum Grip Aperture) were analyzed using
17 standard tests, such as repeated measures ANOVA and t-test.

18

19 **Results**

20 In this study, we sought to compare two estimates of 3D shape: an explicit estimate from a perceptual task, requiring
21 participants to report the perceived size of cardboard prisms through a manual matching task; an implicit estimate from a
22 motor task, where the same individuals reached to touch the prisms with their index and thumb. The amount of depth cues
23 available on the surface of the stimulus in each trial was varied by rotating and oscillating the prism so to hide or show
24 disparity and motion signals and a combination thereof, according to four visual conditions (Focus, Disparity, Motion and
25 Combined).

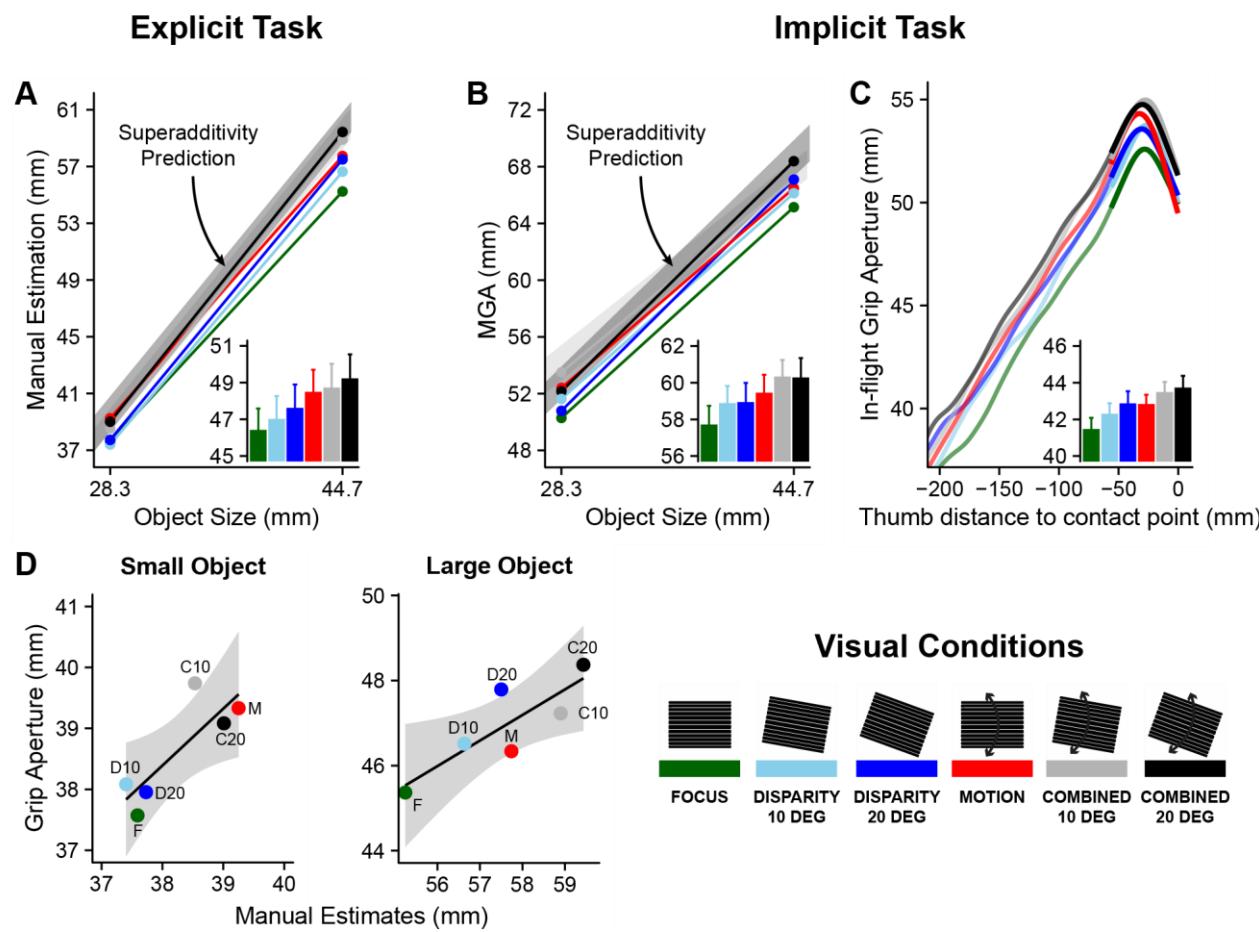
26 Explicit manual reports of 3D size were significantly affected by the amount of depth information available in a given
27 trial (fig. 4A). Participants estimated both small and large objects with the smallest ME in the Focus condition, the largest
28 ME in the Combined conditions, while the Disparity and Motion conditions fell in the middle, as visible in the bar graph in
29 the inset of figure 4A. A repeated measures ANOVA on the Manual Estimations (ME) found a main effect of cue
30 combination ($F_{(5,210)} = 18.6$, $p < .01$, $\eta_p^2 = .31$), a main effect of the object size ($F_{(1,42)} = 287.49$, $p < .01$, $\eta_p^2 = .87$) and a
31 significant interaction between cue combination and object size ($F_{(5,210)} = 6.22$, $p < .01$, $\eta_p^2 = .13$).

32 A Bonferroni-Holm corrected post-hoc analysis confirmed the pattern suggested by the visual inspection of the data.
33 The Focus condition yielded the smallest ME (smaller than the Disparity conditions averaged together: $F_{(1,42)} = 12.48$, $p <$
34 $.01$; than Motion: $F_{(1,42)} = 29.74$, $p < .01$; than the Combined conditions averaged together: $F_{(1,42)} = 35.71$, $p < .01$). The

1 average MGA in the Combined conditions was larger than in the Disparity conditions ($F_{(1,42)} = 27.20$, $p < .01$) but not larger
 2 than in the Motion condition ($F_{(1,42)} = 2.63$, $p = .11$). The ME of the Motion condition was also significantly greater than
 3 that of the Disparity condition ($F_{(1,42)} = 15.65$, $p < .01$).

4 The interaction term explains why the increase in depth estimation in the Combined condition relative to Motion did
 5 not reach statistical significance. An additional post-hoc analysis showed a significant difference for the large object ($F_{(1,42)} = 11.35$, $p < .01$) but not for the small object ($F_{(1,42)} = 1.48$, $p = .46$). This is consistent with the IC model, which posits that
 6 the depth estimate of a given combination of depth cues is proportional to the amount of physical depth that that composite
 7 signal is estimating. As it can be seen in figure 4A, the range of the average manual estimates for the small object was less
 8 than half of the range for the large object. The small object was too small to allow the individual visual conditions to clearly
 9 spread out and separate in virtue of their different strengths, as instead happened for the large object.

10
 11 The effect of tilting the object from 10 to 20 degrees was significant ($F_{(1,42)} = 6.98$, $p = .01$, $\eta_p^2 = .14$) and consistent
 12 in both Disparity and Combined conditions. Overall, the ME gradually increased as more depth information was available
 13 (inset of figure 4A).



14
 15 Figure 4. (A) Explicit 3D shape estimation: Manual Estimates as a function of the stimulus size, for each visual condition (color legend
 16 in the “Visual Conditions” panel). The inset shows the same data averaged across the stimulus size (error bars indicate between-subjects

1 SEM). The gray areas represent the 95%CI of the IC model predictions for superadditivity. (B) Implicit 3D shape estimation: Maximum
2 Grip Aperture as a function of the stimulus size, for each visual condition (data averaged across stimulus size in the inset). (C) Average
3 Grip Aperture over the course of the reaching trajectory, as function of the instantaneous Euclidean distance of the thumb to its final
4 contact location, for each visual condition (portions in full opacity denote effect of cue combination with p-value < .05). The inset shows
5 the GA averaged across the entire trajectory. (D) Correlation between mean explicit reports (Manual Estimates, in abscissa) and implicit
6 measures (mean Grip Aperture, in ordinate), for the small and the large objects (left and right panel, respectively).

7 The results from the implicit task agreed almost exactly with those from the explicit task (figure 4B and 4C). This
8 congruence in the behavior in the two cases highlights the importance that depth information has for the perception-action
9 coupling, especially considering that even tenuous variations in the 3D structure of the scene had yet comparable effect
10 sizes on both tasks. Participants approached both small and large objects with the smallest Maximum Grip Aperture (MGA)
11 in the Baseline condition, the largest MGA in the Combined conditions, while the Disparity and Motion conditions fell in
12 the middle (Fig. 4A). A repeated measures ANOVA on the MGA found a main effect of object size ($F_{(1,42)} = 426.73$, $p <$
13 $.01$, $\eta_p^2 = .91$), of cue combination ($F_{(5,210)} = 7.73$, $p < .01$, $\eta_p^2 = .16$), and a significant interaction between the two ($F_{(5,210)} =$
14 2.28 , $p = .048$, $\eta_p^2 = .05$).

15 The post-hoc results were also essentially the same. The Focus condition yielded the smallest MGA (smaller than the
16 Disparity conditions averaged together: $F_{(1,42)} = 6.65$, $p = .04$; than Motion: $F_{(1,42)} = 13.04$, $p < .01$; than the Combined
17 conditions averaged together: $F_{(1,42)} = 19.31$, $p < .01$). The average MGA in the Combined conditions was larger than in the
18 Disparity conditions ($F_{(1,42)} = 20.26$, $p < .01$) but not larger than in the Motion condition ($F_{(1,42)} = 2.64$, $p = .22$). The MGA
19 of the Disparity and Motion conditions were not different from each other ($F_{(1,42)} = .84$, $p = .36$).

20 A repeated measures ANOVA on a subset including the MGA of the Disparity and Combined conditions also found
21 a significant two-way interaction between the tilt angle and the object's size ($F_{(1,42)} = 14.19$, $p < .01$, $\eta_p^2 = .25$), as the slope
22 relating the MGA to the object's size was greater when the object was tilted 20 degrees than 10 degrees. This means that
23 the subject's sensitivity to changes in the object's size increased when the binocular disparities were more reliable. Again,
24 a clearer picture of the results can be obtained by averaging across the two objects sizes, as per the inset of figure 4B, where
25 the gradual increase of the MGA as the 3D information accumulates is most noticeable.

26 The Grip Aperture during the implicit task was then analyzed over the course of the entire movement, from its onset
27 to the instant when the thumb touched the stimulus (the thumb is typically the first digit to contact an object in a front-to-
28 back precision grasp; Volcic and Domini, 2014). A first visual inspection of the temporal evolution of the GA revealed
29 essentially the same pattern of results found with both MGA and ME (figure 4C). This distribution was qualitatively constant
30 throughout most of the reaching path, with a visible cue combination effect which eventually turned significant in the final
31 part of the grasp (when the thumb was closer than 50 mm from the object).

32 A repeated measures ANOVA on the grip aperture averaged across the entire trajectory (inset of figure 4C) using cue
33 combination as factor confirmed the main effect ($F_{(5,210)} = 13.89$, $p < .01$, $\eta_p^2 = .25$). A post-hoc analysis also confirmed the

1 pattern suggested by the visual inspection. The Focus condition yielded the smallest GA (smaller than Disparity: $F_{(1,42)} = 16.75$, $p < .01$; than Motion: $F_{(1,42)} = 17.82$, $p < .01$; than Combined: $F_{(1,42)} = 49.87$, $p < .01$). The Combined condition
2 yielded the largest GA (larger than Disparity: $F_{(1,42)} = 17.02$, $p < .01$; than Motion: $F_{(1,42)} = 13.01$, $p < .01$). The grip apertures
3 of the Disparity and Motion conditions were in the middle, not different from each other ($F_{(1,42)} = .76$, $p = .39$). Finally, there
4 was a close-to-significant effect of tilting the object from 10 to 20 degrees in both Disparity and Combined conditions ($F_{(1,42)} = 3.27$, $p = .078$, $\eta_p^2 = .07$).

5 As per the main aim of the study, the results from the Combined conditions in both explicit and implicit tasks were
6 tested for superadditivity according to the IC model. For each condition of each task of each participant, the slope of the
7 psychometric function relating the estimated size to the physical size was calculated through a linear fit. Each one of these
8 slopes expressed the strength of the depth information in each condition (i.e., the term k_i in equation [2]). The k terms
9 obtained from the Focus, the Disparity (10 and 20 degrees separately) and the Motion conditions were then combined
10 according to equation [3], resulting in a prediction of the strength of the depth information in the Combined condition (k_C).
11 This calculation was performed for the small and large objects separately, for a total of four k_C terms for each task (small-
12 10deg, small-20deg, large-10deg, large-20deg). These proportionality factors were finally used to scale the physical values
13 of the stimuli to retrieve an estimate of the stimulus size, and of the resulting psychophysical function by interpolation.
14 Figures 4A and 4B show the 95% confidence intervals of the predicted object's size: the light gray areas depict the predicted
15 range for the Combined-10 condition, the dark gray areas that of the Combined-20 condition. The empirical values of both
16 the ME and the MGA in the Combined condition were in accordance with the IC model, providing evidence in support of
17 the idea that depth information is integrated so to maximize the sensitivity to 3D shape information. Two additional pieces
18 of evidence further reinforce the idea that this combination rule appears to be a fundamental mechanism underlying the
19 analysis of depth information in the brain. First, the same prediction method was successfully applied independent of the
20 level of conscious processing of depth information, as the model predicted both the perceptual and the motor results. Second,
21 the symmetry of the results from this study with the results of the previous experiments in VR (Campagnoli et al., 2019),
22 demonstrate that our combination rule well describes how depth signals are translated into estimates in the visual system,
23 and is independent of the medium (virtual versus real objects) conveying depth information to the brain.

24 The congruence between the perceptual and the motor performances was further assayed statistically. A 6x2 factorial
25 ANOVA was performed crossing cue combination (6 visual conditions) with type of task (explicit or implicit). The
26 interaction term was not significant ($F_{(5,210)} = 1.43$, $p = .22$, $\eta_p^2 = .03$), suggesting that the effect of cue combination was the
27 same in both tasks. The ANOVA results were further confirmed by a correlation analysis between the perceptual and the
28 motor data. For each task, the responses were averaged across each of the object sizes and each of the six viewing conditions.
29 The datasets from the explicit task and the implicit task (average GA over the reaching) were then compared against each
30 other. Figure 4D shows the resulting highly significant correlation that was found for both object sizes ($r_{smallObj} = .84$, $p =$
31 $.04$; $r_{largeObj} = .84$, $p = .03$; $n = 6$).

32

1 Discussion

2 When analyzing the 3D structure of an object, the visual scene provides the observer with a rich set of depth cues,
3 ranging from monocular gradients (i.e., texture and motion) to binocular disparity. In a recent study we found that estimates
4 of depth do not scale accurately with an object's 3D shape, but rather they increase with the amount of depth information
5 that specifies that shape (Campagnoli and Domini, 2019). In the present study we set out to investigate whether this behavior
6 is the signature of a genuine superadditivity effect characterizing the processing of depth information by the visuomotor
7 system, or instead an artifact due to top-down control caused by immersion in a virtual environment. In previous studies we
8 proposed that the above results conform to a model of combination of depth signals termed Intrinsic Constraints model of
9 cue integration (IC model – Domini & Caudek, 2011, 2013; Di Luca et al., 2010; Kemp et al., 2018). While the model is
10 able to capture the behavior quantitatively, here we address the question about the nature of the phenomenon, by testing the
11 model's combination rule in a real-world context rather than in VR. Furthermore, to investigate a possible influence of
12 cognition on the responses, we asked participants to report the size of cardboard prisms both through an explicit task (manual
13 estimation) and an implicit task (reaching to touch). Participants performed both tasks under four different combinations of
14 focus, disparity and motion depth cues. The IC model was used to predict the responses in the condition where all cues were
15 present at the same time (Combined), based on the responses in the three intermediate conditions (Focus, Disparity and
16 Motion).

17 The first significant evidence is represented by the main effect of cue combination: given the same physical prism,
18 subjects' estimates increased as more depth information was accessible in the scene. This pattern appeared to be independent
19 of cognitive control, as it was detected in both explicit and implicit tasks. In the explicit task, participants increased the size
20 of their Manual Estimates (ME) progressively from the Focus condition through Disparity, Motion and Combined. The
21 results of the implicit task shows that this progression was inherent, systematic and robust, as it was found throughout the
22 entire reaching trajectory and with two different object sizes. As soon as the hand was close enough to the target object, the
23 vision of the hand allowed participants to guide their fingers onto the physical surface of the object. This is in line with
24 previous evidence showing that what ensures a successful grasp movement is a combination of online visual control during
25 a given trial and sensorimotor error-prediction to adjust the movements of the following trials (Bozzacchi et al., 2014, 2016;
26 Bozzacchi & Domini, 2015; Campagnoli & Domini, 2019; Campagnoli et al., 2012; Campagnoli et al., 2015; Campagnoli
27 et al., 2017; Cesanek et al., 2018; Cesanek & Domini, 2017; Foster et al., 2011; Volcic et al., 2013; Volcic & Domini, 2014,
28 2016).

29 The small effect size of cue combination remarks the strong interconnection between the perceptual system and the
30 motor system during the processing of depth information. The magnitude of the effects between visual conditions was in
31 the order of few millimeters, and yet even such subtle responses to changes in the 3D layout of the environment reverberated
32 with similar intensity on both the grip aperture and the manual estimates.

33 The second major evidence is the adherence of the data to the predictions of the IC model. The model considers depth
34 signals as scalar components of a multidimensional vector. Since in this representation each signal is assumed to be

1 proportional to the 3D properties it encodes (e.g. depth), so it is the magnitude of the global vector. The fundamental
2 assumption of this model is that once the scalar components have been appropriately scaled, the magnitude of the vector is
3 the sole determinant of the perceived 3D property. It follows from this description that adding depth cues to a stimulus is
4 equivalent to “switching on” components of this multidimensional vector with the effect of increasing its intensity, hence
5 perceived depth. This principle can be formalized into equations predicting how depth should be estimated when multiple
6 3D cues are available in the scene, knowing the effect of the cues on depth estimation when acting separately. Figure 4
7 shows that the IC model’s predictions correctly captured the results of the Combined conditions, using the results from the
8 Disparity, Motion and Focus conditions.

9 IC is therefore sufficient to explain the phenomenon of superadditivity – depth estimates increase with the amount of
10 depth information – found both in this study and in VR (Campagnoli & Domini, 2019). Moreover, it can do so without the
11 necessity of further ad-hoc assumptions such as prior-to-flatness (Di Luca et al., 2010; Domini and Caudek, 2003, 2009,
12 2010, 2011, 2013; Domini et al., 2006; Domini et al., 2011; Kemp et al., 2018).

13 The role of flatness information has been widely discussed in the context of a Bayesian explanation for superadditivity
14 of disparity and texture cues. Under this framework, the combination of these particular cues has been assumed and modeled
15 as linear, such that, if z_d and z_t indicate the depth estimates from disparity and texture, the combined estimate \hat{z} is a weighted
16 sum $\hat{z}_c = w_t z_t + w_d z_d$. The weights are inversely proportional to the variance of the noise associated with the depth from
17 texture (σ_t^2) and from disparity (σ_d^2) estimates and they must sum to 1. It follows that, as long as the two cues specify the
18 same depth value $z = z_t = z_d$, the combined estimate is on average the same as that of each individual cue. Formally,
19 $\hat{z}_c = \hat{z}_t = \hat{z}_d$ (since $w_t + w_d = 1$). To reconcile the empirical finding that $\hat{z}_c > \hat{z}_t$ and $\hat{z}_c > \hat{z}_d$, it has been hypothesized
20 that, in virtual displays, accommodation cues and the gradient of the retinal blur are in conflict with the rest of the depth
21 signals, in that they indicate zero depth, that of the CRT monitor (Buckley & Frisby, 1993; Frisby et al., 1995; Hoffman et
22 al., 2008; Watt et al., 2005; Willemse et al., 2008). Once these cues are taken into consideration, their weight can potentially
23 affect the output of the above weighted sum. That is, if the reliability of the focus cues is sufficiently great, they are able to
24 pull the depth estimate towards zero with greater strength when the individual depth cues are seen in isolation, relative to
25 when the same depth cues are combined. It is therefore not surprising then that adding depth cues increases the depth
26 estimates, because the total information about depth increases its prevalence over the flatness information carried by the
27 focus cues. Within this framework, superadditivity should be considered as restoring flattened percepts.

28 Although this explanation accounts for several viewing conditions, there is one representative example case where
29 flatness cues did not have an effect on perceived depth. In a study by Vishwanath and Hibbard (2013), observers were asked
30 to judge the depth of a hemi-cylinder depicted through texture information. No other depth information was available on
31 the stimulus, which was rendered on a CRT screen located at 50 cm from the observers. As expected, when the cylinder
32 was viewed with one eye (monocular condition), the perceived depth reliably increased with the simulated depth. Most
33 notably, depth judgements were not significantly different when the display was viewed with both eyes opened (binocular
34 condition), which is remarkable since binocular vision introduces disparities specifying the flat surface of the screen, unlike

1 monocular vision. Note also that a disparity field indicating a flat frontoparallel surface is highly reliable. In fact, the flatness
2 impression derived from binocular disparity should be at least as reliable as the impression of depth conveyed by texture
3 information, and certainly much more reliable than the flatness impression specified by the focus cues. Why did the
4 observers not perceive the depth of the binocularly viewed cylinder as considerably flatter than the depth of the monocularly
5 viewed cylinder? The logic of the Bayesian approach fails to explain both this phenomenon and superadditivity. This is
6 because a critical requirement of a Bayesian estimator is that it must assess the reliability of all the sources of 3D information
7 at all times. If focus cues are present, the estimator needs to evaluate their reliability in relation to those of the disparity and
8 texture cues. If it detects a zero disparity field, it needs to first determine whether it arises from a close-by frontoparallel
9 surface or from an undetermined structure seen from far away. Paradoxically, this fundamental feature of the Bayesian
10 estimator seems to work in the first case, where a faint signal flattens reliable texture and disparity estimates, explaining
11 why combined stimuli were perceived deeper than motion or disparity stimuli. However, it falters in the second case where
12 the signal to flatness is unmistakably reliable and should therefore dramatically reduce the perceived depth of the cylinder
13 in Vishwanath and Hibbard (2013) study.

14 In contrast, the IC model can be successfully applied to both this latter case and that of superadditivity. If only disparity
15 and texture cues are available, then, according to equation [3],

$$\hat{z}_c = \sqrt{k_t^2 + k_d^2} z,$$

17 where k_t , k_d are the gains associated with texture and disparity and z is the simulated depth. When only disparity
18 information is available, the gain associated with texture is nil ($k_t = 0$) and only the disparity gain contributes to perceived
19 depth: $\hat{z}_d = k_d z$. This easily explains superadditivity since $\hat{z}_c > \hat{z}_d$. In the case of picture perception, in order to analyze
20 this phenomenon, we need to generalize equation [3] to the case where texture and disparity stem from different depth
21 values:

$$\hat{z}_c = \sqrt{k_t^2 z_t^2 + k_d^2 z_d^2}.$$

22 When a picture is viewed monocularly, the disparity signal is absent and therefore $k_d = 0$. When the image is viewed
23 with two eyes, depth from disparity is zero ($z_d = 0$). In both cases the second term under the square root vanishes and the
24 depth estimate is solely determined by the texture signal: $\hat{z}_c = k_t z_t$. Going back to superadditivity, note that we are not
25 claiming that focus cues were negligible in those displays, but that they were uninfluential in both the single-cue and two-
26 cue cases. This follows from the fact that focus cues specify the flat surface of the screen and therefore zero depth, thus
27 adding a zero term under the square root. In summary, both the phenomenon of superadditivity and the unaltered depth
28 perception deriving from viewing a picture with two eyes (Vishwanath, 2014) are successfully accounted for by the IC
29 model.

30 The fact that our previous results obtained in VR were successfully replicated using physical objects supports the
31 hypothesis that superadditivity is an integral feature of depth cue combination, capable to affect depth estimation across

1 multiple layers of cognition. Furthermore, this problem taps into a more general discussion over how VR technology can
 2 be utilized as a reliable tool for research on motor control, which has received increasing attention (Freud et al., 2018; Ozana
 3 et al., 2018). The results from this study support the idea that virtual environments have the potential to be as effective as
 4 more traditional scenes, where participants see and touch real objects.

5

6 Appendix

7

8 *Proof 1: The IC model equation maximizes the Signal-to-Noise-Ratio.*

9 Let's consider only two signals, for simplicity, $s_1 = \lambda_1 z$ and $s_2 = \lambda_2 z$. λ_i are unknown multipliers depending on
 10 confounding variables and z is the magnitude of the 3D property in question (e.g., front-to-back depth of an object). We
 11 seek an estimate $\hat{z}_C = f(s_1, s_2)$ that is (1) proportional to z and (2) most sensitive to 3D information while least sensitive
 12 to random fluctuations ε_i of λ_i . If λ_{i0} is the unperturbed value of λ_i : $\lambda_i = \lambda_{i0} + \varepsilon_i$ and $s_{i0} = \lambda_{i0}z$. We assume small random
 13 perturbations due to changes in viewing conditions such that ε_i are Gaussian distributions with zero mean and standard
 14 deviations σ_i . Taking the derivative of $\frac{df(s_1, s_2)}{dz} = \frac{df}{ds_1}(\lambda_{10} + \varepsilon_1) + \frac{df}{ds_2}(\lambda_{20} + \varepsilon_2)$, where $\frac{df}{ds_i}$ are calculated at s_{i0} , we
 15 observe a signal term $S = f_1\lambda_{10} + f_2\lambda_{20}$ (where $f_i = \frac{df}{ds_i}$) and a noise term $E = f_1\varepsilon_1 + f_2\varepsilon_2$ having standard deviation $\sigma_E =$
 16 $\sqrt{f_1^2\sigma_1^2 + f_2^2\sigma_2^2}$. If we minimize the Noise to Signal Ratio $NSR = \frac{\sigma_E}{S}$ with respect to f_i (by solving for f_i the equation
 17 $\frac{dNSR}{df_i} = 0$) we find that the first derivatives of the sought function are $\frac{df}{ds_i} \propto \frac{\lambda_{i0}}{\sigma_i^2}$. It can be easily shown that the derivatives
 18 $\frac{d\hat{z}_C}{ds_i}$ of the equation $\hat{z}_C = \beta \sqrt{\left(\frac{s_1}{\sigma_1}\right)^2 + \left(\frac{s_2}{\sigma_2}\right)^2}$ (calculated at s_{i0}) meet this requirement. By substituting $k_i = \beta \frac{\lambda_i}{\sigma_i}$ we obtain
 19 the IC equation $\hat{z}_C = \sqrt{(k_1 z)^2 + (k_2 z)^2}$ (easily generalizable to n signals).

20

21 *Proof 2: Derivation of \hat{z}_C for real stimuli*

22 In this study, when the stimuli were viewed under the Disparity and the Motion conditions, the total 3D shape
 23 information available resulted from the combination of each of those cues plus the focus cues. As a result, by applying
 24 equation [2], the estimated depth in the Disparity condition can be defined as:

$$25 \quad \hat{z}_{DF} = \sqrt{(k_D z)^2 + (k_F z)^2} = \sqrt{k_D^2 + k_F^2} z.$$

26 Similarly, the estimated depth in the Motion condition can be defined as:

$$27 \quad \hat{z}_{MF} = \sqrt{(k_M z)^2 + (k_F z)^2} = \sqrt{k_M^2 + k_F^2} z.$$

28 Following equation [2], in order to calculate the depth estimate in the Combined condition (\hat{z}_C) from the three sources
 29 of depth (disparity, motion and focus cues), the depth estimate from the Focus condition must be first subtracted from the

1 depth estimates obtained in the Disparity and the Motion conditions. Equation [3] can be thus derived from equation [2]
2 following few simple steps:

3
$$\hat{z}_C = \sqrt{[(\hat{z}_{DF})^2 - (\hat{z}_F)^2] + [(\hat{z}_{MF})^2 - (\hat{z}_F)^2] + (\hat{z}_F)^2} = \sqrt{k_{DF}^2 + k_{MF}^2 - 2k_F^2 + k_F^2} z = \sqrt{k_{DF}^2 + k_{MF}^2 - k_F^2} z.$$

4

5 References

6 Bozzacchi, C., & Domini, F. (2015). Lack of depth constancy for grasping movements in both virtual and real environments.
7 *Journal of neurophysiology*, 114(4), 2242-2248.

8 Bozzacchi, C., Volcic, R., & Domini, F. (2014). Effect of visual and haptic feedback on grasping movements. *Journal of*
9 *neurophysiology*, 112(12), 3189-3196.

10 Bozzacchi, C., Volcic, R., & Domini, F. (2016). Grasping in absence of feedback: systematic biases endure extensive
11 training. *Experimental brain research*, 234(1), 255-265.

12 Buckley, D. Frisby, J. P. (1993). Interaction of stereo, texture and outline cues in the shape perception of three-dimensional
13 ridges. *Vision Research*, 33, 919–933.

14 Campagnoli, C., & Domini, F. (2016). Conscious perception and grasping rely on a shared depth encoding. *Journal of*
15 *Vision*, 16(12), 449-449.

16 Campagnoli, C., & Domini, F. (2019). Does depth-cue combination yield identical biases in perception and
17 grasping?. *Journal of Experimental Psychology: Human Perception and Performance*, 45(5), 659.

18 Campagnoli, C., Cesanek, E., & Domini, F. (2015). Haptic feedback overrides visual size information during repeated
19 grasping. *Journal of vision*, 15(12), 1149-1149.

20 Campagnoli, C., Croom, S., & Domini, F. (2017). Stereovision for action reflects our perceptual experience of distance and
21 depth. *Journal of Vision*, 17(9), 21-21.

22 Campagnoli, C., Volcic, R., & Domini, F. (2012). The same object and at least three different grip apertures. *Journal of*
23 *Vision*, 12(9), 429-429.

24 Cesanek, E., & Domini, F. (2017). Error correction and spatial generalization in human grasp
25 control. *Neuropsychologia*, 106, 112-122.

26 Cesanek, E., Campagnoli, C., Taylor, J. A., & Domini, F. (2018). Does visuomotor adaptation contribute to illusion-resistant
27 grasping?. *Psychonomic bulletin & review*, 1-19.

28 Chessa, M., Maiello, G., Klein, L. K., Paulun, V. C., & Solari, F. (2019, March). Grasping objects in immersive Virtual
29 Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 1749-1754). IEEE.

30 Di Luca, M., Domini, F., & Caudek, C. (2010). Inconsistency of perceived 3D shape. *Vision research*, 50(16), 1519-1531.

31 Domini, F., & Caudek, C. (2003). 3-D structure perceived from dynamic information: A new theory. *Trends in Cognitive*
32 *Sciences*, 7(10), 444-449.

1 Domini, F., & Caudek, C. (2009). The intrinsic constraint model and Fechnerian sensory scaling. *Journal of Vision*, 9(2),
2 25-25.

3 Domini, F., & Caudek, C. (2010). Matching perceived depth from disparity and from velocity: Modeling and
4 psychophysics. *Acta psychologica*, 133(1), 81-89.

5 Domini, F., & Caudek, C. (2011). Combining image signals before three-dimensional reconstruction: The intrinsic
6 constraint model of cue integration. In *Sensory cue integration* (pp. 120-143). Oxford University Press.

7 Domini, F., & Caudek, C. (2013). Perception and action without veridical metric reconstruction: an affine approach.
8 In *Shape Perception in Human and Computer Vision* (pp. 285-298). Springer, London.

9 Domini, F., & Vishwanath, D. (2020). 3-D Cue Integration, Models of. *Encyclopedia of Computational Neuroscience*.

10 Domini, F., Caudek, C., & Tassinari, H. (2006). Stereo and motion information are not independently processed by the
11 visual system. *Vision research*, 46(11), 1707-1723.

12 Domini, F., Shah, R., & Caudek, C. (2011). Do we perceive a flattened world on the monitor screen?. *Acta psychologica*,
13 138(3), 359-366.

14 Foster, R., Fantoni, C., Caudek, C., & Domini, F. (2011). Integration of disparity and velocity information for haptic and
15 perceptual judgments of object depth. *Acta psychologica*, 136(3), 300-310.

16 Freud, E., Macdonald, S. N., Chen, J., Quinlan, D. J., Goodale, M. A., & Culham, J. C. (2018). Getting a grip on reality:
17 Grasping movements directed to real objects and images rely on dissociable neural representations. *Cortex*, 98, 34-48.

18 Frisby, J. P. Buckley, D. Wishart, K. A. Porrill, J. Garding, J. Mayhew, J. E. (1995). Interaction of stereo and texture cues
19 in the perception of 3-dimensional steps. *Vision Research*, 35, 1463-1472.

20 Glennerster, A., Rogers, B. J., & Bradshaw, M. F. (1996). Stereoscopic depth constancy depends on the subject's task. *Vision
21 research*, 36(21), 3441-3456.

22 Harris, J., Nefs, H., & Grafton, C. (2008). Binocular vision and motion-in-depth. *Spatial vision*, 21(6), 531-547.

23 Hecht, H., Van Doorn, A., & Koenderink, J. J. (1999). Compression of visual space in natural scenes and in their
24 photographic counterparts. *Perception & Psychophysics*, 61(7), 1269-1286.

25 Hoffman, D. M., Girshick, A. R., Akeley, K., & Banks, M. S. (2008). Vergence-accommodation conflicts hinder visual
26 performance and cause visual fatigue. *Journal of vision*, 8(3), 33-33.

27 Kemp, J., Cesanek, E., & Domini, F. (2018). Investigating biases in 3D perception and the effects of signal noise on depth
28 discrimination. *Poster presented at the European Conference on Visual Perception, Trieste, Italy*.

29 Ozana, A., Berman, S., & Ganel, T. (2018). Grasping trajectories in a virtual environment adhere to Weber's
30 law. *Experimental brain research*, 236(6), 1775-1787.

31 R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing,
32 Vienna, Austria. URL <https://www.R-project.org/>.

33 Tassinari, H., Domini, F., & Caudek, C. (2008). The intrinsic constraint model for stereo-motion
34 integration. *Perception*, 37(1), 79-95.

1 Tittle, J. S., Todd, J. T., Perotti, V. J., & Norman, J. F. (1995). Systematic distortion of perceived three-dimensional structure
2 from motion and binocular stereopsis. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3),
3 663.

4 Todd, J. T., & Norman, J. F. (2003). The visual perception of 3-D shape from multiple cues: Are observers capable of
5 perceiving metric structure?. *Perception & psychophysics*, 65(1), 31-47.

6 Van Ee, R., & Schor, C. M. (2000). Unconstrained stereoscopic matching of lines. *Vision research*, 40(2), 151-162.

7 Vishwanath, D. (2014). Toward a new theory of stereopsis. *Psychological review*, 121(2), 151.

8 Vishwanath, D., & Hibbard, P. B. (2013). Seeing in 3-D with just one eye: Stereopsis without binocular
9 vision. *Psychological science*, 24(9), 1673-1685.

10 Volcic, R., & Domini, F. (2014). The visibility of contact points influences grasping movements. *Experimental brain
research*, 232(9), 2997-3005.

11 Volcic, R., & Domini, F. (2016). On-line visual control of grasping movements. *Experimental brain research*, 234(8), 2165-
12 2177.

13 Volcic, R., Fantoni, C., Caudek, C., Assad, J. A., & Domini, F. (2013). Visuomotor adaptation changes stereoscopic depth
14 perception and tactile discrimination. *The Journal of Neuroscience*, 33(43), 17081-17088.

15 Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision*, 5(10),
16 7-7.

17 Willemsen, P., Gooch, A. A., Thompson, W. B., & Creem-Regehr, S. H. (2008). Effects of stereo viewing conditions on
18 distance perception in virtual environments. *Presence: Teleoperators and Virtual Environments*, 17(1), 91-101.

20