Acta Psychologica 136 (2011) 300–310

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Integration of disparity and velocity information for haptic and perceptual judgments of object depth

Rachel Foster [a,b], Carlo Fantoni [b], Corrado Caudek [b,c], Fulvio Domini [a,b,*]

[a] Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI, USA
[b] Center for Neurosciences and Cognitive Systems, Italian Institute of Technology, Rovereto, Italy
[c] Department of Psychology, Università degli Studi di Firenze, Firenze, Italy

ARTICLE INFO

Article history:
Received 16 March 2010
Received in revised form 9 December 2010
Accepted 10 December 2010
Available online 14 January 2011

PsycINFO codes:
2330 Motor Processes
2323 Visual Perception

Keywords:
Prehension
Motor control
Somatosensory
Haptic
Visual feedback
Shape perception
Cue integration

ABSTRACT

Do reach-to-grasp (prehension) movements require a *metric representation* of three-dimensional (3D) layouts and objects? We propose a model relying only on *direct* sensory information to account for the planning and execution of prehension movements in the absence of haptic feedback and when the hand is not visible. In the present investigation, we isolate relative motion and binocular disparity information from other depth cues and we study their efficacy for reach-to-grasp movements and visual judgments. We show that (i) the amplitude of the grasp increases when relative motion is added to binocular disparity information, even if depth from disparity information is already veridical, and (ii) similar distortions of derived depth are found for haptic tasks and perceptual judgments. With a quantitative test, we demonstrate that our results are consistent with the *Intrinsic Constraint* model and do not require 3D metric inferences (Domini, Caudek, & Tassinari, 2006). By contrast, the linear cue integration model (Landy, Maloney, Johnston, & Young, 1995) cannot explain the present results, even if the flatness cues are taken into account.

1. Introduction

It is commonly believed that visually guided behavior relies on a three-dimensional (3D) *metric* representation of the environment and the objects in it (Glover, 2004; Greenwald & Knill, 2009). It is also believed that this 3D depth map is found by reversing the physics of image generation to infer the outside world from sensory data (Helmholtz, 1867/1962; Landy, Maloney, Johnston, & Young, 1995; Landy, Banks, & Knill, in press; Poggio, Torre, & Koch, 1985). The solution of the so-called "inverse-optics" problem by a biological system, however, is extremely difficult because of the underdetermination of the required information. Horizontal binocular disparities, for instance, are not sufficient to recover an object's depth unless the viewing distance is known (Mayhew & Longuet-Higgins, 1982; Fantoni, 2008). Similarly, optic flow is not sufficient to recover surface slant unless additional parameters are known (i.e., the angular displacement between the observer and the surface and the amount of surface rotation) — see Fantoni, Caudek, and Domini (2010). Moreover, even sufficient constraints provided by multiple cues do not guarantee unique percepts (Todd, 2004).

For these reasons, some researchers have questioned the assumption that visuomotor processes rely on metric representations of target distances. Instead, they have hypothesized that (1) the brain relies mainly on image measurements that specify 3D properties directly, without building an explicit metric representation of the environment, and (2) appropriate body–environment interactions emerge as a consequence of adaptive mechanisms, not as the solution of the "inverse-optics" problem (Braunstein, 1994; Domini & Caudek, 2003; Robert, Zeller, Faugeras, & Hébert, 1997; Thaler & Goodale, 2010; Todd, 2004). In prehension movements aimed at reaching and grasping visual objects, for instance, the (haptic and/or visual) feedback resulting from the contact between the hand and the target provides an error signal for calibration that improves the accuracy of subsequent reaches (e.g., Mon-Williams & Bingham, 2007). Thus, visuomotor actions (such as prehension and pointing) may not require the recovery of the full 3D metric depth map, but instead be based on simpler mechanisms of conditional associative learning. If this is true, we should expect that perceptual metric judgments and motor actions in novel stimulus situations *with no haptic feedback* should be systematically distorted, which indeed has been found to be the case (e.g., Cuijpers, Brenner, & Smeets, 2008).

* Corresponding author. Tel.: +1 401 863 1356.
E-mail addresses: rachel.mary.foster@gmail.com (R. Foster), carlo.fantoni@iit.it (C. Fantoni), corrado.caudek@unifi.it (C. Caudek), fulvio.domini@brown.edu (F. Domini).

0001-6918/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.actpsy.2010.12.003

In the current investigation, we carried out a *cue combination experiment* in which human performance was measured in three stimulus conditions: with disparity-only information, motion-only information, or both (see also Tittle, Norman, Perotti, & Phillips, 1998). In different blocks of trials, participants either performed a grasping task or provided a perceptual judgment.

Two models of cue integration are considered here. In the first model, image measurements, diagnostic of 3D depth, but insufficient for metric reconstruction, are utilized (*intrinsic constraints*). The second model, instead, is based on the assumption that the brain uses metric structure (i.e., distance and direction) to represent locations (*linear cue integration*). In the next sections, we will describe the two models and show how it is possible to empirically validate their predictions by using the results of the present experiments.

2. Intrinsic constraints

The *intrinsic constraint* (IC) model proposes that, rather than deriving the full *metric* depth map, it is more advantageous for an organism to derive the best estimate of the *local affine* structure and use haptic feedback to calibrate ordinally scaled distance estimates (Di Luca, Domini, & Caudek, 2007; Domini & Caudek, 2010; Domini & Caudek, in press; Domini, Caudek, & Tassinari, 2006; Tassinari, Domini, & Caudek, 2008; see also Bingham & Pagano, 1998; Thaler & Goodale, 2010).

Retinal signals like relative disparity d are direct measures of the local affine structure, because $d \propto z$, where z is the depth map. The precision of the estimate of the affine structure is given by the signal-to-noise ratio (SNR) d/σ_d. We have shown that the best estimate of the affine structure is found through a linear combination of the retinal signals (not the depth estimates recovered from the signals) that maximizes the “information content” of the combined signal (i.e., the SNR — see also MacKenzie, Murray, & Wilcox, 2008). Once retinal signals are combined through this optimal combination rule, they determine a composite signal that encodes the affine structure, but with better precision (i.e., larger SNR) than either would have in isolation. This composite signal has been termed ρ. We propose that visually guided behavior depends upon this combined signal, which is scaled through calibration and perceptual learning from haptic feedback.

In the absence of haptic feedback, we also hypothesize that both perceptual judgments and motor actions are a monotone function of ρ (Domini et al., 2006). Consequently, we expect both perceptual judgments and motor actions to be systematically distorted, because unbiased estimations of 3D properties and target locations cannot be derived from ρ.

2.1. Disparity and motion integration

In the present investigation, we study the integration of disparity and motion information for both a motor task and a perceptual judgment. In both cases, according to IC, in the absence of haptic feedback, the amount of recovered depth z' should be a monotone function of the combined signal ρ:

$$z' = f_\rho(\rho). \tag{1}$$

Domini et al. (2006) showed that ρ is equal to the scores of the first principal component computed from the standardized retinal signals. Consequently, flatness cues are disregarded.

When only one signal is present, ρ is equal to the standardized value of that signal. For disparity-only stimuli, therefore, $\rho_d = \frac{d}{\sigma_d}$, where d is the relative disparity and σ_d is the measurement noise. The amount of depth recovered from disparity is

$$z'_d = f_\rho(\rho_d). \tag{2}$$

For motion-only stimuli, $\rho_v = \frac{v}{\sigma_v}$, where v is the relative velocity and σ_v is the measurement noise. The amount of depth that is recovered from motion information is given by

$$z'_v = f_\rho(\rho_v). \tag{3}$$

When both cues are present, we have that

$$\rho_c = \sqrt{\frac{v^2}{\sigma_v^2} + \frac{d^2}{\sigma_d^2}} \tag{4}$$

and

$$z'_c = f_\rho(\rho_c). \tag{5}$$

If we assume that the function $f_\rho(\rho)$ is linear for the range of depth magnitudes used in the present experiment, then

$$f_\rho(\rho) \approx a_\rho + k_\rho \rho. \tag{6}$$

Therefore,

$$z'_d = a_\rho + k_\rho \frac{d}{\sigma_d}, \tag{7}$$

$$z'_v = a_\rho + k_\rho \frac{v}{\sigma_v}, \tag{8}$$

$$z'_c = a_\rho + k_\rho \rho_c. \tag{9}$$

Considering that

$$d \approx \frac{IOD}{z_f^2} z = k_d z, \tag{10}$$

$$v \approx \frac{\omega}{z_f^2} z = k_v z, \tag{11}$$

where z is the distal relative depth, IOD is the observer's interocular distance, z_f is the fixation distance, and ω is the object's rotation velocity about a vertical axis, it follows that

$$z'_d = a_\rho + k_\rho \frac{k_d}{\sigma_d} z = a_\rho + K_d z, \tag{12}$$

$$z'_v = a_\rho + k_\rho \frac{k_v}{\sigma_v} z = a_\rho + K_v z, \tag{13}$$

where $K_d = k_\rho \frac{k_d}{\sigma_d}$ and $K_v = k_\rho \frac{k_v}{\sigma_v}$. For the disparity–motion (combined) condition, we can thus write

$$\begin{aligned} z'_c &= a_\rho + k_\rho \sqrt{\frac{v^2}{\sigma_v^2} + \frac{d^2}{\sigma_d^2}} z \\ &= a_\rho + k_\rho \sqrt{\frac{k_v^2 z^2}{\sigma_v^2} + \frac{k_d^2 z^2}{\sigma_d^2}} \\ &= a_\rho + z \sqrt{\frac{k_\rho^2 k_v^2}{\sigma_v^2} + \frac{k_\rho^2 k_d^2}{\sigma_d^2}} \\ &= a_\rho + z \sqrt{K_v^2 + K_d^2}. \end{aligned} \tag{14}$$

Eq. (14), therefore, provides a criterion for testing the IC model. K_d and K_v are the slopes of the linear functions relating recovered and distal depth magnitudes for the disparity-only and motion-only conditions, respectively. If the IC model is consistent with human performance, then the slope (K_c) of this linear relation in the disparity–motion condition must be equal to $\sqrt{K_v^2 + K_d^2}$. In the absence of haptic feedback, we expect that participants will perform in a similar manner in both action and perceptual tasks.

In conclusion, the recovered depth magnitudes increase when motion is added to disparity information because the SNR of the combined signal is larger. According to IC, this increase in the amount of recovered depth with the addition of cues does not have an upper bound corresponding to veridical performance.

Note that, for simplicity, we have denoted the output of the model as the "derived depth" z'. This does not mean, however, that IC provides a *metric reconstruction* of the full Euclidean space. Instead, IC performs a *local analysis* and produces an output that is equal to the SNR of the combined retinal signals. If this output is interpreted in terms of "depth" and these "depth" estimates were integrated over the visual scene, then they would be *internally inconsistent*. As a consequence, IC does not produce a *Euclidean metric representation* of 3D layout and object depth. The IC model is *local* and *non-metric*; its purpose is to provide a *direct* account of human performance on the basis of retinal information alone, not to recover a faithful Euclidean representation of distances and locations from sensory data.

3. Linear cue integration

The linear cue integration model assumes that our brain represents distances and locations in a metric format and that this metric representation is used to generate various kinds of responses (Ernst & Banks, 2002; Greenwald & Knill, 2009; Landy et al., in press). According to this approach, in order to obtain a metric representation, the human brain integrates information from multiple sources in order to reduce the uncertainty associated with any one of the available depth cues. If unbiased estimates of depth can be derived from each individual cue, then an unbiased estimate with *minimum variance* can be determined by a weighted, linear combination in which the weights are inversely proportional to the variances of the corresponding cues (Cochran, 1937). This combination rule also satisfies other statistical criteria of optimality: it is the maximum likelihood estimator and also the MAP estimator (Yuille and Bülthoff, 1996).

Note, however, that the *minimum-variance rule of combination* is only meaningful when the depth estimates obtained from the individual cues are *unbiased* (Helbig & Ernst, 2007; Hillis, Watt, Landy, & Banks, 2004; Oruc, Maloney, & Landy, 2003). Indeed, if the depth estimates obtained from the single cues were biased, it would make sense to minimize the bias of the final estimate of the depth, not its variance (for a discussion, see Domini & Caudek, in press).

3.1. Disparity and motion integration

Here we show how the linear combination model can be used to account for our observers' behavior in the three stimulus conditions (disparity-only, motion-only, and disparity–motion). In the case of computer-generated stimuli, it is necessary to incorporate a prior for frontoparallel and/or residual flatness cues (Watt, Banks, Ernst, & Zumer, 2002). Cues to flatness, arising from stimulus presentation on a flat monitor, can be modelled as a normal random variable with zero expectation and standard deviation σ_f (Adams & Mamassian, 2004). The binocular disparity likelihood z'_d is modeled as a Gaussian random variable centered at the true depth z with a standard deviation σ_d: $z'_d \sim \mathcal{N}(z, \sigma_d)$. By combining disparity information with the prior for flatness, we obtain

$$z'_{df} = w_s z'_d + w_f z'_f. \qquad (15)$$

The expected value of z'_{df} is

$$\mathcal{E}\left(z'_{df}\right) = w_s \mathcal{E}\left(z'_d\right) + w_f 0, \qquad (16)$$

where $w_s = \frac{\sigma_f^2}{\sigma_d^2 + \sigma_f^2}$. Eq. (16) can be rewritten as

$$\begin{aligned}\mathcal{E}\left(z'_{df}\right) &= \frac{\sigma_f^2}{\sigma_d^2 + \sigma_f^2} z \\ &= \frac{1}{1 + \frac{\sigma_d^2}{\sigma_f^2}} z \\ &= \frac{1}{1 + r_d^2} z,\end{aligned} \qquad (17)$$

where $r_d^2 = \frac{\sigma_d^2}{\sigma_f^2}$. In a similar manner, in the presence of residual flatness cues, for depth from motion, we can write

$$\mathcal{E}\left(z'_{vf}\right) = \frac{1}{1 + r_v^2} z, \qquad (18)$$

where $r_v^2 = \frac{\sigma_v^2}{\sigma_f^2}$ and σ_v is the spread of the motion likelihood. Also in this case, it is necessary to assume that $\mathcal{E}\left(z'_v\right) = z$. Finally, by combining disparity and motion information with a prior for frontoparallel, we obtain

$$\mathcal{E}\left(z'_{cf}\right) = \frac{1}{1 + r_c^2} z. \qquad (19)$$

Researchers who advocate linear combination argue that biases in 3D shape perception may be due to the flatness cues present in the stimulus displays (Watt et al., 2002). When the reliability of cues to flatness is not negligible, the depth from disparity-only, motion-only, and disparity–motion stimuli will be underestimated. Depth from the "disparity and motion" stimuli, however, should not be underestimated to the same extent as depth from disparity alone or motion alone. With two cues, in fact, there is more depth information available in the stimulus, and, thus, any priors to flatness and/or residual cues will have less influence (Adams & Mamassian, 2004). In fact,

$$\begin{aligned}r_c^2 = \frac{\sigma_c^2}{\sigma_f^2} &= \frac{1}{\sigma_f^2} \frac{\sigma_d^2 \sigma_v^2}{\sigma_d^2 + \sigma_v^2} (\text{a}) \\ &= \frac{r_d^2 r_v^2}{r_d^2 + r_v^2} (\text{b})\end{aligned} \qquad (20)$$

and $r_c^2 < r_d^2$ and $r_c^2 < r_v^2$.

In our investigation, we asked participants to judge the depth separation between two (virtual) rods by performing either a prehension movement or a perceptual task. In order to apply the linear combination model to our data, we need to establish the relation between the participants' response and the model's parameter z'. In our kinematic analysis of prehension movements, we focus on final grip aperture (FGA). Accordingly, we assume that the function FGA $\rightarrow z'$ is not an identity, but only a linear function:

$$\text{FGA} = a + b\ z'. \qquad (21)$$

This introduces two *free parameters* (a and b), thus increasing the possibility of the model to give a better fit to the empirical data. By taking into account Eq. (21), we can write Eqs. (17), (18), and (19) as follows:

$$\mathcal{E}\left(\text{FGA}_d\right) = a + \frac{b}{1 + r_d^2} z = a + K_d z, \qquad (22)$$

$$\mathcal{E}\left(\text{FGA}_v\right) = a + \frac{b}{1 + r_v^2} z = a + K_v z, \qquad (23)$$

$$\mathscr{E}\left(\mathrm{FGA}_c\right) = a + \frac{b}{1+r_c^2}z = a + K_c z, \tag{24}$$

where $\mathscr{E}\left(\mathrm{FGA}_d\right)$, $\mathscr{E}\left(\mathrm{FGA}_v\right)$ and $\mathscr{E}\left(\mathrm{FGA}_c\right)$ are the empirical estimates of the final grip aperture obtained in the disparity-only, motion-only, and disparity–motion conditions, respectively.

In our experiment, we obtained five empirical estimates $\mathscr{E}(\mathrm{FGA})$ for each stimulus condition. Thus, Eqs. (22), (23), and (24) define an underdetermined system of linear equations (i.e., a system containing fewer independent equations than unknowns) from which the coefficients a, K_d, K_v, and K_c can be estimated by means of a "least-squares" criterion. By using Eq. (20b), we can then solve the system of equations

$$r_d^2 = \frac{b}{K_d} - 1, \tag{25}$$

$$r_v^2 = \frac{b}{K_v} - 1, \tag{26}$$

$$r_c^2 = \frac{b}{K_c} - 1, \tag{27}$$

for b, r_d^2, and r_v^2.

In conclusion, the knowledge of $r_c^2 = \frac{\sigma_c^2}{\sigma_f^2}$, $r_d^2 = \frac{\sigma_d^2}{\sigma_f^2}$, and $r_v^2 = \frac{\sigma_v^2}{\sigma_f^2}$ allows us to test the biological plausibility of the linear integration model. In our experiment, in fact, the stimulus parameters and the viewing conditions were chosen in such a way that the cues to flatness were negligible. In these circumstances, therefore, the empirical estimates of r_c^2, r_d^2, and r_v^2 obtained from the data should be very small.[1]

3.2. Distinctions between the two models

The principal contrasting ideas of the IC and linear cue integration models include:

(1a) In the IC model, spatial structure – the affine spatial relations among the component dots, lines, etc. – is the fundamental information obtained from motion, disparity, and haptic properties; and the depth scale is a derived property, derived by some unspecified process.

(1b) The linear cue integration model assumes that depth is fundamental, and the spatial structure is derived from the set of depths.

(2a) The IC model assumes merely that the spatial information is affine (preserving ordinal depth).

(2b) The linear cue integration model uses the stronger assumption that the spatial information estimates metric (perhaps Euclidean) depth relations. A considerable amount of experimental evidence supports the former assumption (2a), even though the assumption about the primacy of depth persists in many parts of the literature. The primacy of depth is an assumption that is poorly supported by empirical evidence.

(3a) In the IC model, relative motion and disparity provide information about affine spatial structure; and this structure is what is combined from the motion and disparity cues. The "intrinsic constraint" is this affine spatial structure, which is the same for both motion and disparity properties, despite differences in resolution.

(3b) The linear cue integration model combines depths.

(4a) In the IC model, the resolutions of spatial relations add as independent variables. If the spatial resolution is denoted as ρ, then $\rho_c = \sqrt{\rho_v^2 + \rho_d^2}$, for the combined, motion, and disparity variables, respectively.

(4b) In the linear cue integration model, depth estimates from the motion and disparity depth cues are weighted by their resolutions: $z(c) = w_v z(v) + w_d z(d)$, where $w_v^2 + w_d^2 = 1.0$. These are qualitatively different predictions about the relative performance in the disparity, motion, and combined conditions.

4. Experiment

We asked the participants to perform two tasks: (1) to reach out to grasp a target object in the absence of haptic feedback, but with the stimulus always visible during the execution of reach-to-grasp movements (the hand was never visible), and (2) to performed a Manual Size Estimation (MSE) task: participants indicated the depth of the target object with index finger and thumb while holding their hand away from the target. MSE is interpreted as a measure of *perceptual* depth information in the visual system, in contrast to depth information used by the motor system in visually guided grasping (Franz, 2003). The target objects were defined by disparity-only information, motion-only information, or both.

Stimulus properties and viewing conditions were chosen to make the amount of perceived depth for the MSE task as close to veridical as possible in the disparity-only condition. This was done to best contrast the two cue integration models discussed above. Our goal was to test two hypotheses.

Hypothesis H1 states that human reach-to-grasp movements directed toward a virtual target will reveal the same distortions that have been found for perceptual judgments. Specifically, we expect to find an increase in the FGA when motion is added to disparity information, even if performance with disparity-only stimuli is already veridical (see Domini et al., 2006).

Hypothesis H2 states that both veridical performance and systematic errors in reach-to-grasp movements without haptic feedback, as well as the responses in the MSE task, can be accounted for by the IC model. A quantitative test of this hypothesis will be performed by using the derivations presented in the previous sections.

4.1. Participants

Five undergraduate students at the University of Parma as well as the first three authors participated in the experiment. All participants had normal or corrected-to-normal vision. Apart from the authors, all observers were naïve to the purposes of the experiment.

4.2. Apparatus

Each participant was seated in a darkened room in front of a high-quality, front-silvered 150×150 mm mirror (95% incident light reflected). The mirror was slanted at 45° relative to the observer's sagittal body mid-line and reflected the image displayed on a ViewSonic 9613, 19" CRT monitor (0.24 mm pitch) placed directly to the left of the mirror (see Fig. 1). The vertical position of the mirror was adjusted so that its center corresponded to the vertical center of the monitor (400 mm above the tabletop). The observer's eyes were 128 mm from the mirror. The distance from the mirror to the CRT was 242 mm. This produced the illusion that the image displayed

[1] Greenwald, Knill, and Saunders (2005) proposed an implementation of the linear combination model that is slightly more complex than discussed here — see also Greenwald and Knill (2009). They were interested in analyzing the time-varying orientations of subjects' fingers in a grasping task. To this purpose, a Kalman filter approach was used to update the internal estimate of a world's property (in their case, surface slant) by combining the estimates from different cues with the value predicted from its previous internal estimate. Consistent with what is discussed here, however, Greenwald et al. assumed that (1) the estimates obtained from the individual cues are unbiased, and (2) the single-cue estimates are combined through a weighted average with weights proportional to their reliabilities (and to the reliability of the previous internal estimate).

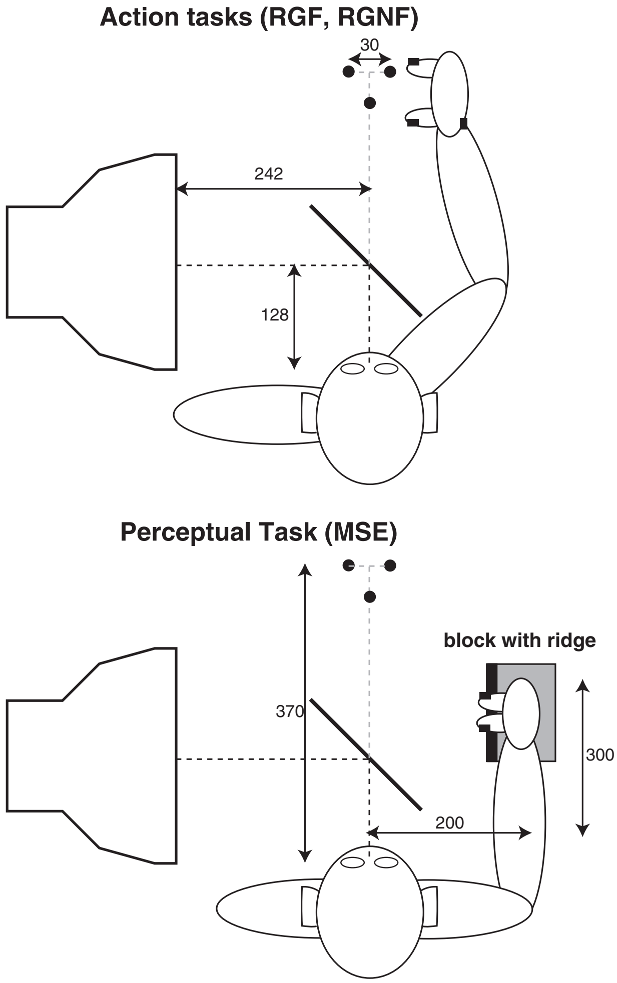

Fig. 1. Schematic of the bird-view of experimental apparatus. Top panel: participants reached behind a semi-silvered mirror to grasp a virtual object they could see in the mirror. In the trials with and without haptic feedback, the image reflected in the mirror was located 45 mm above or below the line of sight of the observer at rest, respectively. In the intermixed trials with haptic feedback in the RGF block, a physical object was present and perfectly aligned with the virtual stimulus. The physical object was never visible by itself. Bottom panel: resting their hand on a wooden platform, participants moved their fingers along a raised ridge to estimate the depth of the target object. The measures reported in the figure are expressed in mm.

on the screen was an object located at a distance of 370 mm behind the mirror. Eye-level was aligned with the center of the mirror.

Participants viewed the stimuli through liquid-crystal-diode LCD shutter glasses (FE-1 Goggles manufactured by Cambridge Research Systems) synchronized with the monitor so that, depending on the viewing condition, the shutter over the non-dominant eye was opened or closed electronically. This electronically driven shutter made it possible to randomly switch between viewing conditions during an experimental session without the observer noticing whether she was viewing the display monocularly or binocularly. The effect of using the glasses was that the effective CRT refresh rate was halved (60 Hz).

A physical object placed behind the mirror (wholly occluded from the observer) was used in the *Reach to Grasp with intermixed haptic Feedback* (RGF) block of trials. The physical object was perfectly aligned with the virtual stimulus (see the Stimuli section) and was made up of two metal rods (each 50 mm long) oriented vertically. One of the rods was aligned along the median axis of the observer's head at rest; the other rod was positioned 12.5 mm to the right of the center line. The central rod was 30 mm closer to the observer than the flanking rod, which was positioned 370 mm from the observer. The vertical midpoints of the rods were positioned 45 mm above the observer's cyclopean line of sight at rest.

The table allowed participants to reach comfortably behind the mirror. During the RGF, *Reach to Grasp with No haptic Feedback* (RGNF), and MSE blocks of trials, observer's right hand rested on a 65 mm high wooden block attached to the tabletop, which served as a starting point for subjects' prehensile movements. The block was shifted relative to the body of the observer at rest by about 200 mm along the coronal axis and 300 mm along the sagittal axis of the body at rest. A raised metal ridge was placed on the left side of this wooden block. Participants moved their fingers along this ridge when asked to perform the MSE task.

Kinematic markers of prehensile movement as well as the matching distance in the visual matching task were acquired on-line by using an Optotrak 3020 Certus system with two position sensors. The two position sensors recovered the signal (3D position data) from infrared-emitting diodes (IREDs) with sub-millimeter resolution. Position sensors were placed at an optimal distance and oriented so that the focus detection regions converged on the observer's rest position, which fell within the field-of-view of both sensors.

The positions of the fingers and the wrist during reach-to-grasp movements were recorded using six IREDs attached to a latex glove on the right hand (all participants were right-handed). Two IREDs were placed on the nail of the index finger, on the tip of the thumb, and on the center of the back of the hand.

4.3. Stimuli

According to the current literature, there are two causes for the misperception of stereoscopic displays (e.g., Held & Banks, 2008). First, the retinal images produced by the rendering display may not be the same as those produced by the original scene. Second, visual cues such as vergence and accommodation may be inconsistent with the information provided by the retinal images. Our stimuli were generated as follows so as to avoid both these problems: (1) disparity information was carefully calibrated by the observer's position and her/his interocular distance, (2) vergence information was always consistent with the simulated depth (Hoffman, Girshick, Akeley, & Banks, 2008), and (3) focus cues were negligible, given the size of the simulated objects (Banks, Akeley, Hoffman, & Girshick, 2008; Watt, Akeley, Ernst, & Banks, 2005). In these cue-consistent conditions, we expect stereoscopic depth constancy to be "essentially as much as with real surfaces" (Watt et al., 2005, p. 852).

Stimuli were presented and responses were recorded by means of a custom C++ program combined with Optrotrak API routines. The virtual stimuli were comprised of three dotted vertical lines embedded in a cloud of random dots (Fig. 2, top panel). These stimuli were arrangements of 800 high-luminance, anti-aliased red dots. Half of the dots coincided with three (invisible) vertical lines, each 50 mm long (see the three transparent cylinders in the top of Fig. 2). One of the three vertical lines was aligned with the vertical line passing through the center of the monitor's screen. The other two vertical lines were positioned 12.5 mm to either side of the center line. The simulated depth of the two lateral structures was equal to the fixation distance (370 mm from the observer at rest). The simulated depth of the central vertical structure varied from 360 mm to 320 mm in 10 mm steps, thus defining five relative depth magnitudes (10, 20, 30, 40 and 50 mm). In order to aid stereoscopic fusion and to increase the global amount of motion information, the remaining 400 dots were randomly placed within a volume 50 mm wide, 50 mm tall and 25 mm deep (see Domini et al., 2006). The volume was centered and aligned with the two lateral structures. The vertical midpoint of the whole stimulus configuration was positioned 45 mm below the center of the screen.

On 25% of the trials of the RGF block, the virtual stimuli were vertically displaced by 90 mm from the position of the no-haptic

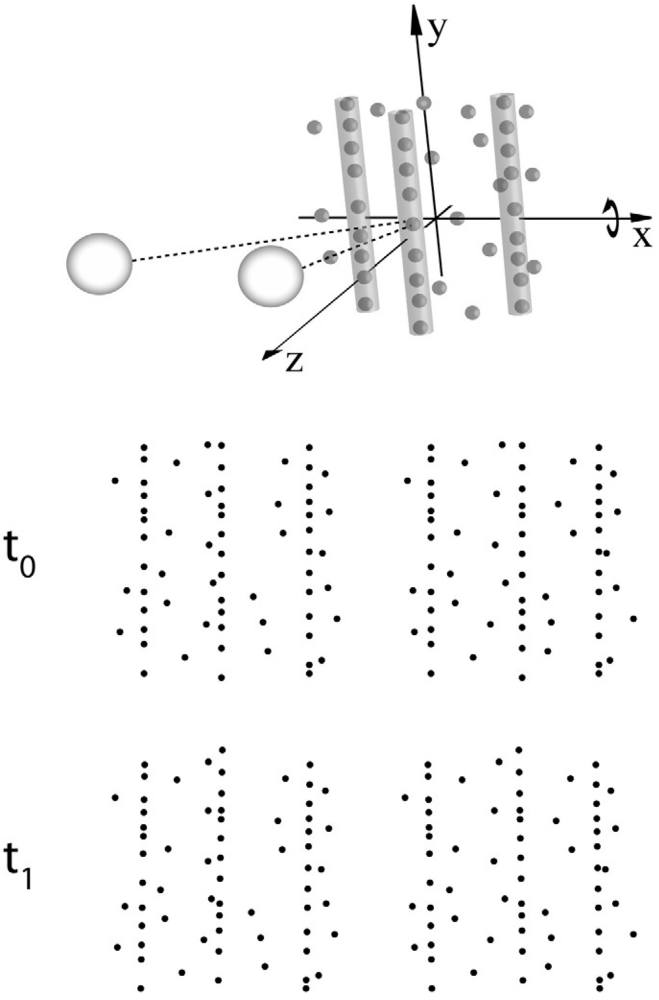

Fig. 2. (Top panel) Schematic representation of the stimulus and the viewing geometry used in the experiment. The stimulus displays and the point of view use a reference frame with the *xy*-plane co-planar with the monitor's screen, the *x*-axis pointing to the participant's right, the *y*-axis pointing upward, and the *z*-axis pointing toward the observer. The origin of the reference frame is set at the center of the monitor's screen. The three transparent cylinders highlight the three vertical dotted lines present in the stimulus displays. The central dotted line is centered on the *z*-axis and is closer to the observer. The flanking dotted lines are positioned at the screen depth level (*x*–*y* plane). The axis of rotation of the motion-only and disparity–motion stimuli coincides with the *x*-axis (arrow line). The point of view is located above the stimulus' center. (Bottom panels) Two stereograms representing a two frame sequence (t_0 and t_1) of a simplified version of a motion–disparity stimulus used in the experiment (cross-fuse).

feedback trials (45 mm above the screen's center), and the center and right lines were perfectly aligned with the two metal rods. This subset of stimuli was specified by disparity-only information simulating a relative depth of 30 mm, equal to the relative depth of the physical object located behind the mirror. These trials provided congruent haptic feedback. We used this condition to reduce response uncertainty in absence of haptic feedback (Bingham, Coats, & Mon-Williams, 2007).

Depth was specified by disparity-only, motion-only, or disparity–motion information. In the motion-only and disparity-motion conditions, the 2D motion of the dots was created by simulating the rigid rotation of the whole arrangement of 800 dots by 15° in either direction around a horizontal axis located in the center of the stimulus and at the same depth level of the screen. Each motion cycle lasted 60 frames (1 s). The stimulus was shown for 150 frames (2.5 s or two and a half full motion cycles). Each participant's IOD was measured and used to calculate horizontal disparities and the overall amount of motion parallax. Perspective projection was used by taking the observers' eyes as the center of projection.

In a departure from previous investigations on goal-directed prehension, we removed proprioceptive extra-retinal information resulting from observers' self-motion that was often available (Bingham, 2005; Bingham, Crowell, & Todd, 2004; Bingham & Pagano, 1998; Watt & Bradshaw, 2003). We achieved this by using motion-only and disparity-motion displays generated by a continuous back-and-forth rotation of the stimulus rather than by the motion of the observer. In this way, disparity and motion information were optimally isolated from any other cues that are normally available within natural viewing conditions.

4.4. Design

Each participant completed three blocks of trials: RGF, RGNF, and MSE. In all blocks, participants were shown 5 simulated depths magnitudes (10, 20, 30, 40, and 50 mm) specified by disparity-only, motion-only, or disparity–motion information.

In the 25% of the trials of the RGF block, a physical object coinciding with the simulated 3D structure on the monitor was felt by the observer when she correctly performed the reach-to-grasp movement. A post hoc analysis of our data revealed that just 0.5% of these reach to grasp movements were incorrectly performed. Incorrect reaches were subsequently eliminated from the analysis.

The RGF block comprised 200 trials, resulting from 50 presentations of the calibration stimulus plus 5 Depth magnitudes × 3 Cue conditions × 10 repetitions. The RGNF and MSE blocks consisted of 150 trials each, the calibration trials being absent.

4.5. Procedure

Participants were tested individually in total darkness, so that only the luminous dots on the simulated 3D object were visible. The observer's head position was stabilized by means of a chin-and-forehead locating apparatus. The chin-rest, parallel to the horizontal dimension of the monitor's reflected screen, was adjusted in height to position the participant's cyclopean eye at the screen's center.

Participants performed the haptic task in two blocks of trials. In one block, haptic feedback was provided on 25% of trials (RGF); in another block, haptic feedback was never provided (RGNF). In both blocks, the target was visible for the entire reach-to-grasp movement, while the hand was never visible. We used this hybrid closed-loop condition to remove the occlusion of the object by the hand and the relative disparity between the hand and the object that are present in the standard closed-loop condition (where the hand and the object are always visible) — see Bingham (2005), Bingham, Bradley, Bailey, and Vinner (2001), Mon-Williams and Tresilian (1999), Tresilian, Mon-Williams, and Kelly (1999), and Tresilian and Mon-Williams (2000).

For the RGF and RGNF blocks, participants were instructed to begin each trial with their hand on the wooden platform touching their index finger and thumb together. As soon as the stimulus appeared (signaled by a high pitched sound), participants were to make quick, accurate, and natural reaches so as to grasp the virtual or physical object front-to-back with their thumb and index finger. Observers held their fingers in position until the stimulus disappeared, at which time they were to return to the starting position. Trials advanced automatically without any direct input from the participants. The stimulus was replaced by a blank screen during the 2 s temporal asynchrony between trials.

In the RGNF block, participants were informed that the stimulus was a virtual object, but they were instructed to make a natural movement as if grasping a physical object. Participants were informed that, in some trials of the RGF block, the target would appear above rather than below their eye level and that furthermore, in these trials only, they would contact a physical object. Participants were allowed to practice with a virtual target and a physical object (20 trials).

In the MSE block, participants rested their hand on the wooden platform throughout the duration of the experiment. They were instructed to move their fingers along the raised ridge to estimate the depth of the object from tip to back when the display appeared and bring their fingers back together when it disappeared. Stimulus duration and temporal asynchrony between trials and acoustic signals were the same as in the RGF and RGNF blocks.

4.6. Results

4.6.1. Data analysis

Information provided by the IREDs placed on the participant's hand and fingers was used to describe the transport and the grasp components of the prehensile movements (Jeannerod, 1981) and the response in the MSE task. The 3D spatio/temporal coordinates of the IREDs placed on the participant's hand and fingers were used to compute the dependent variables considered as markers of reach-to-grasp kinematics: (*i*) the amplitude of the final grip aperture (FGA)[2]; (*ii*) the peak grip aperture, (*iii*) the amplitude of maximum wrist peak velocity, and (*iv*) the movement duration[3] (see Fig. 3).

4.7. Test for hypothesis H1

By using a perceptual task, Domini et al. (2006) found an overestimation of depth when motion was added to binocular disparity, even when disparity information alone yielded veridical performance (see also Domini & Caudek, in press). Here, we asked whether the same mis-estimation of depth occurs when participants use their hand to reach out and grasp an object. To prevent calibration of prehension movements via haptic feedback, we used a virtual target (Mon-Williams & Bingham, 2007; Mon-Williams, Coats, & Bingham, 2004); it is therefore natural to focus our attention on the FGA.[4]

Since grip sizes were initially measured from markers located on the nails of the finger and thumb, we calculated the real apertures (in mm) between the inner surfaces of the two digits at the FGA for each individual participant, in order to remove the effect of digit thickness. This was done by applying a correction to each participant's raw data, based on the size of their mean terminal grip aperture on the 30 mm physical object used in the RGF block. This individualized correction factor was subtracted from each FGA obtained from each participant. The average FGA in the different experimental conditions is shown in Fig. 4.

The results of the perceptual (MSE) task replicate those of Domini et al. (2006). The mean amount of perceived depth in the disparity-only condition was equal to 29.43 mm, 95% *C.I.* = [21.46, 36.52].[5] The amount of perceived depth in the motion-only condition was 3.48 mm smaller than that in the disparity-only condition, 95% *C.I.* = [−5.44, −1.56]. When both cues were present, perceived depth increased by 5.49 mm with respect to the disparity-only condition, 95% *C.I.* = [3.49, 7.39].

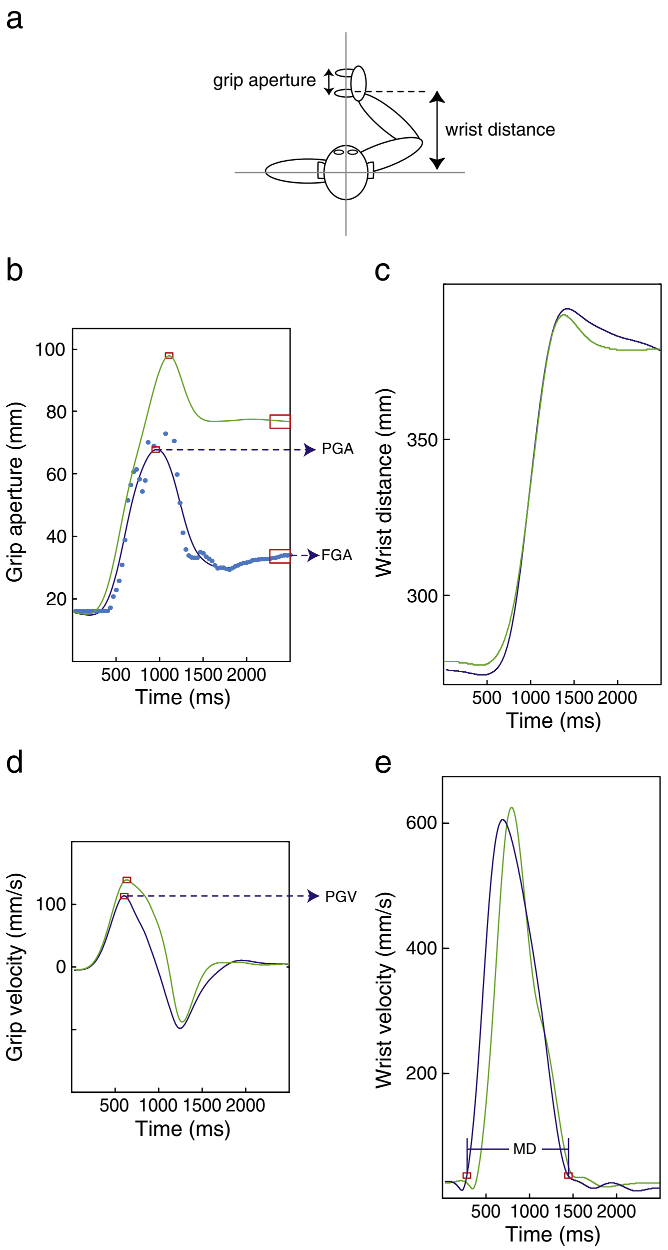

Fig. 3. Panel a provides a schematic representation of grip aperture and wrist distance. For the other panels, the blue and green colors code the spatial and temporal components of the reach to grasp movements for a 10 mm and 50 mm deep object, respectively. The figure shows the data of two trials for one representative observer. Panel b shows the time course of the grip aperture. The red insets highlight the portion of the trajectories used to compute the FGA and the Peak Grip Aperture (PGA). The blue dots represent the 75 measurements obtained for a 10 mm target. The continuous lines represent the cubic splines interpolation used to compute the kinematic markers employed in the analyses. Panel *c* represents the interpolated time course of the wrist distance. Panel *d* represents the velocity profiles of the grasp obtained by time-differentiating the interpolated trajectories of panel *b*. Panel *e* represents the velocity profiles of the wrist trajectory obtained by time-differentiating the trajectories of panel *c*. From the velocity profiles of the grip, we computed Peak Grip Velocity (PGV). From the velocity profile of the wrist, finally, we computed the Movement Duration (MD).

[2] FGA was calculated by averaging the distance between the markers on the index and the thumb during the last 33 ms of the interpolated temporal profile of the grip aperture.

[3] Movement duration measured the difference between onset and conclusion of the wrist movements. Tangential speed thresholds of 20 mm/s was used to mark the beginning and end points of the wrist movements, respectively (e.g., Loftus et al., 2004).

[4] Other spatial components of the grasp have been considered in the literature, such as peak grip aperture and peak grip velocity, and it has been shown that these different variables produce a very similar pattern of results (see Gentilucci, Benuzzi, Gangitano, & Grimaldi, 2001; Jakobson & Goodale, 1991; Jeannerod, 1984;1988; Servos, Goodale, & Jakobson, 1992). In the present study, the correlations between these three variables is equal to $r_{PGA,FGA} = 0.76$, $r_{PGA,PGV} = 0.78$, and $r_{PGV,FGA} = 0.51$.

[5] All the analyses were performed with Linear Mixed-Effects models with participants as random effects (Baayen, Davidson, & Bates, 2008). We evaluate significance by computing the deviance statistic (minus 2 times the log-likelihood; change in deviance is distributed as chi-square, with degrees of freedom equal to the number of parameters deleted from the model) and with the help of 10,000 samples from the posterior distributions of the coefficients using Markov chain Monte Carlo sampling. From these samples, we obtained the 95% Highest Posterior Density confidence intervals, and the corresponding two-tailed *p*-values.

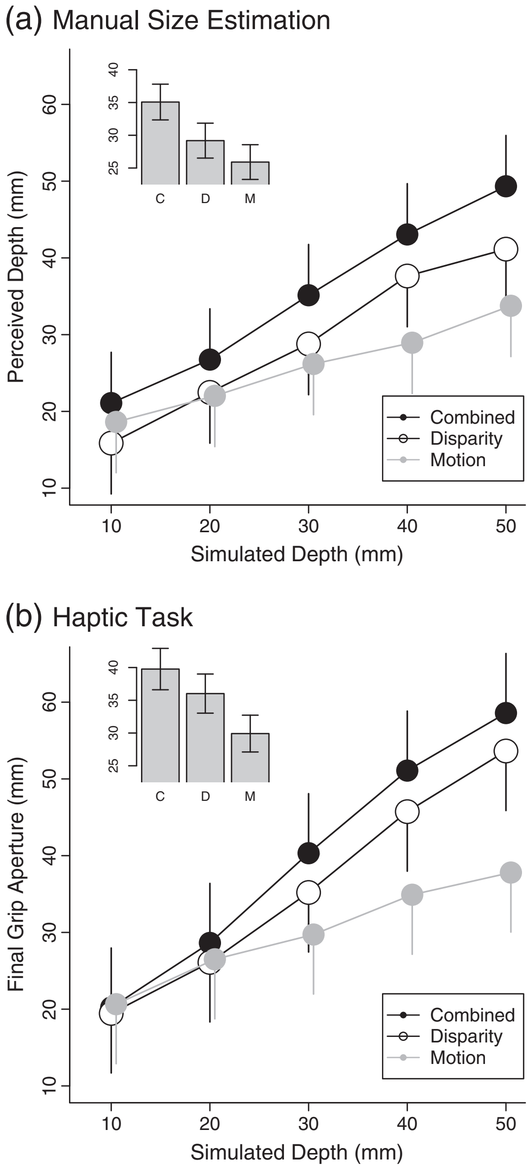

Fig. 4. Final grip aperture (FGA) as a function of simulated depth and cue. The left panel shows the results for the MSE task; the right panel shows the results for the haptic task averaged over the RGF and RGNF blocks. The insets represent the average responses as a function of the cue condition (C: combined; D: disparity; M: motion). Error bars represent one standard error of the mean.

Consistent with hypothesis H1, a similar pattern was found for the results of the haptic task. In the disparity-only condition, mean perceived depth was equal to 36.08 mm, 95% *C.I.* = [27.52, 44.63]. Perceived depth was on average 6.18 mm less in the motion-only condition, 95% *C.I.* = [−8.01, −4.27]. Importantly, perceived depth was 4.04 mm greater on average in the combined-cue condition, 95% *C.I.* = [2.20, 5.89]. This increase produced an overestimation of depth, *C.I.* = [31.85, 48.99].

There were, however, some differences between the haptic and visual tasks, as revealed by a significant three-way interaction Depth × Cue × Task, $\chi_4^2 = 9.96$, $p < 0.05$. Thus, we analyzed the MSE and the RGF and RGNF trials separately.

In the RGF and RGNF trials, the Depth × Cue × Block interaction was not significant, $\chi_2^2 = 1.48$, $p = 0.48$. We thus analyzed the RGF and RGNF trials together. At the simulated depth of 30 mm, the mean response for the disparity-only stimuli was veridical, 95% *C.I.* [28.80, 43.28]. However, depth magnitudes smaller than 30 mm were overestimated and depth magnitudes larger than 30 mm were underestimated. By centering the *x*-axis at 30 mm, the regression of perceived and simulated depth magnitudes had a slope of 0.90, *S.E.* = 0.04, $t_{2040} = 25.11$, $p < 0.001$. At 30 mm, the average response for the disparity–motion stimuli was 3.92 mm larger than for the disparity-only stimuli; by coding the data as we did before, the slope of the relationship between perceived and simulated depth magnitudes was 0.12 mm larger than for the disparity-only stimuli, *S.E.* = 0.05, $t_{2040} = 2.31$, $p < 0.05$. At 30 mm, the average response for the motion-only stimuli was 6.24 mm smaller than for the disparity-only stimuli; also, the slope of the relationship between perceived and simulated depth magnitudes was significantly shallower, with a decrease of 0.47, *S.E.* = 0.05, $p < 0.001$. A similar pattern of results was obtained for the MSE task, except that the slope of the linear regression of perceived and simulated depth magnitudes was not different across the disparity-only and disparity–motion stimuli, $t_{1101} = 0.99$.

For the simulated depth of 30 mm, we found no difference in the mean FGA across the blocks with and without haptic feedback, $t_{400} = 0.592$, $p = 0.55$. The variability of the FGA, however, was 51% larger in the block without haptic feedback, $F_{1,412} = 10.904$, $p < 0.002$.

The data of Fig. 4 seem to indicate an intercept different from zero for a linear fit. Even though is not appropriate to extrapolate the present results beyond the range of the simulated depth values that had actually been measured, the previous observation could be interpreted as an indication that the relation between simulated and perceived depth is not linear.

4.8. Test for hypothesis H2

The system of linear equations given by Eqs. (22), (23), and (24) can be written in matrix format as $\boldsymbol{Ax} = \boldsymbol{B}$, where $\boldsymbol{B}$ is the vector containing the empirical estimates $\mathcal{E}$(FGA) and $\boldsymbol{A}$ is the matrix containing the coefficients a, K_d, K_v, and K_c of equations Eqs. (22), (23), and (24). The empirical estimates of $\mathcal{E}$(FGA) were computed by averaging the trials of the RGF and RGNF blocks. The Moore–Penrose generalized inverse of matrix $\boldsymbol{A}$ produced the following estimates: $K_d = 0.71$, $K_v = 0.50$, and $K_c = 0.85$, and $a = 12.83$. By solving the system of Eqs. (25), (26), and (27), the following solutions were found: $r_d^2 = 0.88$, $r_v^2 = 1.69$, and $r_c^2 = 0.58$.

Remember that, in the present investigation, we minimize the cues to flatness by using a small display size and low point density, omitting texture cues (the stimuli are defined by a set of points arranged along three vertical lines), and providing consistent vergence and accommodation cues (Hoffman et al., 2008). In these circumstances, we should expect depth constancy similar to that found for real objects. In other words, the cues to flatness should have a negligible effect (Watt et al., 2005).

The r^2 values of Eqs. (25), (26), and (27) represent the ratios of the spread of the single- or combined-cue likelihoods and the spread of the likelihood of the cues to flatness. If the linear cue combination model were consistent with the data, we should expect the empirical estimates of r^2 in the three stimulus conditions to be very small, because in our stimulus conditions the reliability of the cues to flatness was extremely low compared to the reliability of disparity and motion information.

Contrary to this prediction, our data require the reliability of disparity information to be only 14% greater than the reliability of the cues to flatness, the reliability of motion information to be 40% lower than the reliability of the cues to flatness, and the reliability of disparity–motion information to be only 72% greater than the reliability of the cues to flatness. It is obvious that these estimates stand in contrast with the stimulus properties of the current investigation (Watt et al., 2005). We should also stress that the model from which the values r_d^2, r_v^2, and $r_c^2 =$ of Eqs. (25), (26), and (27) were estimated contains two free parameters derived ad hoc.

The unacceptable estimates of linear cue combination can be contrasted with the good fit of IC. In the introduction, we showed how the IC model can be tested by considering the linear relation relating the recovered and the simulated depth magnitudes in the different stimulus conditions. According to Eq. (14), the slope of this linear relation in the disparity–motion condition should be equal to $\sqrt{K_v^2 + K_d^2} = 0.87$. This value is very similar to the empirical estimation of K_c, which is 0.85.

4.9. Further analyses

4.9.1. Movement duration

The analysis of the transport component of the grasp revealed that the delay between the onset and the final hand position was significantly longer for motion-only than for disparity-only and the disparity–motion stimuli — see Fig. 5. On average, this delay was equal to 28 ms, 95% *C.I.* [8.27, 47.99], $p<0.005$. This result is consistent with the hypothesis that the removal of binocular disparity leads to greater uncertainty, resulting in longer duration movements (Loftus, Servos, Goodale, Mendarozqueta, & Mon-Williams, 2004; Melmoth & Grant, 2006).

4.9.2. Virtual and natural targets

When analyzing only the trials in the RGF block in which participants reached for physical and virtual targets that shared the same spatial configuration (30 mm depth, disparity-only condition), we found that the average FGA for a virtual target was not significantly different from that of a corresponding physical object: average FGA = 31.6 mm, *S.E.* = 1.3, 95% *C.I.* [28.6, 34.5]. Without haptic feedback, however, the movement was 15% slower, $t_{280} = 5.42$, $p<0.001$ (see Goodale, Jakobson, & Keillor, 1994).

4.9.3. Effect of intermixed haptic feedback on grasping precision

For the disparity-only, motion-only, and disparity–motion stimuli, the variability of the FGA was 28%, 54%, and 16% larger, respectively, in the trials with no intermixed feedback (Levene test: $F_{1,2044} = 30.97$, $p<0.001$). This result replicates the finding that haptic feedback resulting from contact with actual targets produces calibration and it allows reaches to become more precise (Bingham et al., 2007; Wickelgren, McConnell, & Bingham, 2000; see also Bingham, Zaal, Robin, & Shull, 2000; Bingham & Pagano, 1998; Pagano & Bingham, 1998), even though it does not necessarily correct shape distortions (Bingham et al., 2001). Interestingly, we found an increase in precision even if haptic feedback was provided only to a predictable subset of prehension movements directed to a different spatial target.[6]

5. General discussion

In the present investigation we study reach-to-grasp movements directed towards virtual stimuli. We test two hypotheses. H1: The same depth distortions are found in performance involving action (with no haptic feedback) and perceptual judgments. H2: The IC model can account for the limited effectiveness of disparity and motion information in conveying spatial information.

The results shown in Fig. 4 support H1: The addition of motion significantly increased the FGA, even if prehension movements were veridical for disparity-only stimuli. Prehension movements for virtual objects, therefore, reveal the same systematic distortions of depth that have been found for perceptual judgments (Domini et al., 2006). H2 was confirmed by the excellent fit of the IC model: participants' performance in the combined-cue condition can be predicted from their performance in the single-cue conditions.

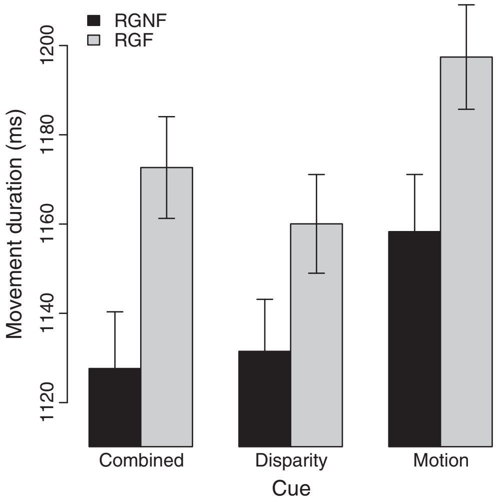

Fig. 5. Movement duration as a function of cue (combined, disparity, and motion) and block (RGNF: Reach to Grasp Without intermixed haptic Feedback trials; RGF: Reach to Grasp with intermixed haptic Feedback trials). Error bars represent one standard error of the mean.

The accurate predictions of IC contrast with the poor fit of the linear combination model. In our investigation, the stimuli and the viewing conditions were chosen so as to maximize depth constancy (e.g., Banks et al., 2008; Held & Banks, 2008; Hoffman et al., 2008; Watt et al., 2005). In fact, almost perfect depth constancy is found for both visual and haptic tasks in the disparity-only condition. In these circumstances, the effectiveness of the flatness cues is minimal. Nevertheless, substantial weights for the cues to flatness must be derived from the linear combination model in order to fit the data; indeed, the estimated weights of the flatness cues are around the size of the weights of the disparity and motion cues. These disproportionate estimates are found even though the linear combination model is provided with two additional free parameters to allow a better fit (see Eq. (21)).

Overall, the qualitative trend of the present results is inconsistent with the linear combination model. By assuming unbiased estimates from single cues, linear cue combination imposes an upper limit to the amount of recovered depth. The distortions of depth are attributed to the cue-conflict between the depth cues and the residual flatness cues and/or to a prior for fronto-parallel. Consequently, linear cue combination is consistent with depth *underestimation* and with an increase in precision as more cues are added to the stimulus display. However, it cannot explain depth *overestimation*, such as we found in the present case.

As discussed elsewhere, the linear combination model makes two assumptions that, in our opinion, are questionable: (1) the assumption that the interactions between the organism and the environment require an *Euclidean metric depth representation*, and (2) the assumption that *unbiased* metric estimates of depth can be derived from single cues (Domini & Caudek, in press). Contrary to the first assumption, it has been argued that the visuo-motor system plausibly uses of an affine representation of 3D space, accompanied with some form of learning-based scaling (e.g., Thaler & Goodale, 2010). Contrary to the second assumption, most of the literature suggests that the single-cue estimates of metric depth are biased (e.g., Todd, 2004). If this is the case, then there is no reason to combine the cues according to their reliability. In fact, the most reliable cue can also have the largest bias. Instead, an "optimal estimator" would require the assignment of the greatest weight to the

[6] According to IC, the amount of recovered depth is a function of the signal-to-noise ratio. In the condition with haptic feedback the variability of the responses is smaller. This does not imply, however, that the amount of recovered depth should be larger. In fact, we need to distinguish between the variance of the errors of measurement of the sensory signals and the variance of the responses due to the planning and execution of a motor movement. The haptic feedback certainly affect the second variance, but there is no reason why it should affect the first one. As a consequence, there is no reason to expect, according to IC, an increase of perceived depth when haptic feedback is provided.

least biased estimator. The problem is that the visual system has no way to determine the amount of bias — unless it relies on some form of learning based on haptic feedback.

It is also necessary to be cautious in interpreting the empirical findings in support to linear combination. The result that the variance of the responses in the combined-cue condition is smaller than in each of the single-cue conditions has been taken as the clearest evidence supporting linear combination (e.g., Hillis et al., 2004). This result is certainly compatible with linear combination, but we must also recognize that this relation among the variances is compatible with a whole class of models, very different with each other, and not only with linear cue combination. For example, Domini et al. (2006) demonstrated that the same relation among the (single-cue and combined-cue) variances holds true for IC as well, regardless to the fact that (*i*) IC is based on a completely different set of assumptions than linear cue combination, and (*ii*) IC makes different predictions about human performance than linear cue combination (see Section 3.2).

Previous studies have revealed that reach-to-grasp movements for virtual stimuli are slower and shortened with respect to the target location, and are characterized by a reduction in maximum grasp aperture (Goodale et al., 1994). The differences between pantomimed and normal prehension movements tend to disappear if participants are allowed to calibrate their reaches across an experimental session (Bingham et al., 2007). In the present study, the hand movements are also slower when haptic feedback is absent.[7] Moreover, the variability of the FGA is larger in the RGNF block than in the intermixed haptic feedback trials. These results suggest that intermixed haptic feedback is not sufficient to produce "normal" grasp movements, despite the fact that it increases the precision of performance. Note that, in contrast to Bingham et al. (2007), the feedback trials in our RGF block was predictable: in the feedback trials, the target was located in a different spatial position than in the other trials. Rather than proper calibration, therefore, our intermixed feedback trials provided a sort of "anchoring" which reduced the variability of the responses but preserved some differences between pantomimed and normal prehension movements.

The lack of haptic feedback is not a limitation in this present study, however. In the majority of previous investigations on reach-to-grasp movements, participants could *grasp, pick up, and manipulate real objects* (e.g., Bennett, Mucignat, Waterman, & Castiello, 1994). However, real objects are not particularly suited for cue-combination experiments, because they make it difficult to isolate specific sources of depth information (e.g., binocular disparity and motion parallax) from other cues. One example is provided by the study of Watt and Bradshaw (2003), where peak grip aperture scaled with object depth also in the static, monocular viewing condition. We chose to use virtual stimuli with no haptic feedback in order to study the effectiveness of motion and disparity information in absence of other cues.

The present results show that, when there is no haptic feedback and the hand is not visible, visual information and proprioceptive feedback can bring the effector in the "ballpark" of the target, but feedback is required for finer adjustments of the grasp movement. The good fit of the IC model suggests that the *planning component* of prehension movement and its execution integrated by proprioceptive feedback may only depend on a subset of visual information — that is, the retinal information that, by itself, does not allow a *metric* reconstruction of the 3D layout and of the object depth.

Finer adjustments are needed for effective prehension movements and they are made by *on-line control* (e.g., Bradshaw et al., 2004) and by calibration from haptic feedback (e.g., Bingham et al., 2000). Both these components of prehension movements were removed from the stimulus situation investigated in the present study. Consistent with the theoretical approach of the IC model, we propose that the on-line control component of prehension movements may be directly driven by retinal information as well (Bingham et al., 2001; Bradshaw, Parton, & Glennerster, 2000; Bradshaw et al., 2004; Jackson, Jones, Newport, & Pritchard, 1997; Melmoth & Grant, 2006; Mon-Williams & Dijkerman, 1999; Watt & Bradshaw, 2000;2003). A *disparity nulling* strategy, for instance, would only require "that the observer [be able to] determine that crossed (near) disparities between hand and target and that uncrossed (far) disparities between target and hand are reducing concurrently" (Melmoth, Storoni, Todd, Finlay, & Grant, 2007, p. 295). As such, a nulling process is not mediated by an "internal" metric representation (see also Bruggeman, Fantoni, Caudek, & Domini, 2010).

Haptic feedback is also important. How calibration deriving from haptic feedback can be integrated into depth and shape judgments within a non-metric model, however, is a question that still needs to be investigated. It should also be kept in mind that haptic feedback does not necessarily produce complete calibration (see also Phillips, Egan, & Perry, 2009; Van Doorn, Richardson, Wuillemin, & Symmons, 2010). It has been suggested that haptic information may be more effective in calibrating judgments of distance than judgments of shape (e.g., Brenner, van Damme, & Smeets, 1997; Coats, Bingham, & Mon-Williams, 2008; Wijntjes, Volcic, Pont, Koenderink, & Kappers, 2009). Moreover, Cuijpers et al. (2008) have reported that complete calibration for grasping of virtual objects is not obtained, even with consistent haptic feedback, when the shape and orientation of virtual objects change every other trial and when grasping involves judging higher-order shape parameters such as surface curvature (see also Bingham et al., 2007; Coats et al., 2008; Smeets & Brenner, 2010). Future research, therefore, should focus on better understanding the distortions of depth present in the planning and execution of reach-to-grasp movements when haptic feedback is provided.

6. Conclusions

In absence of haptic feedback, prehension movements for virtual targets (hand not visible) reveal the same distortions of 3D depth as perceptual judgments. Predictable and spatially displaced intermixed haptic feedback trials are not sufficient to calibrate no-feedback trials. Final grip aperture increases when motion is added to disparity information, even when this corresponds to an overestimation of depth. The present results are consistent with the IC model but not with linear cue combination.

Acknowledgment

We wish to thank Joe Lappin for his constructive comments on a previous version of this manuscript.

References

Adams, W., & Mamassian, P. (2004). Bayesian combination of ambiguous shape cues. *Journal of Vision*, *4*(10), 921–929, doi:10.1167/4.10.7.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59* (4), 390–412.

Banks, M. S., Akeley, K., Hoffman, D. M., & Girshick, A. R. (2008). Consequences of incorrect focus cues in stereo displays. *Information Display*, *24*(7), 10–14.

Bennett, K. M. B., Mucignat, C., Waterman, C., & Castiello, U. (1994). Vision and the reach to grasp movement. In K. M. B. Bennett, & U. Castiello (Eds.), *Insights into the reach to grasp movement* (pp. 171–195). Amsterdam: Elsevier Science.

Bingham, G. P. (2005). Calibration of distance and size does not calibrate shape information: Comparison of dynamic monocular and static and dynamic binocular vision. *Ecological Psychology*, *17*(2), 55–74.

Bingham, G. P., Bradley, A., Bailey, M., & Vinner, R. (2001). Accommodation, occlusion and disparity matching are used to guide reaching: A comparison of actual versus virtual environments. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(6), 1314–1344.

Bingham, G., Coats, R., & Mon-Williams, M. (2007). Natural prehension in trials without haptic feedback but only when calibration is allowed. *Neuropsychologia*, *45*(2), 288–294.

[7] In our study, haptic feedback is provided in a subset of trials to the 30 mm deep objects rendered by disparity-only information. Therefore, a direct comparison between trials with and without haptic feedback is possible solely for these specific stimulus conditions.

Bingham, G. P., Crowell, J. A., & Todd, J. T. (2004). Distortions of distance and shape are not produced by a single continuous transformation of reach space. *Perception & Psychophysics*, *66*(1), 152—169.

Bingham, G. P., & Pagano, C. C. (1998). The necessity of a perception/action approach to definite distance perception: Monocular distance perception to guide reaching. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 145—168.

Bingham, G. P., Zaal, F., Robin, D., & Shull, J. A. (2000). Distortions in definite distance and shape perception as measured by reaching without and with haptic feedback. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(4), 1436—1460.

Bradshaw, M. F., Elliott, K. M., Watt, S. J., Hibbard, P. B., Davies, I. R., & Simpson, P. J. (2004). Binocular cues and the control of prehension. *Spatial Vision*, *17*(1–2), 95—110.

Bradshaw, M. F., Parton, A. D., & Glennerster, A. (2000). The task-dependent use of binocular disparity and motion parallax information. *Vision Research*, *40*(27), 3725—3734.

Braunstein, M. L. (1994). Decoding principles, heuristics and inference in visual perception. In G. Jansson, S. S. Bergstrom, & W. Epstein (Eds.), *Perceiving events and objects*. Hillsdale, NJ: Erlbaum.

Brenner, E., van Damme, W. J., & Smeets, J. B. (1997). Holding an object one is looking at: Kinesthetic information on the object's distance does not improve visual judgments of its size. *Perception & Psychophysics*, *59*(7), 1153—1159.

Bruggeman, H., Fantoni, C., Caudek, C., & Domini, F. (2010). Reaching movement accuracy is mainly determined by visual online control. *Perception*, *39*, 51 ECVP Abstract Supplement.

Coats, R., Bingham, G., & Mon-Williams, M. (2008). Calibrating grasp size and reach distance: Interactions reveal integral organization of reaching-to-grasp movements. *Experimental Brain Research*, *189*(2), 211—220.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, *4*, 102—118 (Suppl.).

Cuijpers, R. H., Brenner, E., & Smeets, J. B. J. (2008). Consistent haptic feedback is required but it is not enough for natural reaching to virtual cylinders. *Human Movement Science*, *27*(6), 857—872.

Di Luca, M., Domini, F., & Caudek, C. (2007). The relation between disparity and velocity signals of rigidly moving objects constrains depth order perception. *Vision Research*, *47*(10), 1335—1349.

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

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

Domini, F., & Caudek, C. (in press). Combining image signals before 3D reconstruction: The Intrinsic Constraint Model of cue integration. In Trommershäuser, J., Landy, M. S. & Körding, K. (Eds.), *Sensory cue integration*. New York: Oxford University Press.

Domini, F., Caudek, C., & Tassinari, H. (2006). Stereo and motion information are not independently processed by the visual system. *Vision Research*, *46*(11), 1701—1723.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429—433.

Fantoni, C. (2008). 3D surface orientation based on a novel representation of the orientation disparity field. *Vision Research*, *48*(25), 2509—2522, doi:10.1016/j.visres.2008.08.015.

Fantoni, C., Caudek, C., & Domini, F. (2010). Systematic distortions of perceived planar surface motion in active vision.*Journal of Vision*, *10*(5):12, 1—20, http://journalofvision.org/content/10/5/12, doi:10.1167/10.5.12.

Franz, V. H. (2003). Manual size estimation: A neuropsychological measure of perception? *Experimental Brain Research*, *151*(4), 471—477.

Gentilucci, M., Benuzzi, F., Gangitano, M., & Grimaldi, S. (2001). Grasp with hand and mouth: A kinematic study on healthy subjects. *Journal of Neurophysiology*, *86*(4), 1685—1699.

Glover, S. (2004). Separate visual representations in the planning and control of action. *The Behavioral and Brain Sciences*, *27*(1), 3—24.

Goodale, M. A., Jakobson, L. S., & Keillor, J. M. (1994). Differences in the visual control of pantomimed and natural grasping movements. *Neuropsychologia*, *32*(10), 1159—1178.

Greenwald, H. S., & Knill, D. C. (2009). A comparison of visuomotor cue integration strategies for object placement and prehension. *Visual Neuroscience*, *26*(1), 63—72.

Greenwald, H. S., Knill, D. C., & Saunders, J. A. (2005). Integrating visual cues for motor control: A matter of time. *Vision Research*, *45*(15), 1975—1989.

Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research*, *179*(4), 595—606.

Held, R. T., & Banks, M. S. (2008). Misperceptions in stereoscopic displays: A vision science perspective. *ACM Transactions on Graphics*, *APGV08*, 23—31.

Helmholtz, H. V. (1867). Handbuch der Physiologischen Optik. *Treatise on physiological optics*. Hamburg: Verlag von Leopold Voss Southall JPC (New York: Dover, 1962).

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.

Hoffman, D. M., Girshick, A. R., Akeley, K., & Banks, M. S. (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, *8*(3), 1—30 33.

Jackson, S. R., Jones, C. A., Newport, R., & Pritchard, C. (1997). A kinematic analysis of goal-directed prehension movements executed under binocular, monocular, and memory-guided viewing conditions. *Visual Cognition*, *42*, 113—142.

Jakobson, L. S., & Goodale, M. A. (1991). Factors affecting higher-order movement planning: A kinematic analysis of human prehension. *Experimental Brain Research*, *86*(1), 199—208.

Jeannerod, M. (1981). Intersegmental coordination during reaching at natural visual objects. In J. Long, & A. Baddeley (Eds.), *Attention and performance IX* (pp. 153—168). Hillsdale, NJ: Erlbaum.

Jeannerod, M. (1984). The timing of natural prehension movements. *Journal of Motor Behavior*, *16*(3), 235—254.

Jeannerod, M. (1988). *The neural and behavioural organization of goal directed movements.* Oxford, England: Oxford University Press.

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*(3), 389—412.

Landy, M. S., Banks, M. S. & Knill, D. C. (in press). Ideal-observer models of cue integration. In Trommershuser, J., Landy, M. S. & Koerding, K. (Eds.), *Sensory cue integration*. New York: Oxford University Press.

Loftus, A., Servos, P., Goodale, M. A., Mendarozqueta, N., & Mon-Williams, M. (2004). When two eyes are better than one in prehension: Monocular viewing and end-point variance. *Experimental Brain Research*, *158*(3), 317—327.

MacKenzie, K. J., Murray, R. F., & Wilcox, L. M. (2008). The intrinsic constraint approach to cue combination: An empirical and theoretical evaluation. *Journal of Vision*, *8*(8), 1—10, doi:10.1167/8.8.5.

Mayhew, J. E. W., & Longuet-Higgins, H. C. (1982). A computational model of binocular depth perception. *Nature*, *297*, 376—378, doi:10.1038/297376a0.

Melmoth, D. R., & Grant, S. (2006). Advantages of binocular vision for the control of reaching and grasping. *Experimental Brain Research*, *171*(3), 317—388.

Melmoth, D. R., Storoni, M., Todd, G., Finlay, A. L., & Grant, S. (2007). Dissociation between vergence and binocular disparity cues in the control of prehension. *Experimental Brain Research*, *183*(3), 283—298, doi:10.1007/s00221-007-1041-x.

Mon-Williams, M., & Bingham, G. P. (2007). Calibrating reach distance to visual targets. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(3), 645—656.

Mon-Williams, M., Coats, R., & Bingham, G. P. (2004). Reaching with feeling. *Journal of Vision*, *4*(8), 411a, doi:10.1167/4.8.411.

Mon-Williams, M., & Dijkerman, C. (1999). The use of vergence information in the programming of prehension. *Experimental Brain Research*, *128*(4), 578—582.

Mon-Williams, M., & Tresilian, J. R. (1999). The size–distance paradox is a cognitive phenomenon. *Experimental Brain Research*, *126*, 578—582.

Oruc, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, *43*(23), 2451—2468.

Pagano, C. C., & Bingham, G. P. (1998). Comparing measures of monocular distance perception: Verbal and reaching errors are not correlated. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1037—1051.

Phillips, F., Egan, E., & Perry, B. (2009). Perceptual equivalence between vision and touch is complexity dependent. *Acta Psychologica*, *132*, 259—266, doi:10.1016/j.actpsy.2009.07.010.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, *317*, 314—319, doi:10.1038/317314a0.

Robert, L., Zeller, C., Faugeras, O., & Hébert, M. (1997). Applications of non-metric vision to some visually-guided robotics tasks. In Y. Aloimonos (Ed.), *Visual navigation: From biological systems to unmanned ground vehicles* (pp. 89—134). Lawrence Erlbaum Associates.

Servos, P., Goodale, M. A., & Jakobson, L. S. (1992). The role of binocular vision in prehension: A kinematic analysis. *Vision Research*, *32*(8), 1513—1521.

Smeets, J. B. J., & Brenner, E. (2010). Vision for action is not veridical. *Cognitive Neuroscience*, *1*(1), 69.

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

Thaler, L., & Goodale, M. A. (2010). Beyond distance and direction: The brain represents target locations non-metrically. *Journal of Vision*, *10*(3), 1—27, doi:10.1167/10.3.3.

Tittle, J. S., Norman, J. F., Perotti, V. J., & Phillips, F. (1998). The perception of scale-dependent and scale-independent surface structure from binocular disparity, texture, and shading. *Perception*, *27*(2), 147—166.

Todd, J. T. (2004). The visual perception of 3D shape. *Trends in Cognitive Sciences*, *8*(3), 115—121, doi:10.1016/j.tics.2004.01.006.

Tresilian, J. R., & Mon-Williams, M. (2000). Getting the measure of vergence weight in nearness perception. *Experimental Brain Research*, *132*, 362—368.

Tresilian, J. R., Mon-Williams, M., & Kelly, B. M. (1999). Increasing confidence in vergence as a distance cue. *Proceedings of the Royal Society B*, *266*, 39—44.

Van Doorn, G. H., Richardson, B. L., Wuillemin, D. B., & Symmons, M. A. (2010). Visual and haptic influence on perception of stimulus size. *Attention, Perception & Psychophysics*, *72*, 813—822, doi:10.3758/APP.72.3.813.

Watt, S. J., Akeley, K., Ernst, M. O., & Banks, M. S. (2005). Focus cues affect perceived depth. *Journal of Vision*, *5*, 834—862.

Watt, S. J., Banks, M. S., Ernst, M. O., & Zumer, J. M. (2002). Screen cues to flatness do affect 3D percepts. *Journal of Vision*, *2*(7), 297a, doi:10.1167/2.7.297, http://www.journalofvision.org/2/7/297/.

Watt, S. J., & Bradshaw, M. F. (2000). Binocular cues are important in controlling the grasp but not the reach in natural prehension movements. *Neuropsychologia*, *38*(11), 1473—1481.

Watt, S. J., & Bradshaw, M. F. (2003). The visual control of reaching and grasping: Binocular disparity and motion parallax. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 404—415.

Wickelgren, E. A., McConnell, D., & Bingham, G. P. (2000). Reaching measures of monocular distance perception: Forward versus side-to-side head movements and haptic feedback. *Perception & Psychophysics*, *62*(5), 1051—1059.

Wijntjes, M. W. A., Volcic, R., Pont, S. C., Koenderink, J. J., & Kappers, A. M. L. (2009). Haptic perception disambiguates visual perception of 3D shape. *Experimental Bain Research*, *193*, 639—644, doi:10.1007/s00221-009-1713-9.

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill, & W. Richards (Eds.), *Bayesian perspectives on visual perception* (pp. 123—161). Cambridge University Press.