Pose Estimation Through Cue Integration: A Neuroscience-Inspired Approach

Eris Chinellato, *Member, IEEE*, Beata J. Grzyb, and Angel P. del Pobil

***Abstract*—The aim of this paper is to improve the skills of robotic systems in their interaction with nearby objects. The basic idea is to enhance visual estimation of objects in the world through the merging of different visual estimators of the same stimuli. A neuroscience-inspired model of stereoptic and perspective orientation estimators, merged according to different criteria, is implemented on a robotic setup and tested in different conditions. Experimental results suggest that the integration of multiple monocular and binocular cues can make robot sensory systems more reliable and versatile. The same results, compared with simulations and data from human studies, show that the model is able to reproduce some well-recognized neuropsychological effects.**

***Index Terms*—Biological system modeling, grasping, intelligent robots, robot vision systems, stereo vision.**

I. Introduction

PRIMATES possess a superior ability in dealing with objects in their environment. One of the keys for achieving such ability is the continuous concurrent use of multiple estimators deriving from different cues, particularly of visual nature. Cue integration is indeed a major principle in the primate sensory cortex. Visual information is processed in a highly parallel way, and different estimators for the same stimulus are processed, compared, and merged in order to provide increased estimation reliability through redundancy [24], [45]. Often, motion and texture cues are at least as informative as stereoptic data, and a method for integrating all the available information for obtaining the most likely estimate is required. Although some modern artificial vision approaches build strongly on biological concepts [40], these principles have not been exploited up to their potentialities in robotics. With this work, we put forward a proposal for improving the reliability of artificial systems in the estimation of visual features in 3-D based on neuropsychological concepts.

In a previous related work, the thorough study of neuroscience research related to the integration of monocular and binocular retinal information for estimating object pose allowed us to define a modular computational structure composed of various estimators and different ways of merging them [7]. In this paper, we apply the aforementioned computational model to a real robotic platform, where a robot is required to observe boxlike shapes of different size and proportion and estimate the features useful for a potential grasping action. Estimation of object slant and distance is performed merging a number of different stereoptic and perspective estimators. In fact, our working hypothesis is that a compound estimator merging multiple estimators of different nature, both monocular and binocular, should provide the robot with superior estimation capabilities. The employment of visual cues that work differently in varying visual and pose conditions is expected to provide an important advantage for obtaining stable and reliable measures.

The experimental results presented in this paper, regarding pose estimation tests performed by the robotic visual system with various objects in different visual conditions, confirm the aforementioned hypothesis. More specifically, our experiments show that our global merged estimator, obtained by appropriately weighting the contribution of each simple estimator, provides very good performances and is robust across working conditions, offering a solution to pose estimation in real environments inspired on biological concepts. We show that such performances are not attainable by a simple average of stereoptic and perspective cues and, even less, by each estimator alone.

In addition to the contribution to robotics, our work provides interesting insights to the study of human visual mechanisms. We had previously reproduced in simulation the effect of different driving factors on estimation reliability, as reported by the neuroscience literature [7]. The new robot experiments presented here offer a good approximation of those same effects, providing further support to the plausibility of our computational model of pose estimation.

The background of this research, both in neuroscience and computational vision, is introduced in Section II. Section III describes the model on which our implementation is based. The robotic implementation is explained in Section IV, and experimental results are presented and discussed in Section V.

Manuscript received August 12, 2010; revised March 24, 2011 and August 22, 2011; accepted September 10, 2011. Date of publication October 20, 2011; date of current version March 16, 2012. This work was supported in part by the Spanish Ministerio de Ciencia y Innovación (DPI-2008-06636, FPU Grant AP2007-02565), by Fundació Caixa-Castello-Bancaixa (P1-1B2008-51), and by World Class University Program through the National Research Foundation of Korea, Ministry of Education, Science and Technology, under Grant R31-2008-000-10062-0. This paper was recommended by Associate Editor N. Cassimatis.

E. Chinellato was with the Robotic Intelligence Laboratory, Jaume I University, 12071 Castellón de la Plana, Spain. He is now with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: e.chinellato@imperial.ac.uk).

B. J. Grzyb is with the Robotic Intelligence Laboratory, Jaume I University, 12071 Castellón de la Plana, Spain (e-mail: grzyb@uji.es).

A. P. del Pobil is with the Robotic Intelligence Laboratory, Jaume I University, 12071 Castellón de la Plana, Spain, and also with the Department of Interaction Science, Sungkyunkwan University, Seoul 110-745, Korea (e-mail: pobil@uji.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2011.2168952

1083-4419/$26.00 © 2011 IEEE

II. Background

For both natural and artificial agents, the ability of estimating distance, size, and shape of surrounding objects is highly supported by, if not fully dependent on, the use of binocular or stereoscopic vision [28]. Binocular vision consists in the acquisition of two different images taken from viewpoints, the eyes, that are always at the same short distance. The process allows to obtain a fast and accurate estimation of object distance, size, and motion, through the interpretation of binocular disparities.

Despite its fundamental importance, stereoptic information alone is often not enough, and motion, texture, shading, and other cues are used to complement it. Indeed, in each modality, the brain seems to efficiently use a large set of different cues at the same time [31]. Cue evaluation and integration is a major principle in the primate sensory cortex, and particularly in vision, in order to obtain the most likely final estimates of stimulus properties. Visual information is processed in a highly parallel way, and different cues for the same stimulus are processed, compared, and merged in order to provide increased estimation reliability through redundancy [24]. In this section, vision science concepts and approaches related to cue generation and integration both in natural and artificial systems are reviewed.

A. *Visual and Visuomotor Brain Areas*

The basic mechanisms of stereoscopic vision have been studied for long time and are discussed in fundamental works [22], [29]. Neuronal responses to disparity stimuli in cortical visual areas have also been thoroughly investigated [10], [33]. Disparity detection is a fundamental aspect of visual processing that begins already in primary visual areas of the primate brain. In higher visual regions, disparity coding spans areas of the visual field wide enough to provide a proper interpretation of stereoptic information, both in monkeys and in humans [44]. Advanced visual areas are thought to be in charge of processing both higher order disparities and basic perspective cues [48].

Of special interest for the research presented in this paper is an area of the posterior parietal cortex of primates, the *caudal intraparietal sulcus* (CIP), which is dedicated to the extraction and description of visual features suitable for grasping purposes. Its neurons are strongly selective for the orientation of visual stimuli, represented in a viewer-centered way. Selectivity toward disparity-based orientation cues is predominant in monkey and human's CIP [38], [44]. On the other hand, many CIP neurons also respond (some exclusively) to perspective-based orientation cues. The evidence suggests that CIP integrates stereoptic and perspective cues for obtaining better estimates of orientation [45], [48]. This sort of processing performed by CIP neurons is the logical continuation of the simpler orientation responsiveness found in earlier visual areas and makes CIP the ideal intermediate stage toward the grasping-based object representations of downstream associative areas [3], [38].

Distance and location estimation of target objects is also performed in primates' posterior parietal cortex. More exactly, according to psychophysiological research in humans [42], what is actually estimated and used is the reciprocal of distance, that is, nearness. In the parietal cortex, distance and disparity are processed together, the former acting as a gain modulation variable on the latter [14]. This mechanism allows to properly interpret stereoscopic visual information [30], as described in Section III-A.

B. *Cue Integration*

Cue integration, or combination, is one of the main working principles of the human sensory systems. Restricting to unimodal cue integration, vision is probably the best example of the complexity reached in the process of getting the best estimate of a stimulus from concurring and often discordant cues. Several models have been proposed for explaining how such best estimate is obtained, but most phenomena can be modeled by a simple linear weighting of concurrent cues, aimed at maximizing the likelihood of the final estimate [24]. The main underlying principles that allow us to achieve this goal seem to be two: cue reliability and cue correlation, or discrepancy [42].

Cue reliability is probabilistic; it depends on environmental conditions, on the estimate itself, and, sometimes, on other ancillary measures [24]. Considering the case of interest in this thesis, i.e., orientation estimation, stereoscopic cues are considered less reliable outside a certain range of disparity but also at longer distances, being distance in this case an ancillary cue. Often, ancillary cues directly affect the estimation process through gain modulation, such as in the mentioned distance/disparity example [43]. This seemingly simple and safe mechanism may nevertheless suffer because of a second-order uncertainty, the problem of assessing the reliability of the ancillary cue itself. In any case, reliability rules have to be learnt by a subject in her/his interaction with the environment and can be misleading in the case of unusual situations, such as in optical illusions.

The second principle, cue correlation, considers the degree to which concurrent cues conflict or coincide and gains importance with increasing number of cues. In fact, there is no way to choose between two conflicting cues only on the base of cue correlation, but if a cue is the only one in disagreement with a number of coincident cues, it is very reasonable to consider it untrustworthy. Fortunately, vision systems often provide many cues quite different from each other, so that correlation can be a perfect criterion for weighting the cues in the final estimate [1].

The available models for extraction and integration of visual cues usually focus on early visual processing [37] or on very specific aspects, such as conflicting stimuli [46], maximum-likelihood cue integration [20], temporal integration according to cue reliability [17], and extraction of local surface slant [21]. Apparently, no published models on the subject provide details for practical implementation on robotic vision setups.

C. *Artificial Vision and Robotics*

Object orientation (or slant) estimation is a common, and difficult, problem in artificial vision [27]. Nevertheless, no research works similar to the proposed approach are available in the literature. Existing techniques for pose estimation still build on the fundamental concepts described by Goddard [16]. The available approaches differ depending on the type and location of the sensors, the illumination requirements, the object or scene feature on which the pose is calculated, and the relative motion between robot and object. Sometimes, noise sources and uncertainty factors are modeled in an attempt to improve the robustness and accuracy of the results. Among various methods, geometry- or model-based techniques are most common. These methods use an explicit model for the geometry of the object in addition to its image in determining the pose. The object is modeled in terms of points, lines, curves, planar surfaces, or quadric surfaces [34]. Methods of this kind have been proved useful also with moving targets [26] and even with articulated shapes [25]. Some of these methods build on cognitive science concepts, like Peters' [32], in which viewpoint-based sparse representations of objects are used. Often, the use of markers substitutes explicit modeling [13]. Model-based techniques can be combined with others, where appearance-based methods are used for a rough initial estimate followed by a refinement step [11]. A model-based approach can also be connected with range images, for example, matching a 3-D model to a range representation of the scene [15]. The managing of range data is anyway quite different from vision research, and works which locate parallel surfaces to grip from range images, such as [47], are interesting but have little in common with the current approach. In a work more related to this paper, Xu *et al.* [49] consider parallel lines to self-calibrate a pair of cameras and estimate the pose of geometric features. Some of their initial assumptions are similar to ours, but they do not include perspective data and their approach is not biologically inspired.

For what concerns stereo slant estimation inspired on human physiology, Ferrier [12] describes a method based on binocular disparities which make use of a model for computing orientation of features based on eye orientation. The results they obtain are consistent with, and complementary to, those presented in this paper. The idea of integrating stereoptic and perspective cues in artificial vision is not novel [9], but only one robotic platform is nowadays making use of both visual cues at the same time [39]. In Saxena *et al.*' setup, a vision system is trained to estimate scene depth through monocular data using supervised learning, and a joint monocular/binocular estimator is generated. The authors show that integration of monocular and stereopsis data performs better than either cue alone. The main difference between this work and our approach is the use of traditional computer vision techniques as opposed to our biologically inspired computational neuroscience-based implementation. Other works, focused on object tracking [41] and on visual servoing [23], perform cue integration, but their visual analysis is model based and their goal is feature matching and not feature extraction.

III. Computational Model

As part of a computational model for vision-based grasping based on neuroscience findings [5], [8], we developed and implemented a model of distance and orientation estimation inspired by human visual mechanisms. Here, we briefly explain only those concepts necessary to properly understand the robotic implementation; please refer to the original paper for additional references and details on computational issues [7]).

The model is based on the integration of monocular and binocular cues. The implemented cues are perspective under the assumption of edge parallelism and width disparity. We assume that the target object has straight parallel edges and is standing upright. This is reasonable from a neuropsychological point of view, as the primate brain is actually "programmed" to better assess vertical and horizontal edges, most common in nature. Indeed, experiments on monkeys [45] and humans [2] have shown that, even for purely perspective pose estimations, a frontoparallel trapezoid is usually interpreted as a rectangular shape slanted in depth. The model does not take into account cyclorotation movements, which are usually not implemented in robot vision systems, although their inclusion would constitute an interesting challenge for future developments [19].

A. Orientation Estimation

We provide here a brief description of how we compute slant estimation from stereopsis and perspective visual information, and distance from proprioceptive eye data.

1) Perspective: The slant of an object can be estimated using only monocular data, as shown in Fig. 1(a), in which the origin of the axes is one of the eyes. As explained earlier, given a rectangle slanted in depth, we can exploit the assumption of parallelism and equality of opposite edges (PS and RQ in the image). Angles β in the figure represent the vertical retinal angles associated to such edges. The equation which leads from retinal angles to orientation estimation can be derived from Fig. 1(a) [7], and can be referred entirely to either the left or the right eye (monocular separation $\psi_Q = (\alpha_Q - \alpha_P)/2$):

$$\tan\theta = \frac{\tan\beta_{QR}}{\tan\beta_{PS}\sin\psi_Q} - \frac{1}{\tan\psi_Q}. \qquad (1)$$

2) Stereopsis: In Fig. 1(b), a viewing scene is seen from above: Object PQ of length l is slanted about a vertical axis with orientation θ. Its extreme P is the fixation point, placed straight ahead from the cyclopean eye (a point lying slightly behind the midpoint between the eyes). All α angles represent the retinal projections of points P and Q on the left and right eyes, I is the interocular distance, ψ_Q is the binocular separation of points P and Q (being $\psi_P = 0$), and γ_P is the vergence angle. The horizontal slant θ of an object can be computed only from retinal angles using the following expression, which can be derived from the image [7]:

$$\tan\theta = \frac{(\tan\alpha_{rQ} - \tan\alpha_{lQ}) - (\tan\alpha_{rP} - \tan\alpha_{lP})}{\tan\alpha_{lP}\tan\alpha_{rQ} - \tan\alpha_{lQ}\tan\alpha_{rP}}. \qquad (2)$$

B. Distance Estimation

The distance of a fixated point from a viewer can be estimated by either retinal and/or proprioceptive information

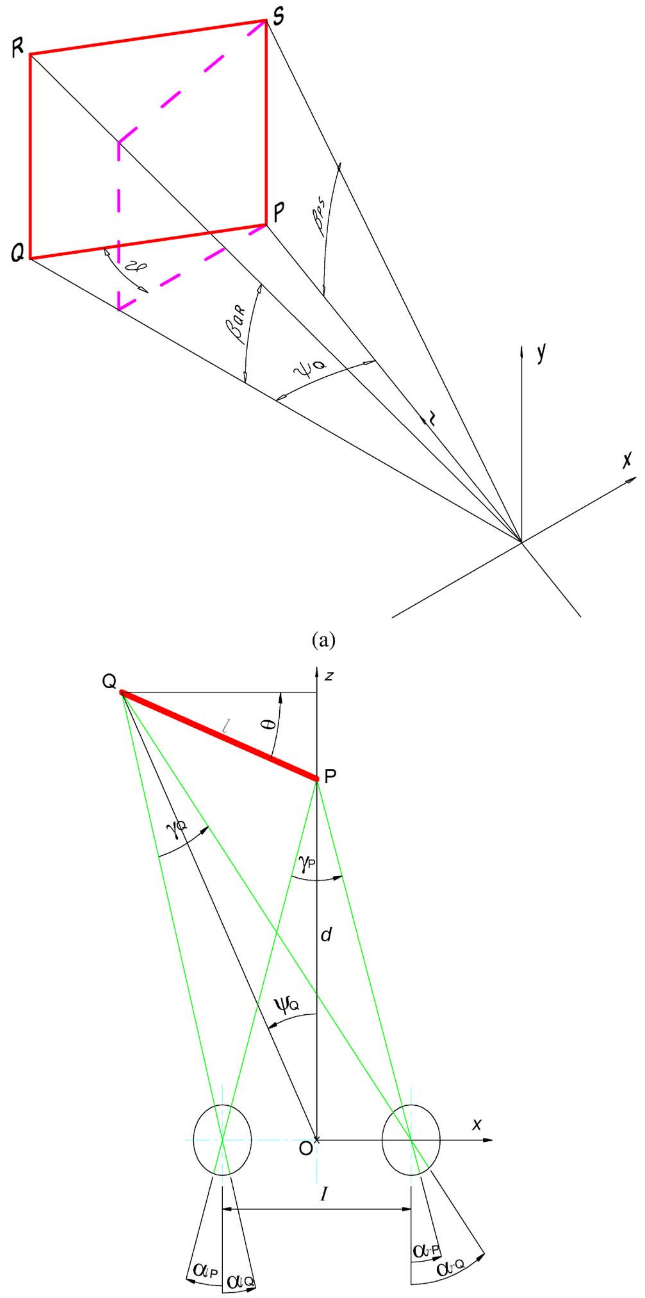

Fig. 1. Schemes for deriving slant from perspective and stereopsis, adapted from [7]. (a) Perspective. (b) Stereopsis.

(accommodation and vergence). Proprioceptive cues are preferably used when retinal data are not available or considered not reliable, and for short distances [1]. Neuropsychological experiments [42] suggest that distance estimation is most probably performed in our brain using *nearness* units instead of distance units. Nearness is the reciprocal of distance, so that a point at infinite distance has zero nearness and a point at the maximum vergence angle has nearness of one (or 100%). Although, in our computational implementation with radial basis functions [7], we showed that nearness is a more convenient measure, for practical purposes, we will use distance in this paper. The distance estimator that we designed is based on proprioceptual vergence data, according to the following simple equation,

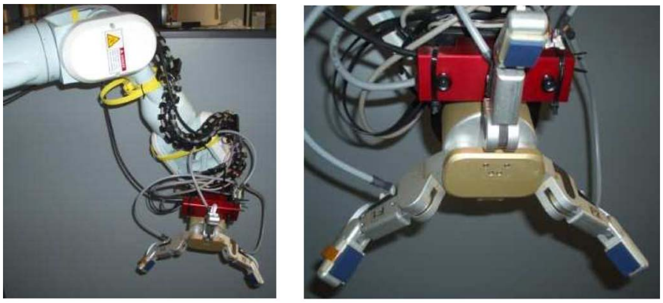

Fig. 2. Robotic setup with arm, hand, and stereoscopic camera. (a) Robotic arm and hand. (b) Detail of hand with stereo camera.

where d is the viewing distance, I is the interocular interval, and γ is the vergence angle:

$$d = \frac{I}{2\tan(\gamma_P/2)}. \tag{3}$$

In [7], we implemented with neural networks a multiple cue orientation estimator which makes use of the equations provided in this section and different cue merging methods. The data from neuropsychological experiments that we were able to reproduce are explained in Section V-A. In the next section, we extend such approach to robotics research.

IV. Robotic Pose Estimation

The extension of our computational model to robotics has been done with two purposes. The first goal is to obtain an orientation estimator that is very robust and reliable to use in a robotic vision-based grasping system. The second goal is to try and reproduce the mentioned effect with real experimental data, thus further validating the model.

From a robotic point of view, our approach for computing the orientation of an object is original in that it pursues estimation reliability through the merging of different estimation methods, as in the primate brain. We have implemented the described computational method on our robotic setup and performed a number of different experiments to verify how the ideal results change when the model has to face the uncertainties of the real world.

A. Setup and Visual Processing

Our robotic setup, shown in Fig. 2, consists of a seven-degree-of-freedom (DOF) Mitsubishi PA10 arm, on which are mounted a three-finger four-DOF Barrett hand and a stereoscopic camera Videre Design (eye-in-hand style). The robot world is a dark environment in which boxlike clear shapes are placed on a table at variable positions and orientations (see Fig. 3). The range of possible positions includes those that allow to view the object and also keep it at reaching distance for the hand. The system is able to estimate distance, pose, and size of the object without using models, only exploiting the assumption, supported by neuroscience studies, that what looks like a trapezoid is most likely a slanted shape having parallel and equal edges [38].

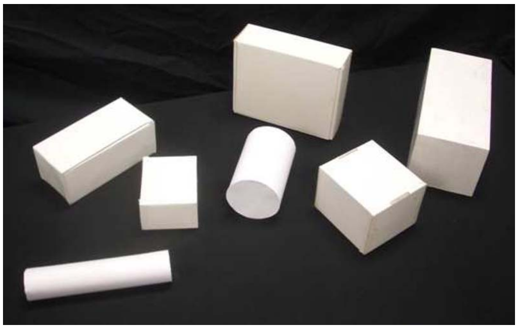

Fig. 3. Workspace with possible target objects.

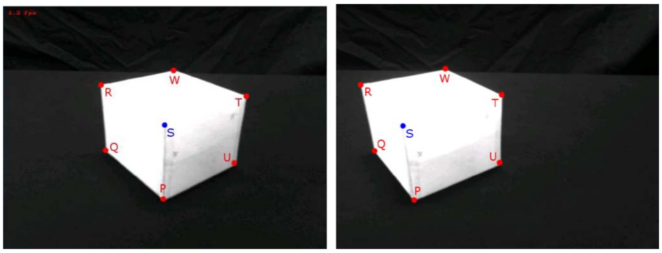

Fig. 4. Left and right images from the initial position, with labels of detected corners.

The choice of object and background color was driven by the need of keeping image processing as fast and lean as possible. Given the image of a target object, we binarize it, extract the contour, and detect its salient points, corresponding to the object corners (Fig. 4). Object faces are not segmented separately, so the number of detected corners ranges from four to six depending on point of view and object pose. Even with this simplified setup, to reliably detect the salient points, we perform a double search on the contour, combining the information given by two different algorithms for corner [4] and edge detection, to maximize the chance of finding all six visible corners of the object when possible. Although a system able to segment the three faces of the object separately would probably provide a better estimate, good results were obtained with this simple approach.

The three variables that identify object position and orientation are the distance d, the slant angle with respect to the frontoparallel position θ (see Fig. 1), and the direction of view with respect to the horizontal plane ϕ. This last variable is known by the robot and is computed by the vestibular system in primates. We restricted the viewing direction angle in our experiments, to allow a clear perspective view without simplifying too much the task as it happens for large angles (in such cases, the slant is very similar to what can be estimated simply using the inclination of segments in the 2-D image). The final working range is about $15^\circ < \phi < 50^\circ$, and these are very plausible values even for a human subject looking at an object with grasping purposes. For what concerns the slant θ, we only rule out those situations that would reduce the interest of the slant estimation (for angles very close to 0° and 90°) and which can anyway be detected quite easily by the system, from the number and distribution of the defining corners.

The process of distance, pose, and size estimation begins with the arm moving until the object is placed horizontally at the center of the image, in order to minimize distortions due to the cameras' optic. Left and right images at this position are then processed: Corners P, Q, R, W, T, and U are found as previously explained, and the position of S is estimated through a two-point perspective method (Fig. 4). At this point, the coordinates of the defining points are transformed in angles with respect to the center of the image, using the camera focal lens and image size in pixels as parameters. The nonlinearity of the camera optic is the reason to avoid getting close to the image borders, where distortions could affect the transformation process.

TABLE I
ORIENTATION ESTIMATORS

$\imath$	**Estimator**	**Computation Method**
1	Perspective I	Segments PS/QR, left eye
2	Perspective II	Segments PS/QR, right eye
3	Perspective III	Segments UT/PS, left eye
4	Perspective IV	Segments UT/PS, right eye
5	Stereopsis I	Segment PQ
6	Stereopsis II	Segment SR
7	Stereopsis III	Segment UP
8	Stereopsis IV	Segment TS
9	Merged ($\hat{\theta}_P$)	Perspective Only Average, $i = 1..4$
10	Merged ($\hat{\theta}_S$)	Stereopsis Only Average, $i = 5..8$
11	Merged ($\hat{\theta}_A$)	$\hat{\theta}_P$ and $\hat{\theta}_S$ Simple Average, $i = 9, 10$
12	Merged ($\hat{\theta}_G$)	Global Simple Average, $i = 1..8$
13	Merged ($\hat{\theta}_W$)	Global Weighted Average, $i = 1..8$

B. Cue Estimation and Integration

Once the six points identifying the two frontal faces of the object for both cameras have been detected, the actual slant estimation process can begin. Eight different estimators are calculated using the equations provided in Section III-A. Equation (1) for slant estimation through perspective information is applied to the couples of segments PS/QR and UT/PS for both the left and right eyes. Stereoptic slant estimation [(2)] is applied instead to segments PQ, SR, TS, and UP. In this way, we obtain the first eight estimators, four perspective and four stereoptic, in Table I.

As explained in Section II, there is a good evidence that primates make use of many different monocular and binocular cues and merge them according to their expected reliability and correlation. In our experiment, we start from a situation in which no information is available regarding the reliability of the different cues in the various working conditions. Thus, to begin with, there are only two solutions readily available without the need of performing a training session for learning the cue weights. The first is to compute a simple nonweighted average of the simple estimators (estimator 12 in Table I)

$$\hat{\theta}_G = \frac{1}{8}\sum_{i}^{8} \hat{\theta}_i. \tag{4}$$

The same can be done for perspective and stereoptic estimators separately, to obtain estimators 9–10 in Table I (estimator 11 is thus the average of the two). The second, slightly more complex, method is to compute an average in which weights are calculated using cue correlation (estimator 13). In our case, we use the deviation of each estimator from the simple average of all estimators, thus reinforcing estimators that are supposedly more representative of the whole measure

$$\hat{\theta}_W = \sum_i^8 w_i \hat{\theta}_i, \quad w_i = \frac{|\hat{\theta}_i - \hat{\theta}_G|}{\sum_j^8 |\hat{\theta}_j - \hat{\theta}_G|}. \tag{5}$$

The first experiments performed with the robot revealed that, sometimes, measurements were not stable and that one or more estimators could strongly deviate from the average in an unpredictable way. For this reason, before calculating the final merged estimator, we check for possible outliers (completely wrong estimations). In nature, bad estimations could be due to momentary occlusions, unusual light conditions, sudden movements, etc. In our simple setup, we discovered that any previous processing step could affect the final results, so, again, illumination issues, imperfections in the binarization, or corner detection can cause one or more cues to deviate hugely from the average estimate. For example, a bad detection of point U would affect estimators 3, 4, and 7. The average of all estimators would still be reasonably good, as deviations are usually random around the correct value, but estimators 3, 4, and 7 would differ substantially from the average and their exclusion would improve the global measure.

Outlier detection is a full subbranch of statistics [36], and many different methods are available. The methods we tried did not give significantly different results, and we finally chose the classical Rosner's many outliers test [35], widely used in the literature for similar problems. We tested the method with different values of the significance level α and obtained the best results for $\alpha = 0.01$, which gave a final estimation improvement of more than 5% compared to the implementation without outlier rejection. The possibility of performing outlier rejection is offered by the use of multiple estimators, which make sure that bad values of a small subset of all estimators do not fully corrupt the global average.

C. Distance and Size Estimation

As we assume that there is no previous knowledge regarding object dimension, it is not possible to disambiguate the pair distance/size only from retinal data. We thus make use of (3) and only have to estimate the proprioceptive vergence angle γ. Our stereo camera does not allow vergence movements of the eyes, so we have to simulate them. The simple procedure we adopt is to center point P of our object in one of the images first and rotate the camera around the cyclopean eye, in order to center again P on the other image without changing the actual distance. To take advantage of this movement, we take left and right images both from the initial and the final position and consider them as two independent slant estimation experiments. We actually observed that estimations from the initial position are slightly better than those from the final, probably because, although horizontal centering is performed in both cases, only the first experiment starts from an ideal vertical alignment of the object.

Regarding size estimation, the relative size of the object (proportion between its edges) can be detected from orientation and separation angles alone. Once distances have been estimated, the actual dimensions of the object can be computed through simple geometric equations, as the ambiguity size/distance has been resolved.

V. Results and Discussion

A. Simulation Results

Experiments with human subjects tell that distance, as an ancillary cue, and slant itself are the two most important driving factors for slant estimation reliability [1], [20]. With increasing distance, both perspective and stereoptic estimators become less reliable, but stereoscopic cues are more affected. The effect of orientation is more complex, as perspective methods are more sensitive and precise for pronounced slants, which generate higher differences in vertical disparities. Disparity methods also prefer high slants at long distances, but for short distances, the ones we use in all our vision-for-grasping experiments, their error is minimum for low slant values, which grant higher binocular disparities [20].

The purpose of our previous work [7] was to reproduce with our model some of the aforementioned effects. For pursuing this goal, we trained a set of neural networks with the equations described in the previous section and computed slant from different monocular and binocular visual features. Introducing random noise, we observed how the estimation performance was affected with variations of distance and slant itself. The similarity of the obtained results to what is described in the literature was remarkable [7]. The second effect we could reproduce was the improved performance obtained through a maximum-likelihood merged estimator in which cues were weighted according to their reliability (experimentally learnt), as explained in Section II. In this paper, we plan to further validate our model using real robot experiments. We want to test if the described trend is still valid when applying our equations to an actual hardware system acting in a real 3-D environment. Compared to the simulated tests, in which random noise was employed, we deal now with real-world uncertainty, presenting our slant estimation method with an entirely new level of complexity. At the same time, we want to check if the proposed bio-inspired approach has not only theoretical meaning but also practical usefulness in a real robotic application.

B. Experimental Results

Overall, we executed 422 slant estimation experiments with different values of slant and distance, as shown in Table II. The global average estimation errors of all executed experiments are provided in Table III. Perspective estimator $\hat{\theta}_P$ and stereopsis estimator $\hat{\theta}_S$ are calculated merging the four estimators of each modality alone. The simple average $\hat{\theta}_A$ is the mean between the two, and the global average $\hat{\theta}_G$ is the mean of all eight initial estimators. It can be observed how the combination of multiple cues, particularly when they come from different kinds of visual information, strongly improves the estimation performance. The worst merged estimator $\hat{\theta}_P$ performs better than the best single cue estimator, i.e., Stereopsis I; the global average $\hat{\theta}_G$

TABLE II
NUMBER OF EXPERIMENTS PER DISTANCE AND SLANT

Distance	Count	Slant	Count
450-500	14	10	12
500-550	40	20	80
550-600	66	30	96
600-650	74	40	80
650-700	88	50	92
700-750	94	60	48
750-800	28	70	14
800-850	18		

TABLE III
OVERALL AVERAGE ERRORS

ı	Estimator	Error (°)
1	Perspective I	8.63
2	Perspective II	6.67
3	Perspective III	12.75
4	Perspective IV	9.59
5	Stereopsis I	4.73
6	Stereopsis II	7.89
7	Stereopsis III	6.31
8	Stereopsis IV	5.41
9	Merged ($\hat{\theta}_P$)	4.71
10	Merged ($\hat{\theta}_S$)	3.92
11	Merged ($\hat{\theta}_A$)	3.78
12	Merged ($\hat{\theta}_G$)	2.91
13	Merged ($\hat{\theta}_W$)	2.68

improves the merged stereopsis estimator $\hat{\theta}_S$ by more than 25%. The cue correlation weighted average estimator $\hat{\theta}_W$ shows a further improvement of around 8% compared to $\hat{\theta}_G$, bringing the overall mean error close to 2.5°, which constitutes a very good pose estimation for a robotic grasping system. It is worth mentioning that even the least precise estimators, such as Perspective III, are not bad in all cases. This means that, most of the times, they still provide a useful contribution to the estimation process. On the other hand, when they deviate from the general trend, they are either ruled out as outsiders or their contribution to the weighted average is much reduced.

It is interesting to compare the error distributions obtained in the real practical experiments with the theoretical ones and those obtained in the simulation. Fig. 5 shows the average error plotted as a function of slant [Fig. 5(a)] and distance [Fig. 5(b)]. The error is thus averaged across distance in Fig. 5(a) and across slant in Fig. 5(b). The variability of the setup did not allow to obtain clean smooth curves, and some slant and distance values are probably affected by some external factors; see, for example, the bad quality of stereopsis and, consequently, of the merged estimators, for $slant = 60$. It is very difficult to understand why there are sudden drops or improvements in performance for certain values of slant or distance. We believe that they are due to particular shading properties for certain conditions that make corner detection more or less difficult. A more robust visual processing would probably solve, at least in part, this issue and would provide smoother curves. Nevertheless, the trends are quite clear, and the expected effect of slant and distance on the different estimators is reproduced. In Fig. 5(a), the improvement in perspective estimation and the deterioration in stereoptic estimation with increasing slant are clearly visible, and the weighted average is definitely the best available estimator. Fig. 5(b) shows that stereoptic estimation gradually decreases its precision with distance, while perspective seems nearly uncorrelated with it, apart for extreme values. As in the simulation, the weighted average presents a clearly advantageous behavior in all conditions. The application of our model on the robot has hence obtained the same general trends indicated by the cognitive science literature and by the simulated tests.

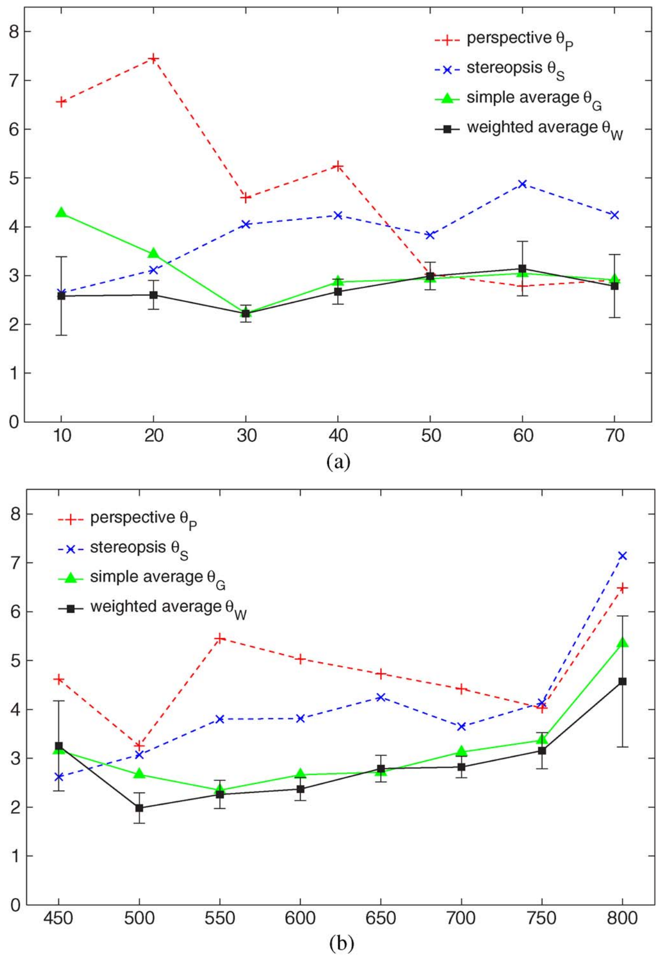

Fig. 5. Slant estimation error as a function of slant and distance; experimental results. For clarity, errors on errors are plotted only for θ_W. (a) Error (°) versus slant (°). (b) Error (°) versus distance (in millimeters).

Fig. 5 shows that the weighted estimator maintains its reliability across conditions. Error bars of θ_W are always small, apart for extreme conditions. Errors for other estimators (not plotted for clarity reasons) are always quite larger. This is a very important aspect for a robotic application, as there are no "blind spots" for which its estimation capabilities become unreliable. The implementation of a multiple cue estimation method thus provides a robotic system with a robustness hardly achievable with perspective or stereopsis alone.

For what concerns distance estimation, the global average error for all experiments is of 33.4 mm, and the error distribution shown in Fig. 6, although noisy, follows the expected trend, showing decreasing estimation precision with increasing distance. Size estimation revealed to be less precise compared to slant and distance estimation. In part, this is due to the fact

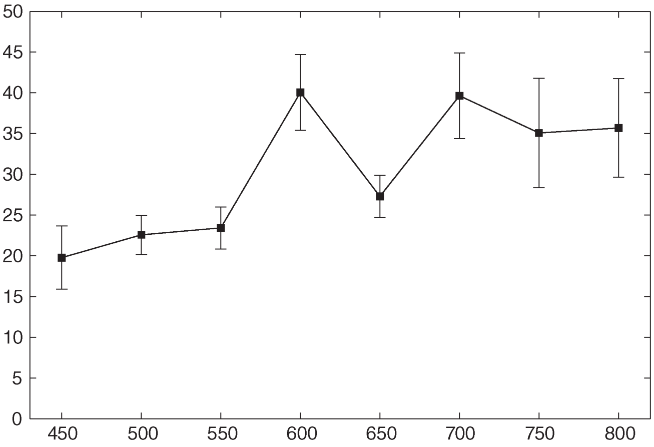

Fig. 6. Distance estimation error; error (in millimeters) versus distance (in millimeters).

that it makes use of two estimators and the theoretical final error is the product of the two initial errors. Moreover, for high slants and for small objects, the edges of the least visible side have very short separation angles, for which the relative error is much higher. Anyway, the worst case error is never larger than a few centimeters, and this is enough for reliable grasping by the robot hand, as it remained clear from experiments executed with our robotic grasping setup [5], [18].

VI. Conclusion and Extensions

The robotic implementation of a computational model for estimating object features in 3-D permitted us to achieve two important results. On the one hand, we provided our robotic grasping system with a very reliable and versatile visual estimation of slant, distance, and size of target objects. On the other hand, we could reproduce at a reasonable level of approximation effects described in neuropsychological data. Cue integration is the fundamental principle which allowed us to obtain such results, through the efficient merging of a set of stereoscopic and perspective estimators.

The research presented in this paper has been carried out for further developments in both engineering and scientific aspects. For what concerns the goal of modeling primate visual estimation of the properties of nearby objects, a full model of the information flow through the visual and visuomotor cortices has been completed and selectively implemented, using visual input to obtain reaching plans and candidate grips for a target object [5], [6].

From the pragmatic point of view of robot grasping performance, the model has been extended to allow the system to deal with other simple objects, such as cylinders and spheres. Successful reaching and grasping experiments have been performed using our estimated measures as inputs [18]. The accurateness of the final action is assessed through tactile feedback, and we plan to use it as a way to learn the reliability of each estimator in different conditions. The next-generation estimator will thus perform cue integration using both correlation and reliability, as in the primate brain.

Other two important extensions for improving our research are currently under development. The first is to make feature estimation more stable and less prone to lighting and shading conditions by using more elaborate visual processing techniques. With this extension, we should obtain clearer error trends and even better overall performances. A second interesting extension is to make our slant estimation system able to deal with moving targets or obstacles. Indeed, in our current research, we are pursuing the integration of a sequence of visual stimuli in time. Thus far, we have dealt with the effect of saccadic eye movements on the perception of the surrounding visual environment [6]. We are currently exploring the way we can employ concepts of visual integration through a sequence of eye movements to apply them to moving visual targets.

References

[1] B. T. Backus and M. S. Banks, “Estimator reliability and distance scaling in stereoscopic slant perception,” *Perception*, vol. 28, no. 2, pp. 217–242, 1999.

[2] G. J. Brouwer, R. van Ee, and J. Schwarzbach, “Activation in visual cortex correlates with the awareness of stereoscopic depth,” *J. Neurosci.*, vol. 25, no. 45, pp. 10 403–10 413, Nov. 2005.

[3] U. Castiello and C. Begliomini, “The cortical control of visually guided grasping,” *Neuroscientist*, vol. 14, no. 2, pp. 157–170, Apr. 2008.

[4] D. Chetverikov and Z. Szabo, “A simple and efficient algorithm for detection of high curvature points in planar curves,” in *Proc. Workshop Austrian Pattern Recognit. Group*, 1999, pp. 175–184.

[5] E. Chinellato, “Visual neuroscience of robotic grasping,” Ph.D. dissertation, Universitat Jaume I, Castelló de la Plana, Spain, 2008.

[6] E. Chinellato, M. Antonelli, B. J. Grzyb, and A. P. del Pobil, “Implicit sensorimotor mapping of the peripersonal space by gazing and reaching,” *IEEE Trans. Autonom. Mental Dev.*, vol. 3, no. 1, pp. 43–53, Mar. 2011.

[7] E. Chinellato and A. P. del Pobil, “Distance and orientation estimation of graspable objects in natural and artificial systems,” *Neurocomputing*, vol. 72, no. 4–6, pp. 879–886, Jan. 2009.

[8] E. Chinellato, Y. Demiris, and A. P. del Pobil, “Studying the human visual cortex for achieving action-perception coordination with robots,” in *Artificial Intelligence and Soft Computing*, A. P. del Pobil, Ed. Anaheim, CF: Acta Press, 2006, pp. 184–189.

[9] J. J. Clark and A. L. Yuille, *Data Fusion for Sensory Information Processing Systems*. New York: Springer-Verlag, 1990.

[10] B. G. Cumming and G. C. DeAngelis, “The physiology of stereopsis,” *Annu. Rev. Neurosci.*, vol. 24, pp. 203–238, 2001.

[11] S. Ekvall, F. Hoffmann, and D. Kragic, “Object recognition and pose estimation for robotic manipulation using color cooccurrence histograms,” in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, Oct. 2003, pp. 1284–1289.

[12] N. J. Ferrier, “Determining surface orientation from fixated eye position and angular visual extent,” in *Proc. IEEE Int. Conf. Robot. Autom.*, May 1999, pp. 938–943.

[13] S. K. Gehrig, H. Badino, and P. Paysan, “Accurate and model-free pose estimation of small objects for crash video analysis,” in *Proc. Brit. Mach. Vis. Conf.*, Edinburgh, U.K., Sep. 2006, pp. 1009–1018.

[14] A. Genovesio and S. Ferraina, “Integration of retinal disparity and fixation-distance related signals toward an egocentric coding of distance in the posterior parietal cortex of primates,” *J. Neurophysiol.*, vol. 91, no. 6, pp. 2670–2684, Jun. 2004.

[15] M. Germann, M. D. Breitenstein, I. K. Park, and H. Pfister, “Automatic pose estimation for range images on the GPU,” in *Proc. Int. Conf. 3-D Digit. Imag. Model.*, Aug. 2007, pp. 81–90.

[16] J. S. Goddard, “Pose and motion estimation using dual quaternion-based extended Kalman filtering,” in *Proc. SPIE: Three-Dimensional Image Capture Appl.*, 1998, vol. 3313.

[17] H. S. Greenwald, D. C. Knill, and J. A. Saunders, “Integrating visual cues for motor control: A matter of time,” *Vision Res.*, vol. 45, no. 15, pp. 1975–1989, Jul. 2005.

[18] B. J. Grzyb, E. Chinellato, A. Morales, and A. P. del Pobil, “A 3D grasping system based on multimodal visual and tactile processing,” *Ind. Robot J.*, vol. 36, no. 4, pp. 365–369, 2009.

[19] M. Hansard and R. Horaud, “Cyclorotation models for eyes and cameras,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 151–161, Feb. 2010.

[20] J. M. Hillis, S. J. Watt, M. S. Landy, and M. S. Banks, “Slant from texture and disparity cues: Optimal cue combination,” *J. Vis.*, vol. 4, no. 12, pp. 967–992, Dec. 2004.

[21] D. G. Jones and J. Malik, "Determining three-dimensional shape from orientation and spatial frequency disparities," in *Proc. Eur. Conf. Comput. Vis.*, 1992, pp. 662–669.

[22] B. Julesz, *Foundations of Cyclopean Perception*. Cambridge, MA: MIT Press, 1971.

[23] D. Kragic and H. I. Christensen, "Cue integration for visual servoing," *IEEE J. Robot. Autom.*, vol. 17, no. 1, pp. 18–27, Feb. 2001.

[24] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young, "Measurement and modeling of depth cue combination: In defense of weak fusion," *Vision Res.*, vol. 35, no. 3, pp. 389–412, Feb. 1995.

[25] B. Li, Q. Meng, and H. Holstein, "Articulated pose identification with sparse point features," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 3, pp. 1412–1422, Jun. 2004.

[26] V. Lippiello, B. Siciliano, and L. Villani, "Position and orientation estimation based on Kalman filtering of stereo images," in *Proc. IEEE Int. Conf. Control Appl.*, Sep. 2001, pp. 702–707.

[27] V. Lippiello, B. Siciliano, and L. Villani, "3D pose estimation for robotic applications based on a multi-camera hybrid visual system," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2006, pp. 2732–2737.

[28] A. Loftus, P. Servos, M. A. Goodale, N. Mendarozqueta, and M. Mon-Williams, "When two eyes are better than one in prehension: Monocular viewing and end-point variance," *Exp. Brain Res.*, vol. 158, no. 3, pp. 317–327, Oct. 2004.

[29] D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman, 1982.

[30] M. Mon-Williams, J. R. Tresilian, and A. Roberts, "Vergence provides veridical depth perception from horizontal retinal image disparities," *Exp. Brain Res.*, vol. 133, no. 3, pp. 407–413, Aug. 2000.

[31] J. F. Norman, J. T. Todd, and F. Phillips, "The perception of surface orientation from multiple sources of optical information," *Perceptual Psychophys.*, vol. 57, no. 5, pp. 629–636, Jul. 1995.

[32] G. Peters, "Efficient pose estimation using view-based object representations," *Mach. Vis. Appl.*, vol. 16, no. 1, pp. 59–63, Dec. 2004.

[33] G. F. Poggio, F. Gonzalez, and F. Krause, "Stereoscopic mechanisms in monkey visual cortex: Binocular correlation and disparity selectivity," *J. Neurosci.*, vol. 8, no. 12, pp. 4531–4550, Dec. 1988.

[34] B. Rosenhahn, C. Perwass, and G. Sommer, "Pose estimation of 3D free-form contours," *Int. J. Comput. Vis.*, vol. 62, no. 3, pp. 267–289, May 2004.

[35] B. Rosner, "On the detection of many outliers," *Technometrics*, vol. 17, no. 2, pp. 221–227, May 1975.

[36] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

[37] S. P. Sabatini, F. Solari, G. Andreani, C. Bartolozzi, and G. M. Bisio, "A hierarchical model of complex cells in visual cortex for the binocular perception of motion-in-depth," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 1271–1278.

[38] H. Sakata, M. Taira, M. Kusunoki, A. Murata, K. Tsutsui, Y. Tanaka, W. N. Shein, and Y. Miyashita, "Neural representation of three-dimensional features of manipulation objects with stereopsis," *Exper. Brain Res.*, vol. 128, no. 1/2, pp. 160–169, Sep. 1999.

[39] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 2197–2203.

[40] F. Solari, J. Diaz, E. Ros, K. Pauwels, M. Van Hulle, N. Pugeault, S. P. Sabatini, G. Gastaldi, and N. Krueger, "Compact and accurate early vision processing in the harmonic space," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2007, pp. 213–220.

[41] G. Taylor and L. Kleeman, "Fusion of multimodal visual cues for model-based object tracking," in *Proc. Australasian Conf. Robot. Autom.*, Brisbane, Australia, Dec. 2003.

[42] J. R. Tresilian and M. Mon-Williams, "Getting the measure of vergence weight in nearness perception," *Exp. Brain Res.*, vol. 132, no. 3, pp. 362–368, Jun. 2000.

[43] Y. Trotter, S. Celebrini, B. Stricanne, S. Thorpe, and M. Imbert, "Neural processing of stereopsis as a function of viewing distance in primate visual cortical area V1," *J. Neurophysiol.*, vol. 76, no. 5, pp. 2872–2885, Nov. 1996.

[44] D. Y. Tsao, W. Vanduffel, Y. Sasaki, D. Fize, T. A. Knutsen, J. B. Mandeville, L. L. Wald, A. M. Dale, B. R. Rosen, D. C. Van Essen, M. S. Livingstone, G. A. Orban, and R. B. H. Tootell, "Stereopsis activates V3A and caudal intraparietal areas in macaques and humans," *Neuron*, vol. 39, no. 3, pp. 555–568, Jul. 2003.

[45] K.-I. Tsutsui, M. Jiang, K. Yara, H. Sakata, and M. Taira, "Integration of perspective and disparity cues in surface-orientation-selective neurons of area CIP," *J. Neurophysiol.*, vol. 86, no. 6, pp. 2856–2867, Dec. 2001.

[46] R. van Ee, M. S. Banks, and B. T. Backus, "An analysis of binocular slant contrast," *Perception*, vol. 28, no. 9, pp. 1121–1145, 1999.

[47] A. Weigl, K. Hohm, and M. Seitz, "Processing sensor images for grasping disassembly objects with a parallel-jaw gripper," in *Proc. TELEMAN Telerobotics Conf.*, 1995, pp. 527–532.

[48] A. E. Welchman, A. Deubelius, V. Conrad, H. H. Bülthoff, and Z. Kourtzi, "3D shape perception from combined depth cues in human visual cortex," *Nat. Neurosci.*, vol. 8, no. 6, pp. 820–827, Jun. 2005.

[49] D. Xu, Y. F. Li, Y. Shen, and M. Tan, "New pose-detection method for self-calibrated cameras based on parallel lines and its application in visual control system," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 5, pp. 1104–1117, Oct. 2006.

Eris Chinellato (M'03) received the Ph.D. degree in intelligent robotics from Jaume I University, Castellón de la Plana, Spain, in 2008, the M.Sc. degree in artificial intelligence from The University of Edinburgh, Edinburgh, U.K., in 2002, and the Industrial Engineering degree from Università degli Studi di Padova, Padova, Italy, in 1999.

He is currently with Imperial College London, London, U.K., working on neural models of social interaction to be applied to the iCub humanoid robot. He has published in influential journals and proceedings in robotics, neuroscience, and computational neuroscience and has served as Reviewer and Program Committee Member for international journals and conferences. His interdisciplinary research, integrating robotics with experimental and theoretical neuroscience, focuses on sensorimotor integration in both natural and artificial systems.

Dr. Chinellato was the recipient of the Best A.I. M.Sc. Student Prize from The University of Edinburgh. His Ph.D. thesis "Visual Neuroscience of Robotic Grasping" was among the finalists of the European Robotics Research Network Award for the Best European Ph.D. Thesis in robotics.

Beata J. Grzyb received the M.Sc. degree in computer science from Maria Curie-Sklodowska University, Lublin, Poland. She is currently working toward the Ph.D. degree in the Robotic Intelligence Laboratory, Jaume I University, Castellón de la Plana, Spain.

She has already published in several journals and proceedings. In her research, she follows the approach of cognitive developmental robotics and tackles problems related to body representation, peripersonal space representation, and perception of body effectivities, by means of synthesizing neuroscience, developmental psychology, and robotics.

Angel P. del Pobil received the B.Sc. degree in physics and the Ph.D. degree in engineering from the University of Navarra, Pamplona, Spain, in 1986 and 1991, respectively.

He is currently a Professor with Jaume I University (UJI), Castellón de la Plana, Spain, where he is the founding Director of the Robotic Intelligence Laboratory. He is a World Class University Visiting Professor at Sungkyunkwan University, Seoul, Korea. He has been an Invited Speaker of 49 tutorials, plenary talks, and seminars. He has over 200 scientific publications including ten books. His present research interests include perceptual robotics, robot physical interaction for manipulation, cognitive developmental robotics, robot learning for sensorimotor interaction, and the interplay between neurobiology and robotics.

Dr. del Pobil has chaired two technical committees of the IEEE Robotics and Automation Society and is a member of the Board of Directors of the European Robotics Research Network. He has been a Program or General Chair of six international conferences and some 15 workshops and has served on the program committees of over 100 international conferences.