

Chapter 5

Extraction of Grasp-Related Visual Features

A model for visual cue extraction and merging strongly inspired on primate, and especially human, psychophysiology is described in this chapter. This implements the part of the framework of Chap. 4 dedicated to the extraction of object visual features relevant for grasping purposes. The areas and connections of the model of Fig. 4.6 involved in this process are highlighted in Fig. 5.1.

The distance of a target object is estimated, similarly as in the lateral intraparietal sulcus LIP, using proprioceptive vergence data. The advantages of expressing and calculating distances in nearness units are discussed. Object orientation estimation, executed in the posterior intraparietal sulcus CIP, is performed combining binocular (stereoptic) and monocular (perspective) visual data. A theoretical analysis for deriving plausible expressions for slant estimation is accompanied by an implementation with a set of artificial neural networks. The behavior of the system in simulated noisy conditions suggests that the model is faithful to biological reality.

A first interaction between the two streams is implemented at this point. An object recognition module, representing area V4, classifies the target object into one of three basic shapes: boxes, cylinders and spheres. Even though such classification provides no direct information on object size and proportion, it allows to access a basic knowledge about the target shape which helps in the pose estimation process.

The outcome of applying the computational model to a real robotic platform is presented and discussed. The robot is required to observe target shapes of different size and proportion and estimate the features useful for a potential grasping action. The comparison of the obtained results with experiments described in the neuroscience literature confirms the effect of different driving factors on estimation reliability, showing how stereoscopic, perspective and merged estimators behave in different conditions. The same comparison is done for distance estimation obtained from proprioceptive vergence data.

In order to complement the background information provided in the previous chapters, some important concepts regarding cue extraction and integration, object recognition, and artificial vision methods for pose estimation, are given in the next section.

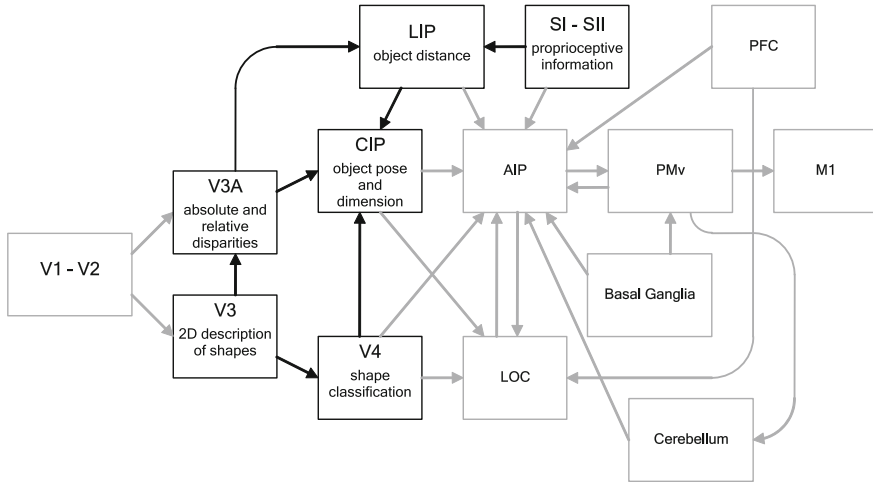


Fig. 5.1 Areas of the model framework involved in extraction of grasp-related visual features

5.1 Extraction and Integration of Object-Related Visual Cues in the Primate Cortex and in Robotics

For both natural and artificial agents, the interaction with the environment requires the ability of estimating distance, size and shape of surrounding objects. Such skill is highly supported by, if not fully dependent on, the use of binocular, or stereoscopic vision (Marotta et al. 1995; Watt and Bradshaw 2003; Bradshaw et al. 2004; Loftus et al. 2004). Binocular vision consists in the contemporaneous acquisition of two different images taken from viewpoints that are always at the same, short distance—the eyes. The process allows to obtain a fast and accurate estimation of object distance, size, motion, through the interpretation of *binocular disparities* (see box).

The difference between the left and right retinal representations of visual features is called *binocular disparity* (Howard and Rogers 2002; Parker 2004). Absolute disparities are simple distances, either horizontal or vertical, between the two retinal positions of the same point-like feature. Various types of higher order disparities can be computed from absolute ones. First order relative disparities represent the difference in disparity between two image points, and thus directly code for feature depth and slant. Horizontal relative disparities are used for estimating object slant about a vertical axis, the most common in nature. Orientation disparities allow instead to calculate slants about horizontal axes (Heeley et al. 2003). Second order disparities are used to estimate object curvature.

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 (Norman et al. 1995). Evaluation and integration of all available cues is performed in order to obtain the most likely final estimates. Cue integration is a major principle in the primate sensory cortex, and especially in vision. Visual information is processed in a highly parallel way, different cues for the same stimulus are processed, compared and merged in order to provide increased estimation reliability through redundancy (Landy et al. 1995; Tsutsui et al. 2001). In this section, vision science concepts related to cue generation and integration which help in complementing the review of Chap. 2 are provided.

5.1.1 Feature Extraction

The basic mechanisms of stereoscopic vision have been studied for long time, and are discussed in fundamental works such as Julesz (1971), Marr (1982). Neuronal responses to disparity stimuli in cortical visual areas have also been thoroughly investigated (Poggio et al. 1988; Cumming and DeAngelis 2001). Disparity detection is a fundamental aspect of visual processing that begins already in V1 and V2 (Gonzalez and Perez 1998; von der Heydt et al. 2000; Thomas et al. 2002; Trotter et al. 2004; Read 2005). It is though from V3 that disparity coding spans areas of the visual field wide enough to provide a proper interpretation of stereoptic information, both in monkeys (Adams and Zeki 2001; Tsao et al. 2003) and in humans (Backus et al. 2001; Welchman et al. 2005; Anzai et al. 2011). For what concerns the processing of higher order disparities, there is a general consensus regarding a prominent role of V3A in representing relative disparities (Backus et al. 2001; Tsao et al. 2003; Rutschmann and Greenlee 2004; Brouwer et al. 2005). An initial, basic perspective processing could also be performed in area V3A (Welchman et al. 2005; Georgieva et al. 2009).

As explained in Sect. 2.3.1, the caudal intraparietal sulcus CIP 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 macaque's CIP, which neurons are selective for first and second order disparities (Sakata et al. 1998; Endo et al. 2000; Taira et al. 2000). fMRI studies showed that the human posterior intraparietal sulcus is responsive to disparity-coded orientation, too (Tsao et al. 2003; Rutschmann and Greenlee 2004; Naganuma et al. 2005). On the other hand, many CIP neurons also respond (some exclusively) to perspective-based orientation cues, both in monkeys (Tsutsui et al. 2001, 2005) and humans (Taira et al. 2001).

The evidence suggests that CIP integrates stereoptic and perspective cues for obtaining better estimates of orientation (Tsutsui et al. 2005; Welchman et al. 2005). This sort of processing performed by CIP neurons is the logical continuation of the

simpler orientation responsiveness found in V3 and V3A, and makes of CIP the ideal intermediate stage toward the grasping-based object representations of AIP (Sakata et al. 1999; Shikata et al. 2001; Bray et al. 2013; Konen et al. 2013).

Another area that projects to CIP is the lateral intraparietal sulcus LIP, which performs distance and location estimation of target objects. More exactly, according to psychophysiological research in humans (Tresilian and Mon-Williams 2000), what is actually estimated and used in the parietal cortex is the reciprocal of distance, that is, nearness. In the intraparietal sulcus, distance and disparity are processed together, the former acting as a gain modulation variable on the latter (Salinas and Thier 2000; Genovesio and Ferraina 2004). This mechanism allows to properly interpret stereoscopic visual information (Dobbins et al. 1998; Mon-Williams et al. 2000), as described in Sect. 5.2.2.

5.1.2 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 (Landy et al. 1995). The main underlying principles that allow to achieve this goal seem to be two: cue reliability and cue correlation, or discrepancy (Tresilian and Mon-Williams 2000; Jacobs 2002).

Cue reliability is probabilistic, it depends on environmental conditions, on the estimate itself and sometimes on other, ancillary measures (Landy et al. 1995). Considering the case of interest for our research, 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 (Trotter et al. 1996). 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 (Backus and Banks 1999).

The available models for extraction and integration of visual cues usually focus on very specific aspects, such as disparity responsiveness with changing distance (Lehky et al. 1990; Lehky and Sejnowski 1990), conflicting stimuli (van Ee et al. 1999), maximum-likelihood cue integration (Hillis et al. 2004), temporal integration according to cue reliability (Greenwald et al. 2005), extraction of local surface slant (Jones and Malik 1992). Apparently, no published models on the subject provide details for practical implementation on robotic vision setups.

5.1.3 Object Recognition in the Ventral Stream

As pointed out in Sect. 2.4.1, object recognition in the ventral stream is performed gradually and hierarchically (Grill-Spector et al. 1998; Bar et al. 2001). Recent findings indicate that object recognition is composed of at least two subsequent stages, categorization and identification (Grill-Spector and Kanwisher 2005). In the first stage, an object is classified as belonging to a given class or family of objects, and such process is strikingly fast. The classification delay is so short that there is probably time to feed category information to the dorsal stream, for improving the online estimation of action-related features. This mechanism is represented by the link projecting from area V4 to CIP in Fig. 5.1. As pointed out in Sect. 2.3.1.3, anatomical and functional evidence supports this early integration between the streams (Perry et al., 2014). The second stage of object recognition is proper identification, performed by LOC, in which object identity is recognized within its category.

A second aspect, relevant for modeling purposes, is the method employed by the ventral stream for performing object recognition (Ullman 1996). At least for the first classification stage, visual input is very likely compared to memorized 2D representations (Bülthoff et al. 1995; Orban et al. 2006), and complex objects are identified by composing simpler features (Thoma and Henson 2011). A classification based on 3D representations would require mental rotation, and this can hardly be performed with the quickness observed in the experiments of Grill-Spector and Kanwisher (2005). Moreover, the consistent preference of some “canonical” views during free and classification-oriented object exploration indirectly supports the existence (if not the dominance) of 2D object representations (Bianz et al. 1999; James et al. 2001).

Various biologically inspired methods for object recognition have been developed in computer vision, and different models of ventral stream processing are available (O’Reilly and Munakata 2000; Riesenhuber and Poggio 2000; Rolls and Webb 2014). Some of them are strongly inspired by neuroscience findings, and use plausible approaches such as radial basis function networks (Pouget and Sejnowski 1997; Deneve and Pouget 2003) or a temporal coherence principle in unsupervised learning (Einhäuser et al. 2005). For the purposes of this work, object recognition is functional to grasping actions, and the interest is not in detailed modeling of ventral stream mechanisms. A simple viewpoint invariant classification is implemented, based on basic 2D global object representations (see Sects. 5.2.3 and 5.4.2).

5.1.4 Orientation Estimation in Artificial Vision and Robotics

Object orientation (or slant) estimation is a common, and difficult, problem in artificial vision (Lippiello et al. 2006). Nevertheless, no research works similar to the proposed approach are available in the literature.

A detailed overview of existing techniques for pose estimation can be found in Goddard (1998). 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, 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 (Rosenhahn et al. 2004). Methods of this kind have been proved useful even with moving targets (Lippiello et al. 2001). Often, the use of markers substitute explicit modeling (Gehrig et al. 2006). These techniques can be combined with others, where appearance based methods are used for the rough initial estimate and followed by a refinement step using model based technique (Ekvall et al. 2003). In Peters (2004) the rough initial estimate is determined on the viewing hemisphere as an initial guess, and then also refined. A model based approach can also be connected with range images, for example matching a 3D model to a range representation of the scene (Germann et al. 2007). 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 Weigl et al. (1995), are interesting but unrelated to the current approach.

For what concerns stereo slant estimation inspired on human physiology, Ferrier (1999) describes a method based on disparities which makes use of a model for computing orientation of features. With the support of camera calibration, which is not used in this work, they obtained similar results. Regarding the integration of stereoptic and perspective cues in artificial vision, although the idea is not novel (Clark and Yuille 1990), there are very few robotic platforms that make use of both visual cues at the same time. For example, in Saxena et al. (2007) 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. Other works, focused on object tracking (Taylor and Kleeman 2003) and on visual servoing (Kragic and Christensen 2001), perform cue integration, but their visual analysis is model-based, and their goal is feature matching and not feature extraction.

5.2 A Model of Distance and Orientation Estimation of Graspable Objects

This section introduces a proposal for distance estimation based on proprioceptive vergence data, and two different orientation estimators obtained from stereoscopic and perspective cues. A hierarchical approach to object classification is also presented.

5.2.1 Distance Estimation Through Proprioceptive Data

The distance of a fixated object from the viewer can be estimated by either retinal and/or proprioceptive cues, *accommodation and vergence* (see box).

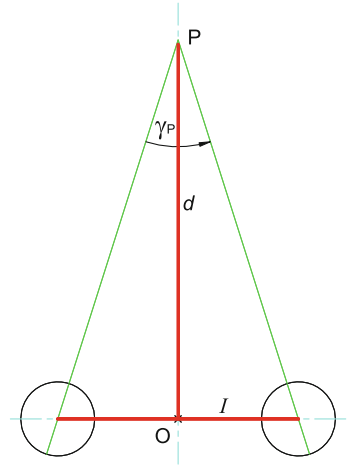
The movement performed by the eyes in order to converge on a given visual target is called *vergence*, or *convergence*. The resultant angle is called *vergence angle*. The adaptation of the shape of the eye crystalline lens in order to change the eye focus is called *accommodation*. Accommodation and vergence are both directly related to the distance of the visual target, and are linked by a reflex. In distance estimation, accommodation and vergence are preferably used when retinal data are not available or considered not reliable, and for short distances (Mon-Williams and Tresilian 1999; Tresilian et al. 1999).

The relation between distance and vergence angle γ_P is simple and depends only on the interocular distance I , which is constant (see Fig. 5.2). The distance d between the fixated point P and the cyclopean eye O, middle point between the two eyes, is given by:

$$d = \frac{I}{2 \tan(\gamma_P/2)} \quad (5.1)$$

Psychophysiological experiments (Tresilian and Mon-Williams 2000) suggest that distance estimation is most probably performed in the human brain using *nearness* units instead of distance units. Nearness is the reciprocal of distance, and a point at infinite distance has 0 nearness. The nearest distance at which vergence can be maintained, called the *near point of vergence*, is usually between 60 and 70 mm (Brautaset and Jennings 2005). Average interocular distance for adults is considered to be between 63 and 65 mm, and thus approximately coincident with the near point of vergence. Setting $I = d$ yields a maximum vergence angle of: $\gamma = 2 \cdot \arctan(I/(d \cdot 2)) = 2 \cdot \arctan(1/2) = 0.927 \text{ rad} = 53^\circ 8'$. The following expression for computing nearness from vergence hence produces nearness values between 0 (for $\gamma_P \rightarrow \infty$) and 1 (for $\gamma_P = 0.927 \text{ rad}$, the maximum vergence angle):

Fig. 5.2 Relation between vergence angle γ_P and distance d . I is the interocular distance, O the position of the cyclopean eye



$$nearness = 2 \tan(\gamma_P/2) \tag{5.2}$$

Such measure is more precise for close distances, and thus especially suitable for dealing with objects in the peripersonal space. Moreover, it is based on the relation between I and the near point of vergence, and does not depend on any constant or auxiliary measures.

Two radial basis function (RBF) networks were designed, for learning the association between vergence and nearness and between vergence and distance. The results can be seen in Fig. 5.3. On the top left, the distance/vergence curve corresponding to (5.1) is shown. Equation (5.2) between vergence and nearness is depicted on the top right of the image, and the corresponding learnt curve appears on the bottom right of Fig. 5.3 (lighter, dashed curve). The reciprocal of the learnt relation is finally depicted on the bottom left, where it can be compared with the true mathematical relation (they practically coincide). In this simplified example, to obtain a similar performance in the estimation of distance, the distance/vergence network requires 11 RBF units, while the nearness/vergence net requires only 4 neurons. This is not surprising, considering the approximate linearity of the relation vergence/nearness, and considering that the brain often employs an economy principle, minimizing the resources required to perform a given task. In the current model, object distance is represented in nearness units, and is used in the next section to modulate the effect of disparity on orientation estimation.

5.2.2 Object Orientation Estimation Through Retinal Data

For estimating object pose, humans combine estimators provided by different visual and proprioceptive cues, both binocular (mainly horizontal and gradient disparity

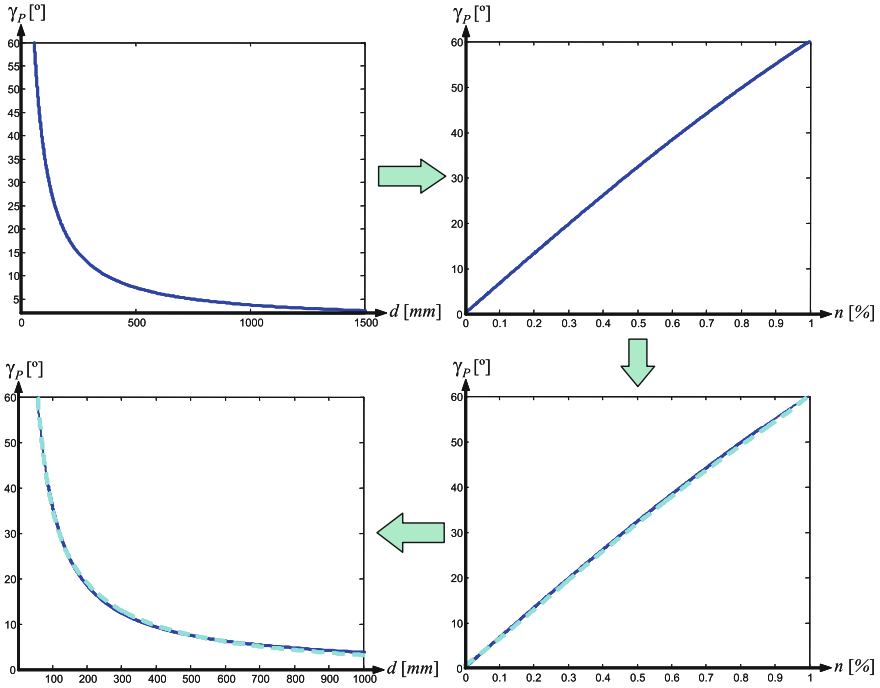


Fig. 5.3 Distance to vergence and nearness to vergence relations: theoretical equations (*upper diagrams*) and learnt curves (*superposed dashed lines in the lower diagrams*)

cues) and monocular (texture or edge perspective cues) (Backus et al. 1999). In this section, the problem of orientation estimation according to different cues is analyzed. First, a couple of expressions both neurologically plausible and useful for a practical implementation are derived. Next, the neural network architecture implemented for the solution of the problem is introduced. Finally, some results which allow to discuss the theoretical and practical implications of the proposed approach are described.

The proposed object orientation estimation process makes use of simple visual information for achieving a geometric 3D selectivity similar to that observed in neuroscience studies. The goal is to develop a modular computational structure, composed of various estimators, which makes use of proprioceptive and retinal cues in order to obtain the geometrical parameters needed for grasp planning. This approach differs from related research (e.g. Jones and Malik 1992) in that it builds upon retinal data: instead of using pixelated images and projective matrices, the only inputs are retinal angles and proprioceptive eye data. The center of the coordinate system is the cyclopean eye, as for humans.

For orientation and basic shape discerning, the approach relies upon one monocular information source, that is, perspective under the assumption of edge parallelism, and one kind of binocular information, width disparity. As explained in the previous section, these data are coded by visual areas V3 and V3A and combined in the

posterior intraparietal cortex CIP. One basic assumption is that objects recognized as boxes or cylinders have actually straight, parallel edges, and are laying on a horizontal table. This is very plausible 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 (Tsutsui et al. 2001) and humans (Brouwer et al. 2005) have shown that, even for purely perspective pose estimations, a frontoparallel trapezoid is usually interpreted as a rectangular shape slanted in depth.

Next, we analyze the sort of computation performed by the human brain during orientation estimation, in the binocular and in the monocular case, and propose plausible transfer functions to obtain estimators from simple retinal angles.

5.2.2.1 Stereoscopic Slant Estimation

In Fig. 5.4a a viewing scene is seen from above: object PQ of length l is slanted about a vertical axis with an orientation θ . Its extreme P is the fixation point, placed straight ahead from the cyclopean eye (in this way γ_P corresponds to the vergence angle). 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 the binocular separation of points P and Q (being $\psi_P = 0$).

The change in slant of segment PQ as point Q moves on the xz reference system can be observed in Fig. 5.5. In the graph, the position of P is fixed at $(x_P = 0, z_P = 30)$. The dot represents θ for Q positioned as in Fig. 5.4a.

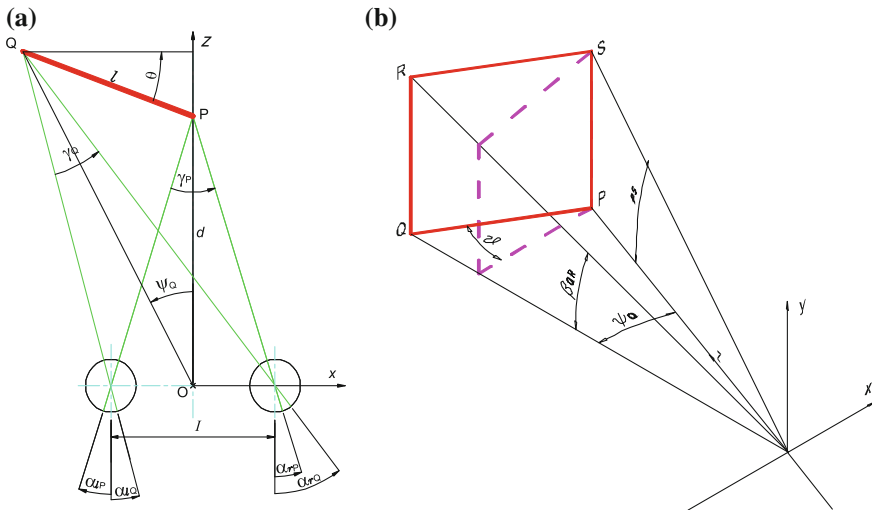
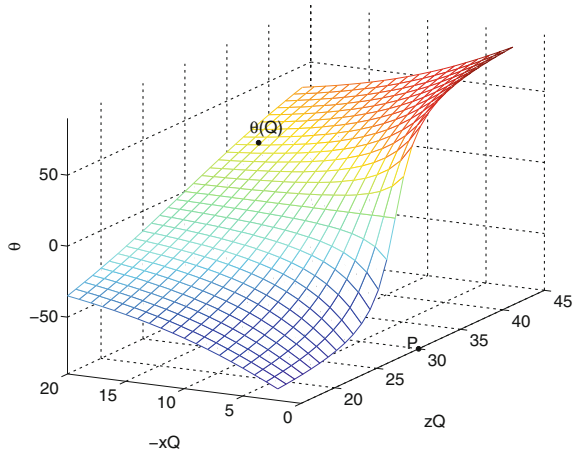


Fig. 5.4 Schemes for deriving slant from stereopsis and perspective. **a** Stereopsis. **b** Perspective

Fig. 5.5 Slant θ as a function of the position of point Q in the xz space



The horizontal slant θ of an object can be computed only from retinal angles using the following expression, which can be derived from Fig. 5.4a:

$$\tan \theta = \frac{(\tan \alpha_{rQ} - \tan \alpha_{lQ}) - (\tan \alpha_{rP} - \tan \alpha_{lP})}{\tan \alpha_{lP} \tan \alpha_{rQ} - \tan \alpha_{lQ} \tan \alpha_{rP}} \quad (5.3)$$

Reminding that P is the fixation point, so that $\alpha_{lP} = -\alpha_{rP} = \gamma_P/2$, the equation can be simplified in this way:

$$\tan \theta = \frac{1}{2 \tan(\gamma_P/2)} \cdot \frac{(\tan \alpha_{rQ} - \tan \alpha_{lQ}) - (\tan \alpha_{rP} - \tan \alpha_{lP})}{(\tan \alpha_{rQ} + \tan \alpha_{lQ})/2} \quad (5.4)$$

Recalling (5.2) and the definitions in the disparities text box, this relation can be expressed by using only quantities that are actually computed in the visual brain areas:

$$\tan \theta = \frac{1}{nearness} \cdot \frac{relative\ disparity}{separation} \quad (5.5)$$

Separating θ , a biologically plausible stereoptic orientation estimator $\hat{\theta}_S$ is obtained:

$$\hat{\theta}_S = \arctan \frac{relative\ disparity}{nearness \cdot separation} \quad (5.6)$$

The interpretation of (5.6) is that the component due to disparity (the fraction relative disparity/separation, which is also called disparity gradient) is modulated by the viewing distance (or nearness), as indicated by neuroscience research. Nearness is probably computed from proprioceptive data, as discussed in the previous section.

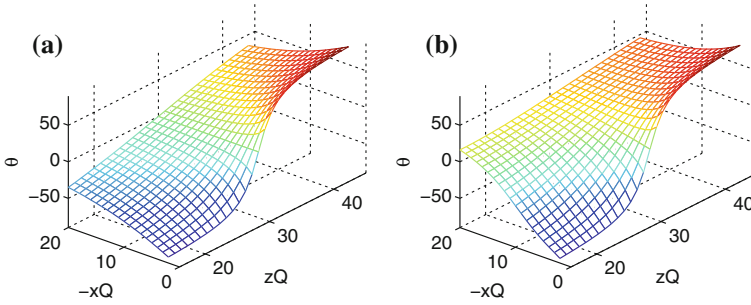


Fig. 5.6 Distorted estimation of θ (b) obtained by approximating $\tan \alpha$ with α

The separation which appears in the formula is binocular, referred to the cyclopean eye.

Other works (Banks et al. 2001; Howard and Rogers 2002) propose expressions similar to (5.6), starting from slightly different assumptions. It is important to point out though that common approximations found in the literature carry to unacceptably wrong estimations in most cases. An example of this can be seen in Fig. 5.6, where an exact reproduction of Fig. 5.5 (Fig. 5.6a), obtained with (5.6), is compared to a distorted one (Fig. 5.6b), obtained with the same equation in which the common solution of approximating the exact value of $\tan \alpha$ with α is employed. The comparison demonstrates that for a real application, and a faithful model, exact expressions need to be used.

5.2.2.2 Perspective Slant Estimation

The slant of an object can be estimated using only monocular data, as depicted in Fig. 5.4b, in which the origin of the axes is one of the eyes. The frontal object face is considered as rectangular, exploiting the reasonable assumption of parallelism and equality of opposite edges (PS and QR in the image). The angles β in the figure represent the vertical retinal angles associated to such edges. The function which leads from retinal angles to orientation estimation can be derived from the draw, and can be referred entirely to either the left or the right eye:

$$\tan \theta = \frac{\tan \beta_{QR}}{\tan \beta_{PS} \sin \psi_Q} - \frac{1}{\tan \psi_Q} \tag{5.7}$$

In this case the monocular separation is: $\psi_Q = (\alpha_Q - \alpha_P)/2$.

Approximating $\sin \psi_Q$ to $\tan \psi_Q$, which is plausible for reasonably small separations, a formula for perspective estimation of θ is obtained:

$$\hat{\theta}_P = \arctan \frac{\text{perspective disparity}}{\text{separation}} \tag{5.8}$$

Perspective disparity is the quantity $\frac{\tan \beta_{QR}}{\tan \beta_{PS}} - 1$, which represents the proportion between the projected sizes of edges QR and PS. Therefore, again, the estimator depends on a separation factor and a disparity factor, this time monocular.

Equations (5.8) and (5.6) will be used, both separately and merged, for simulated (Sect. 5.3) and real orientation estimation on a robotic setup (Sect. 5.4).

5.2.3 Hierarchical Object Classification

The approach to object classification proposed in the model is composed of a three stages process. These stages are initial shape classification, proper object recognition and actual identification of a known object.

1. **Shape classification.** In this stage the target object is classified into one of a number of known classes. For example, a bottle would be classified in the class of cylinders. Simple visual information such as shape silhouette or a basic topographic relation between object features is enough for this task. No actual data regarding the size and the proportion of the object are considered. Nothing is inferred at this point about object composition, utility, meaning. The information recovered at this stage is used by early areas of the dorsal stream in order to estimate the size and pose of the object.
2. **Object Recognition.** Actual object recognition is the goal of this stage. The target object is identified as if the task was to name it. What was a general cylindrical shape in the previous stage is now identified as a bottle. Additional conceptual knowledge is thus added to the previous basic information. Composition, roughness, weight of the object can be inferred if not known for sure. The object proper use in different tasks is also recalled at this point. Object recognition directly affects the process of grip selection, providing a bias toward grasp configurations better suited to the object weight distribution, possible friction and common use.
3. **Object Recall.** In this final stage, a subject recalls a single well-known object which was encountered, and possibly grasped, before. Going back to the cylinder example, here it can be recognized as a wine bottle recently bought, and thus previously known and dealt with by the subject. Compared to the previous one, this stage adds security to the estimation of the object characteristics. To recognize an object as a bottle helps in estimating its weight, whilst to identify a previously encountered bottle provides an exact value of that weight.

In all stages, the classification process has to be viewpoint invariant. A very important issue is that object classification and recognition is always a gradual process, not a binary one, and each classification is accompanied by a confidence value, necessary to clarify its reliability. Any classification having a low confidence should be used prudentially, and if no class or object are clearly identified the system should rather provide a failed classification answer, to clarify that the situation is uncertain and needs further exploration. Feedback from execution outcome can later be used to complete and improve the world knowledge in these situations.

5.3 Neural Network Implementation of a Multiple Cue Slant Estimator

A neural architecture for estimating the orientation θ of a target object according to the concepts described in the previous section has been implemented. The framework includes two sets of neural networks, for stereoscopic estimation and for monocular estimation based on perspective data.

5.3.1 Neural Network Estimators

The whole framework of the neural network implementation is depicted in Fig. 5.7, where the nets are associated to the brain areas that probably perform their functions. Apart from nearness estimation, implemented with the RBF network described in Sect. 5.2.1, all networks are feedforward backpropagation, trained with the Levenberg-Marquardt algorithm. Four neural networks constitute the module for orientation estimation based on stereopsis: they compute nearness, relative disparity and separation (two nets represented as a unique one in the scheme), and the final estimate of $\hat{\theta}_S$ from the outputs of the previous three networks (according to expression (5.6)). The module for orientation estimation based on perspective makes use of two networks for computing the two components of (5.8), and a third for the final calculation of $\hat{\theta}_P$.

5.3.2 Merging the Estimators

Following the insights provided by the neuroscience literature, the final orientation estimator is computed by combining the stereoscopic estimator $\hat{\theta}_S$ and the perspective

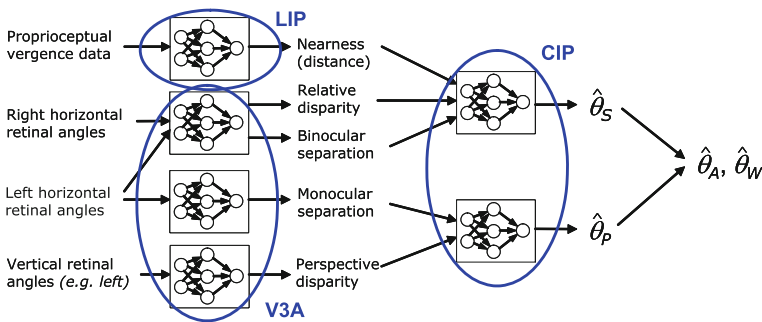


Fig. 5.7 Scheme of the neural network architecture for slant estimation

estimator $\hat{\theta}_P$. A first merging can be done through a simple average of the two output values:

$$\hat{\theta}_A = \frac{\hat{\theta}_S + \hat{\theta}_P}{2} \quad (5.9)$$

Looking for better performances and improved biological plausibility, the two principal driving factors for cue merging, correlation and reliability, have been taken into account. In the present case, cue correlation could not be used, as only two different estimators are available. The chosen solution was thus to experimentally simulate cue combination using only cue reliability. The stereoptic and perspective estimators were trained to learn how their reliability changes in different conditions. According to the literature, the driving factors for the accurateness of orientation estimation are distance and orientation itself (this is not a contradiction: the estimated value can be used as output and, at the same time, as reliability index for the estimation). In fact, although it is known that stereopsis quickly loses its reliability with distance, here the interest is on the near space defined by the arm reaching distance, within which the variation of distance affects the two methods in similar ways. For this reason, the focus is rather put on the effect of orientation, and the goal is to devise a merging method that optimizes the weights given to the two estimators when changing the estimated value of θ .

How the human brain can predict cue reliability is still a matter of debate. Nevertheless, it has been shown that stereoscopic and perspective cues are actually weighted through a maximum-likelihood process (Knill 2007). To emulate this process, the error patterns obtained with the estimation methods alone were saved, and used to generate a joint estimator which is a weighted average of the original ones:

$$\hat{\theta}_W = w_S \hat{\theta}_S + w_P \hat{\theta}_P \quad (5.10)$$

In (5.10), w_S and w_P are functions of θ computed in the following way:

$$w_S = \frac{SSE_P}{SSE_S + SSE_P}; \quad w_P = \frac{SSE_S}{SSE_S + SSE_P} \quad (5.11)$$

where SSE_S and SSE_P are the previously learnt summed squared errors of stereopsis and perspective respectively.

5.3.3 Results of the ANN Simulation

In principle, the neural network implementation allows to achieve any arbitrary precision in the estimation. The study of estimators reliability can thus be done either in the real world or simulating the effect of natural imprecisions introducing stochastic variability in the computation. Before the implementation on a real robotic setup, a

Table 5.1 ANN slant estimation results for different estimators

Method	Estimator	Error (°)
Perspective	$\hat{\theta}_P$	4.49
Stereopsis	$\hat{\theta}_S$	4.17
Simple average	$\hat{\theta}_A$	3.05
Weighted average	$\hat{\theta}_W$	2.93

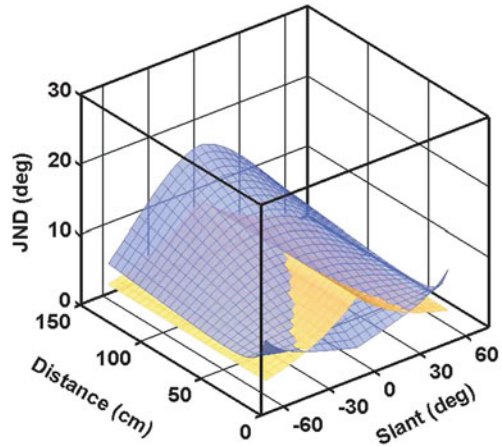
simulation was performed to study the effect of noise on slant estimation performed with stereoptic and perspective methods. With this purpose, random noise was added to all retinal angles which constitute input values for the nets. The error representing the average difference between estimation and true value was calculated all over the test space ($10^\circ < \theta < 70^\circ$, $450 \text{ mm} < d < 850 \text{ mm}$).

In Table 5.1 the improvement of joining the two estimators in this way can be observed. For comparison, consider that the nets were trained so that the average error of the two original estimators $\hat{\theta}_S$ and $\hat{\theta}_P$ before the introduction of noise was less than one degree. Although stereopsis seems to suffer less from the insertion of noise, the contribution of both perspective and stereoptic predictors is very important for improving the final result. In fact, the weighted average $\hat{\theta}_W$ allows to obtain an improvement of almost 30% on the best single cue estimator $\hat{\theta}_S$, suggesting that the combination of different cues is the best solution for pursuing a reliable estimation. Even the simple average $\hat{\theta}_A$, that can be used a priori without exploiting previous experience, improves the $\hat{\theta}_S$ performance by more than 25%. The performance difference between $\hat{\theta}_W$ and $\hat{\theta}_A$, which is rather small in this example, sensibly increases especially in the most extreme situations, when one of the estimators is much better than the other, and the simple average would not take this aspect into account.

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. Figure 5.8, taken from Hillis et al. (2004), depicts the precision of two orientation estimators, perspective and disparity based, as a function of distance and slant. With increasing distance, both estimators become less reliable, but the stereoscopic cue (blue) is clearly more affected. The effect of orientation is more complex. Perspective methods are more sensitive and precise for pronounced slants, that generate higher differences in vertical disparities. At long distances, disparity methods also prefer high slants. On the contrary, for the short distances typical of grasping actions their error is minimum for low slant values, which grant higher binocular disparities.

To check if this pattern of behavior could be reproduced in the simulation, the estimators accurateness was plotted as a function of distance d and as a function of orientation θ . The outcome can be observed in Fig. 5.9, in which the error in stereoptic and perspective estimation is plotted against orientation (Fig. 5.9a) and distance (Fig. 5.9b). The similarity of the obtained results to what is described in the literature is remarkable, as can be observed by comparing the corresponding ranges of Figs. 5.8 and 5.9. Notice that, being Fig. 5.8 symmetrical with respect to slant, in Fig. 5.9a only positive slants are plotted, and Fig. 5.9b considers just reachable

Fig. 5.8 Precision of texture (*orange*) and disparity (*blue*) cues as a function of distance and slant. JND is the Just Noticeable Difference, corresponding to the smaller detectable slant variation. From Hillis et al. (2004)



distances, up to 850 mm. The proposed model looks thus appropriate to reproduce the behavior of stereoptic and perspective estimators.

The second effect that could be reproduced is the improved performance obtained through a maximum likelihood merged estimator in which cues are weighted according to their reliability (experimentally learnt), as explained in Sect. 5.1.2. The better results obtained with the weighted estimator θ_W can also be observed in Fig. 5.9.

The implemented neural architecture hence constitutes an orientation estimator both biologically plausible and practically reliable. The quantities used are employed by the human visual system, but also computationally useful for artificial implementation (e.g. retinal angles). The proposed equations for computing orientation from stereopsis and perspective are plausible transfer functions useful to model the estimation process. The trained neural networks are somehow emulating the behavior of modules pertaining to higher visual brain areas. Indeed, inputs and intermediate results represent quantities that have been observed and measured, and are part of real brain processes (Welchman et al. 2005). This suggests that functions (5.6) and (5.8) are plausible models for stereoptic and perspective slant estimation in the human cortex.

The simulation results indicate that the proposed approach can be suitable for improving the reliability of a real application. The next immediate step is the practical experimentation on a robotic platform.

5.4 Robotic Validation

Orientation and pose estimation are very complex problems in machine vision, especially when the goal is to develop a reliable robotic system which makes use of visual estimates to interact with the environment, such as in object grasping actions (Wandell 1995; Trucco and Verri 1998). In this section, the computational method

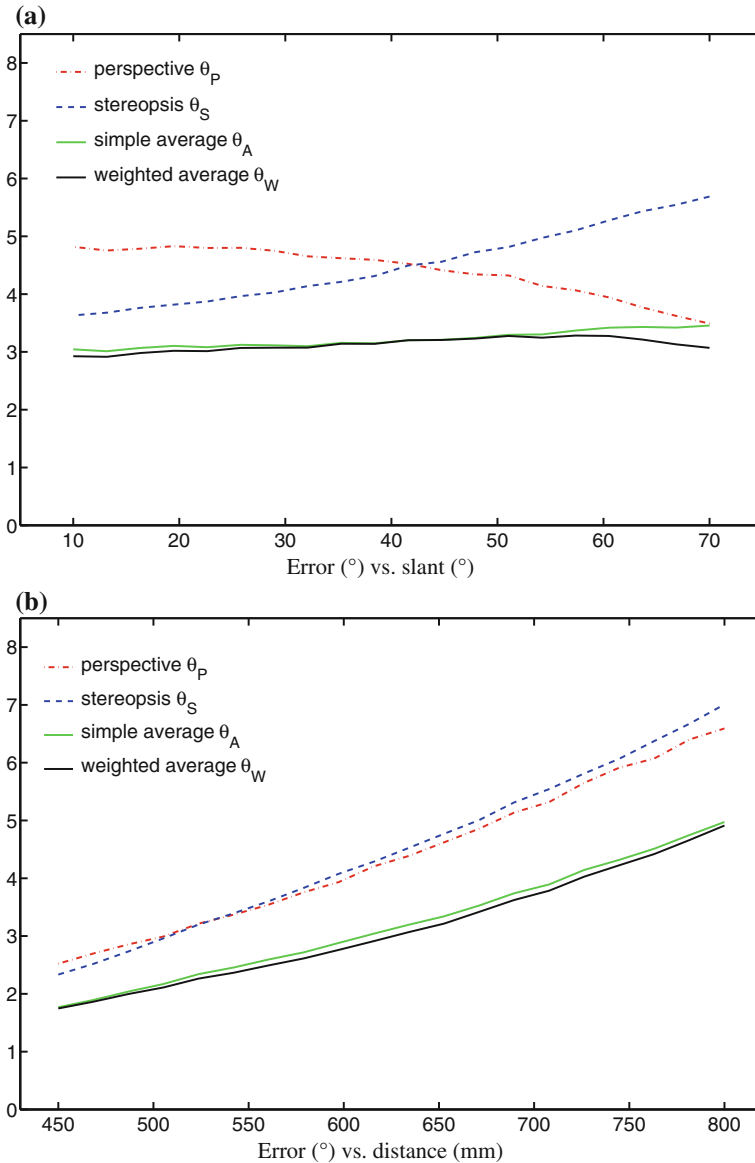


Fig. 5.9 Slant estimation error as a function of slant and distance; neural network simulation (for distance the value is the class lower bound). **a.** Error (°) versus slant (°). **b** Error (°) versus distance (mm)

described above is implemented on a robotic setup. The goal is to obtain an orientation estimator robust and reliable enough to be used in vision-based robotic grasping. A number of different experiments to verify how the ideal results change when the

model has to face the uncertainties of the real world are executed. As a second goal, the implementation tries to reproduce the effects obtained with the ANN simulation with real experimental data, and hence further validate the model.

5.4.1 Robotic Setup

The robotic setup, shown in Fig. 5.10a, consists of a seven degrees-of-freedom (DOF) Mitsubishi PA-10 arm endowed with a Barrett Hand and a JR3 force/torque and acceleration sensor mounted at the wrist, between hand and arm. A stereoscopic, black and white camera Videre Design is coupled to the wrist, eye-in-hand style (Fig. 5.10b). This configuration allows for controlled movements of the vision system without the need of a pan-tilt-vergence robotic head.

The Barrett Hand (see schema in Fig. 5.11) has three-fingers with a total of four controllable degrees of freedom. Each finger possesses two joints which are driven by a single motor. The controlled variables are thus the three finger extensions e_1 , e_2 and e_3 . The fourth degree of freedom controls the opening angle θ of fingers 2 and 3, which are symmetrically placed on either side of finger 1, the *thumb*, which is fixed. When fully abducted, for $\theta = 0^\circ$, fingers 2 and 3 oppose the thumb, when adducted ($\theta = 180^\circ$) they flex in parallel to the thumb.

As it can be observed in Fig. 5.10b, the hand fingertips are equipped with arrays of pressure sensors, designed and implemented by Weiss Robotics (Weiss and Wörn 2004). The sensors are 8×5 cell matrices that cover the inner parts of the distal phalanges of the fingers. Each sensor is able to detect a complete two dimensional force profile by the use of a homogeneous material connected to an adequate electrode matrix.

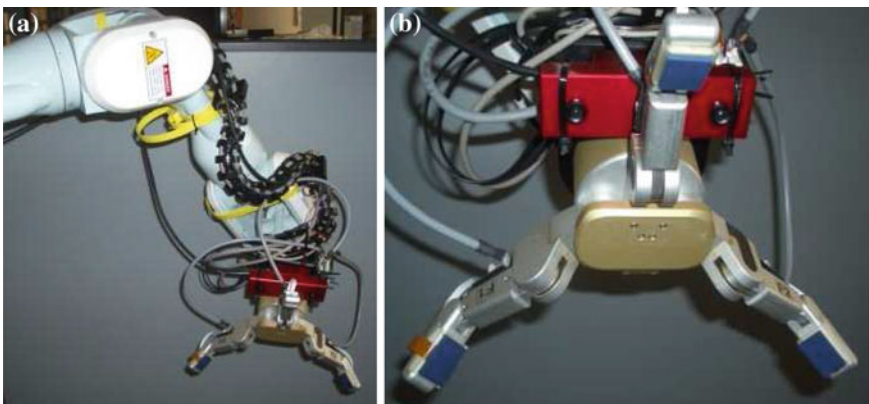
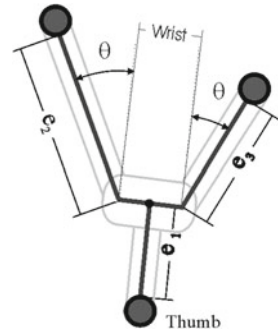


Fig. 5.10 Robotic setup with arm, hand and stereoscopic camera. **a** Robotic arm and hand. **b** Detail of hand with stereo camera

Fig. 5.11 Barrett hand kinematics



The robot world is a dark environment in which clear shapes are placed on a table at variable positions and orientations (see Fig. 5.12). The range of possible positions are those that allow to view the object and also keep it at reaching distance for the hand. Using the estimators previously introduced, the system is able to estimate distance, pose and size of objects without using explicit models, but only common knowledge regarding basic shapes it recognizes, such as the assumption of edge parallelism.

The grasping action begins with the stereo camera facing straight ahead, and having an object in its field of view. Both left and right images are continuously binarized and the object contour tracked. The choice of object and background color is driven by the need of keeping image processing as fast and lean as possible. The point in the image having minimum y coordinate, called P , is selected as reference and starting point for the contour, and one of the images is centered on it. Let us

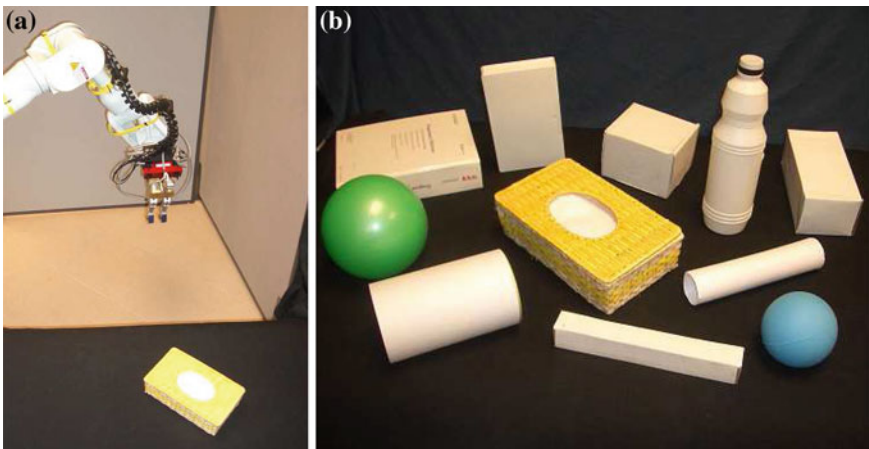


Fig. 5.12 Workspace with robot fixating an object and possible target objects as seen from the robot camera. **a** Robot at fixation position. **b** Workspace with examples of target objects

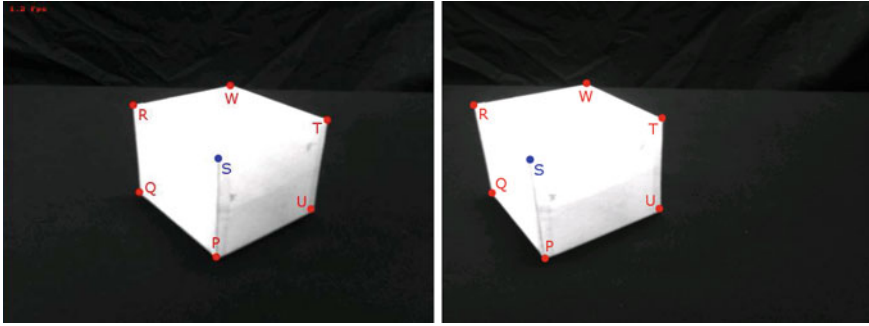


Fig. 5.13 *Left and right object images from the initial position, with labels of detected corners*

assume from now on that processing begins from the left eye, hence point P is centered on the left image (see Fig. 5.13 in Sect. 5.4.3).

The images at this position are processed by two parallel modules, one concerned with classifying the object in a number of known categories, the other dedicated to pose estimation. The first module, emulating the processing of the medium ventral stream, makes use of a global visual representation of the object in order to perform a viewpoint invariant classification. The second module integrates different cues for estimating object distance, size and pose.

5.4.2 Object Classification Experiments

The object classification module has to categorize objects seen from different poses and distances. With this purpose, it has to consider object images globally, rather than focusing on local features. The goal is to classify an object as pertaining to one of three known object classes: parallelepipeds (boxes), cylinders and spheres. This has to be done using only a couple of stereo images, without changing the viewpoint. Moreover, it is important to retrieve a value measuring the confidence in the classification, represented by the percentage of likeliness assigned to each class. Two different approaches were tested, using the extracted object contour as a silhouette of the object.

The first tested method consisted in computing a chain code of the contour, which constitutes a representation that is invariant with respect to size and distance, while maintaining the feature topology necessary to identify the object. The chain was generated extracting a pre-defined, finite number of points regularly spaced along the contour, starting from P . The code c_i corresponding to point P_i is thus the following, normalized so that the range is $[-1, 1]$:

$$c_i = \text{atan} \left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}} \right) \quad (5.12)$$

The number of points to use can be chosen according to the application. The selection of 20–30 points gave the best results. The chain codes of different objects from different points of view constituted the training data. A probabilistic neural network was used to classify the object in one of the three classes. For training, 10 different objects were used, each seen from 19 different positions (apart for the spheres).

This solution did not provide the required behavior. In fact, results on training objects (from different viewpoints) and on different test objects gave recognition success very close to 100%, but test objects were often misclassified. Moreover, even in the wrong cases, confidence was always very high, often above 98–99%. The conclusion is that the method is very good in recognizing known objects, but not in generalizing. The sequential order of different object features, like straight and curved segments, or corners, would be enough for classification. Instead, the chain code representation takes into account and hence classify objects also according to the feature length, distinguishing for example a short cylinder from a long one. Moreover, classification should be much more shaded, with confidence percentages not always close to 100%. As justified in Sect. 5.2.3, a missed classification due to high uncertainty is preferred to a wrong categorization.

For these reasons, a different classification method was tested, based this time on the curvedness of objects. This method is based on only one index representing each object, the curved fraction of its contour, ratio between the length of its curved features and the total contour length. For the shapes in use, experimental data showed that parallelepipeds, cylinders and spheres normally possess linearly separable curvedness values. The classification process begins with a training phase during which the system is presented with five different boxes (B), three cylinders (C) and two spheres (S), again from 19 viewpoints distributed along a 90° range. Average curvedness values μ_K and corresponding standard deviations σ_K are calculated for the three classes, $K \in \{B, C, S\}$.

Given a test point c_i , the curvedness coefficient of object i , its degree of membership m_{iK} to class K is computed as the reciprocal of the relative distance to the class center:

$$m_{iK} = \frac{\sigma_K}{|c_i - \mu_K|} \quad (5.13)$$

At this point, classification percentages for the three classes $K = B, C, S$ are given by:

$$p_{iK} = \frac{m_{iK}}{m_{iB} + m_{iC} + m_{iS}} \quad (5.14)$$

As explained above, a missing recognition response is better than a misclassification. To favor the former over the latter, a high confidence value of 70% is required to assign the object to any class. If no class reach this value, the object is not classified. In such cases, only distance and approximated center of mass (that is in reality the centroid of the visible 2D silhouette) can be estimated and used for grasping.

An exception is the case of uncertainty between boxes and cylinders. If the sum $p_{iB} + p_{iC} > 70\%$, then the object is classified in the less restrictive class, i.e., as a cylinder. For cylinders, only one face can be computed for slant estimation, while for boxes two visible faces are used. A misclassification of a cylinder as a box would thus provide a wrong orientation estimation, whilst a misclassification of a box as a cylinder would just imply that some available information is not used.

Classification results for objects in the training set are provided in Table 5.2. Cases of misclassification are highlighted in bold whilst uncertain cases are marked in italics. For the training set, only two problematic cases are identified, both for cylinders seen from a 0° angle (objects 5 and 6). It is not surprising that this is a difficult condition for the recognition system, as the contour provides limited if any information on curvature, and more elaborate methods which take into account shading would be required for proper classification.

Classification results for test objects are given in Table 5.3. Most cases of missing classification regard the problem observed for the training set. Cylinders seem to be difficult to recognize, especially for extreme viewing angles, in which their silhouette appears as a rectangle or as a circle. Nevertheless, the prudential decision of assigning the object to class *C* in case of uncertainty between box and cylinder, works in nearly all conditions, and provide reliability to the whole pose estimation system. Only objects 14 and 16 from the 0° viewpoint are finally misclassified, the first as a sphere and the second as a cylinder. Object 18 cannot be clearly put in any of the three classes, but it has one face that can be used for slant estimation, as cylinders, hence its classification as a cylinder is the most appropriate from a practical point of view.

5.4.3 Object Pose and Distance Estimation

5.4.3.1 Orientation Estimation

For what concerns pose estimation, this requires the extraction of features as those used in the model. For this reason, a number of salient points on the contour have to be extracted. Object classification biases this process, modeling the influence of ventral stream data on dorsal stream processing, and the contextual nature of 3D perception (Todd 2004).

For boxes, the salient points usually correspond to the object corners (Fig. 5.13). Object faces are not segmented separately, so the number of detected corners ranges from 4 to 6 depending on point of view and object pose. Possible missing points are added according to the shape class. For example, if a box is detected by the classification process and, due to a bad perspective position only five points of the contour are chosen as corners, the sixth will be set according to simple geometric considerations. For a cylinder, only four points are necessary, as those on the curved parts of the contour are not used. For spheres only centroid and apparent diameter are computed.

Table 5.2 Object classification percentages for different slants; training shapes




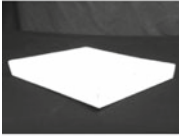

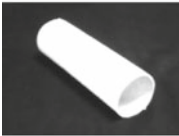



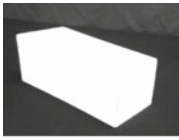
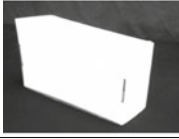
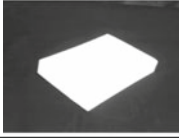
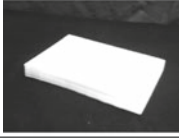





#	Object	Class	0°	30°	60°	90°
1		Box	98.16	86.63	84.81	94.90
		Cylinder	1.64	11.57	13.12	4.44
		Sphere	0.20	1.80	2.07	0.66
2		Box	92.97	85.90	84.81	91.22
		Cylinder	6.13	12.19	13.12	7.63
		Sphere	0.90	1.91	2.07	1.15
3		Box	93.94	84.81	84.84	87.25
		Cylinder	5.29	13.12	13.10	11.04
		Sphere	0.77	2.07	2.06	1.71
4		Box	99.87	86.88	84.81	99.23
		Cylinder	0.11	11.36	13.12	0.68
		Sphere	0.02	1.76	2.07	0.09
5		Box	86.20	0.59	0.26	0.81
		Cylinder	12.44	98.65	97.91	92.94
		Sphere	1.36	0.76	1.83	6.25
6		Box	58.07	1.68	20.81	0.35
		Cylinder	38.70	96.85	74.97	97.96
		Sphere	3.23	1.47	4.22	1.69
7		Box	2.74	2.42	0.63	9.01
		Cylinder	95.19	95.73	94.59	88.51
		Sphere	2.07	1.84	4.77	2.48
8		Box	0.48			
		Cylinder	25.46			
		Sphere	74.07			
9		Box	0.37			
		Cylinder	24.23			
		Sphere	75.40			

Table 5.3 Object classification percentages for different slants; test shapes

#	Object	Class	0°	30°	60°	90°
10		Box	98.83	85.80	85.07	85.55
		Cylinder	1.05	12.53	13.16	12.75
		Sphere	0.12	1.67	1.77	1.70
11		Box	94.60	89.97	85.07	91.08
		Cylinder	4.81	8.88	13.16	7.90
		Sphere	0.59	1.15	1.77	1.02
12		Box	80.49	96.20	86.02	91.72
		Cylinder	17.75	3.38	12.33	7.54
		Sphere	1.76	0.42	1.65	0.74
13		Box	94.48	95.59	90.59	99.43
		Cylinder	4.98	3.92	8.33	0.51
		Sphere	0.54	0.49	1.08	0.06
14		Box	0.59	5.86	0.33	0.51
		Cylinder	30.85	91.42	96.56	51.93
		Sphere	68.56	2.72	3.11	47.56
15		Box	60.35	61.66	35.89	0.32
		Cylinder	36.20	35.02	59.45	99.16
		Sphere	3.45	3.32	4.66	0.52
16		Box	57.89	84.91	93.33	98.05
		Cylinder	38.63	13.04	5.77	1.73
		Sphere	3.48	2.05	0.90	0.22
17		Box	0.82	1.03	0.42	0.71
		Cylinder	98.29	87.37	95.07	90.31
		Sphere	0.89	11.60	4.51	8.98
18		Box	17.89	0.20	3.69	3.81
		Cylinder	77.30	97.37	94.28	93.78
		Sphere	4.81	2.43	2.03	2.41

Even with this simplified setup, to reliably detect the salient points a double search is performed on the contour, combining the information given by different algorithms for corner (Teh and Chin 1989; Chetverikov and Szabo 1999) and edge detection (Ray and Ray 1995), to maximize the chance of finding all visible corners of the object when possible. A system able to segment the three faces of the object separately would provide a better estimate, but the results obtained with this simpler approach, presented below, match the application requirements.

Three variables identify object position and orientation. The distance d is measured between point P and the camera, and is enough to represent full 3D object location, as the object is centered in the image, and there is thus no lateral or vertical displacement. The other two variables are slant angle with respect to the frontoparallel position θ (see Fig. 5.4), 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. The viewing direction angle is restricted in the 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 2D 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 , only those situations that would reduce the interest of the slant estimation (for angles very close to 0° and 90°) are ruled out. These conditions 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 point P of the object is placed horizontally at the center of the image, in order to minimize distortions due to the cameras' optics. Left and right images at this position are then processed: corners P, Q, R, W, T and U are found as explained above, and the position of S is estimated through a two point perspective method (Fig. 5.13). At this point, the coordinates of the defining points are transformed into angles with respect to the center of the image, using the camera focal lens and image size in pixels as parameters. The non-linearity of the camera optics is the reason to avoid getting close to the image borders, where distortions could affect the transformation process.

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 Sect. 5.2.2: (5.8) is applied to the couples of segments PS/QR and UT/PS for both the left and right eye, whilst (5.6) is applied to segments PQ, SR, TS and UP. The first eight estimators, four perspective and four stereoscopic, of Table 5.4 are obtained at this point.

Before calculating the final, merged estimator it is useful to check for possible outliers (completely wrong estimations). In nature, bad estimations could be due to momentary occlusions, unusual light conditions, sudden movements, etc. In the simple setup used, any previous processing step can 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. Outlier detection is a full sub-branch of statistics (Rousseeuw and Leroy 1987), and many different methods

Table 5.4 Slant estimators

#	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, # 1–4
10	Merged ($\hat{\theta}_S$)	Stereopsis Only Average, # 5–8
11	Merged ($\hat{\theta}_A$)	$\hat{\theta}_P$ and $\hat{\theta}_S$ Simple Average, # 9–10
12	Merged ($\hat{\theta}_G$)	Global Simple Average, # 1–8
13	Merged ($\hat{\theta}_W$)	Global Weighted Average, # 1–8

are available. Various techniques were explored, and they did not give significantly different results. The classical Rosner's many outliers test (Rosner 1975), widely used in the literature for similar problems, was finally chosen. The best results were obtained for a significance level $\alpha = 0.01$, which gave a final estimation improved of more than 5% compared to the implementation without outlier rejection.

Following the model, monocular and binocular cues have to be merged according to their expected reliability and correlation. The starting point of the experiments is a situation in which no information is available regarding reliability of the different cues in the various working conditions. Therefore, 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, non-weighted average of a set of simple estimators (Estimators 9–12 of Table 5.4). The second is to compute an average in which weights are calculated using cue correlation (Estimators 13), in this case simply using the deviation of each cue from the simple average of all cues.

5.4.3.2 Nearness and Size Estimation

As no previous knowledge regarding the target object is assumed, it is not possible to disambiguate the pair distance/size only from retinal data. The nearness of the object can be calculated making use of expression (5.1), after estimating the proprioceptive vergence angle γ_P . The available stereo camera does not allow for vergence move-

ments of the eyes, so they have to be simulated. The simple procedure adopted is to center point P of the 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 left and right images are taken both from the initial and the final position, and they are considered as two independent slant estimation experiments. No significant differences were observed regarding the estimation precision from the initial and the final position.

For what concerns 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.

5.4.4 Experimental Results

Overall, 422 experiments were executed with different values of slant and distance, as shown in Table 5.5. The global average estimation errors of all executed experiments are provided in Table 5.6. 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 is quite apparent how the combination of multiple cues, especially 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, Stereopsis I; the global average $\hat{\theta}_G$ 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 quite a good pose estimation for a robotic system, even in these restricted conditions.

Table 5.5 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		
Total	422	Total	422

Table 5.6 Experimental slant estimation results, overall average errors

#	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

5.4.4.1 Comparison with Human and Neural Network Simulation Data

It is interesting to compare error distributions obtained in the real practical experiments with the theoretical ones of Sect. 5.3.3. Figure 5.14 shows the average error plotted as a function of slant (Fig. 5.14a) and distance (Fig. 5.14b), similarly to Fig. 5.9. Some slant and distance values are probably affected by the use of different objects and viewpoints, which were not regularly distributed across conditions; see for example the bad quality of stereopsis, and consequently of the merged estimators, for slant = 60°. Nevertheless, the trends are quite clear, and the expected effect of slant and distance on the different estimators is once again reproduced. In Fig. 5.14a 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. Figure 5.14b shows that stereoptic estimation gradually decreases its precision with distance, whilst perspective seems nearly uncorrelated with it, apart for extreme values. Again the weighted average presents a clearly advantageous behavior in all cases.

It can be noted from both graphs how the weighted estimator maintains its reliability across conditions. Error bars of θ_W are always small apart from extreme conditions (which are also affected by a reduced number of trials). Errors for other estimators, which could not be 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.

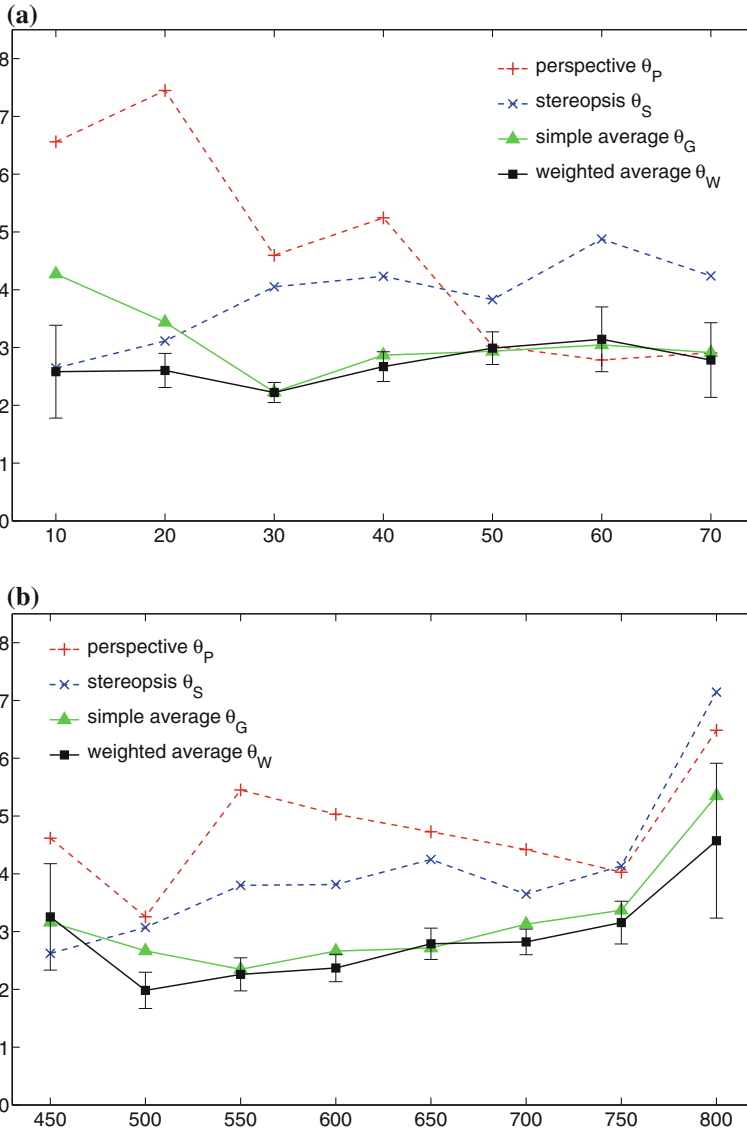


Fig. 5.14 Slant estimation error as a function of slant and distance; experimental results. For clarity, errors on errors are plotted only for θ_W . **a** Error (°) versus slant (°). **b** Error (°) versus distance (mm)

For what concerns distance estimation, the global average error for all experiments is of 33.4 mm, and the error distribution shown in Fig. 5.15, 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 that it makes use of two estimators and the

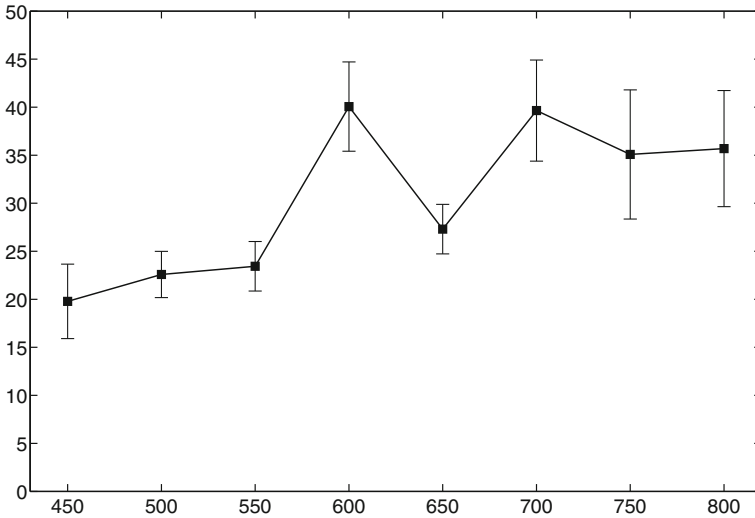


Fig. 5.15 Experimental distance estimation error; Error (mm) versus distance (mm)

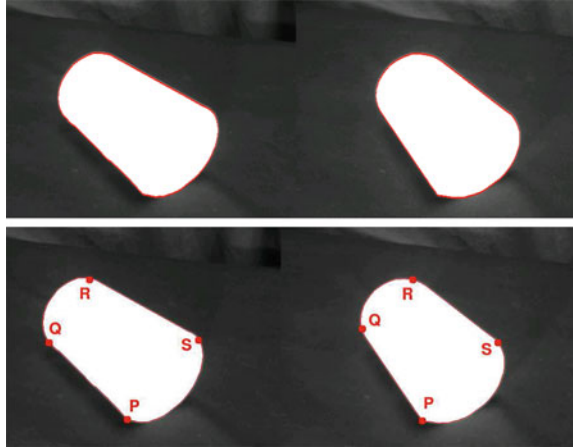
theoretical final error is the product of the two initial errors. Moreover, for high slants and for small objects, the edges of the least visible side have very short separation angles, for which the relative error is much higher. Anyway, the worst case error is never larger than a few centimeters, and this is enough for reliable grasping by the robot hand, as shown in Sect. 6.2.4.

5.4.4.2 Additional Experiments

The second class of objects used to test the system were approximately cylindrical shapes, which still offer parallel edges. In the experimental setup, cylinders are lying on a plane, and the slant to estimate is that of their axis. Four salient points are detected on cylinder contours, those points in which curvature changes from 0 to some positive value, i.e., the transition from straight to curved segments. Those four points are treated as they were the P, Q, R and S of the box shapes (see Fig. 5.16). In this way, estimators 1, 2, 5 and 6 of Table 5.4 can be computed. The results of just few experiments are encouraging, as the average orientation estimation error is around 3.5° . It must be said though that the values chosen for viewpoint, slant and distance were those that gave the most consistent results in the experiments with boxes, and the method has not been tested exhaustively in various different conditions.

For spheres, only centroid, size and distance can be estimated. As for spheres there is no reliable point P with minimum y coordinate, the centroid was used instead in order to detect the vergence angle γ_P and hence the distance of the object. Distance computed with this method was found to have higher precision than for boxes and

Fig. 5.16 Contour and salient points extraction for cylindrical shapes, left and right images



cylinders, and this reflects in an improved size estimation for spheres compared to the other classes.

5.5 Conclusions

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

The experimental results obtained with the robotic visual system confirm the hypothesis that integration of monocular and binocular data provide a robot with superior estimation capabilities. The final merged estimator obtained appropriately weighting the different cues is robust across working conditions, in a way that is probably not attainable by a simple estimator alone.

The following step is to make use of the extracted information regarding object potential grasping features to generate suitable action plans. This is the subject of next chapter.

References

- Adams DL, Zeki S (2001) Functional organization of macaque V3 for stereoscopic depth. *J Neurophysiol* 86(5):2195–2203
- Anzai A, Chowdhury SA, DeAngelis GC (2011) Coding of stereoscopic depth information in visual areas v3 and v3a. *J Neurosci* 31(28):10270–10282. doi:[10.1523/JNEUROSCI.5956-10.2011](https://doi.org/10.1523/JNEUROSCI.5956-10.2011)
- Backus BT, Banks MS, van Ee R, Crowell JA (1999) Horizontal and vertical disparity, eye position, and stereoscopic slant perception. *Vision Res* 39(6):1143–1170
- Backus BT, Fleet DJ, Parker AJ, Heeger DJ (2001) Human cortical activity correlates with stereoscopic depth perception. *J Neurophysiol* 86(4):2054–2068
- Backus BT, Banks MS (1999) Estimator reliability and distance scaling in stereoscopic slant perception. *Perception* 28(2):217–242
- Banks MS, Hooge IT, Backus BT (2001) Perceiving slant about a horizontal axis from stereopsis. *J Vis* 1(2):55–79. doi:[10.1167/1.2.1](https://doi.org/10.1167/1.2.1)
- Bar M, Tootell RB, Schacter DL, Greve DN, Fischl B, Mendola JD, Rosen BR, Dale AM (2001) Cortical mechanisms specific to explicit visual object recognition. *Neuron* 29(2):529–535
- Blanz V, Tarr MJ, Bülthoff HH (1999) What object attributes determine canonical views? *Perception* 28(5):575–599
- Bradshaw MF, Elliott KM, Watt SJ, Hibbard PB, Davies IRL, Simpson PJ (2004) Binocular cues and the control of prehension. *Spat Vis* 17(1–2):95–110
- Brautaset RL, Jennings JAM (2005) Distance vergence adaptation is abnormal in subjects with convergence insufficiency. *Ophthalmic Physiol Opt* 25(3):211–214. doi:[10.1111/j.1475-1313.2005.00274.x](https://doi.org/10.1111/j.1475-1313.2005.00274.x)
- Bray S, Arnold AEGF, Iaria G, MacQueen G (2013) Structural connectivity of visuotopic intraparietal sulcus. *Neuroimage* 82:137–145. doi:[10.1016/j.neuroimage.2013.05.080](https://doi.org/10.1016/j.neuroimage.2013.05.080)
- Brouwer GJ, van Ee R, Schwarzbach J (2005) Activation in visual cortex correlates with the awareness of stereoscopic depth. *J Neurosci* 25(45):10403–10413. doi:[10.1523/JNEUROSCI.2408-05.2005](https://doi.org/10.1523/JNEUROSCI.2408-05.2005)
- Bülthoff HH, Edelman SY, Tarr MJ (1995) How are three-dimensional objects represented in the brain? *Cereb Cortex* 5(3):247–260
- Chetverikov D, Szabo Z (1999) A simple and efficient algorithm for detection of high curvature points in planar curves. In: Workshop of the austrian pattern recognition group, pp 175–184
- Clark JJ, Yuille AL (1990) Data fusion for sensory information processing systems. Springer, Berlin
- Cumming BG, DeAngelis GC (2001) The physiology of stereopsis. *Annu Rev Neurosci* 24:203–238. doi:[10.1146/annurev.neuro.24.1.203](https://doi.org/10.1146/annurev.neuro.24.1.203)
- Deneve S, Pouget A (2003) Basis functions for object-centered representations. *Neuron* 37(2):347–359
- Dobbins AC, Jeo RM, Fiser J, Allman JM (1998) Distance modulation of neural activity in the visual cortex. *Science* 281(5376):552–555
- Einhäuser W, Hipp J, Eggert J, Körner E, König P (2005) Learning viewpoint invariant object representations using a temporal coherence principle. *Biol Cybern* 93(1):79–90. doi:[10.1007/s00422-005-0585-8](https://doi.org/10.1007/s00422-005-0585-8)
- Ekvall S, Hoffmann F, Kragic D (2003) Object recognition and pose estimation for robotic manipulation using color cooccurrence histograms. In: IEEE International conference on intelligent robots and systems, pp 1284–1289
- Endo K, Haranaka Y, Shein WN, Adams DL, Kusunoki M, Sakata H (2000) Effects of different types of disparity cues on the response of axis-orientation selective cells in the monkey parietal cortex. *Nippon Ganka Gakkai Zasshi* 104(5):334–343
- Ferrier N (1999) Determining surface orientation from fixated eye position and angular visual extent. In: IEEE international conference on robotics and automation, pp 938–943. doi:[10.1109/ROBOT.1999.772426](https://doi.org/10.1109/ROBOT.1999.772426)
- Gehrig S, Badino H, Paysan P (2006) Accurate and model-free pose estimation of small objects for crash video analysis. In: British machine vision conference, Edinburgh

- Genovesio A, Ferraina S (2004) Integration of retinal disparity and fixation-distance related signals toward an egocentric coding of distance in the posterior parietal cortex of primates. *J Neurophysiol* 91(6):2670–2684. doi:[10.1152/jn.00712.2003](https://doi.org/10.1152/jn.00712.2003)
- Georgieva S, Peeters R, Kolster H, Todd JT, Orban GA (2009) The processing of three-dimensional shape from disparity in the human brain. *J Neurosci* 29(3):727–742. doi:[10.1523/JNEUROSCI.4753-08.2009](https://doi.org/10.1523/JNEUROSCI.4753-08.2009)
- Germann M, Breitenstein MD, Park IK, Pfister H (2007) Automatic pose estimation for range images on the GPU. In: International conference on 3-d digital imaging and modeling, pp 81–90. doi:[10.1109/3DIM.2007.13](https://doi.org/10.1109/3DIM.2007.13)
- Goddard J (1998) Pose and motion estimation using dual quaternion-based extended Kalman filtering. In: SPIE: three-dimensional image capture and applications, vol 3313
- Gonzalez F, Perez R (1998) Neural mechanisms underlying stereoscopic vision. *Prog Neurobiol* 55(3):191–224
- Greenwald HS, Knill DC, Saunders JA (2005) Integrating visual cues for motor control: a matter of time. *Vis Res* 45(15):1975–1989. doi:[10.1016/j.visres.2005.01.025](https://doi.org/10.1016/j.visres.2005.01.025)
- Grill-Spector K, Kanwisher N (2005) Visual recognition: as soon as you know it is there, you know what it is. *Psychol Sci* 16(2):152–160. doi:[10.1111/j.0956-7976.2005.00796.x](https://doi.org/10.1111/j.0956-7976.2005.00796.x)
- Grill-Spector K, Kushnir T, Hendler T, Edelman S, Itzhak Y, Malach R (1998) A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Hum Brain Mapp* 6(4):316–328
- Heeley DW, Scott-Brown KC, Reid G, Maitland F (2003) Interoocular orientation disparity and the stereoscopic perception of slanted surfaces. *Spat Vis* 16(2):183–207
- Hillis JM, Watt SJ, Landy MS, Banks MS (2004) Slant from texture and disparity cues: optimal cue combination. *J Vis* 4(12):967–992. doi:[10.1167/4.12.1](https://doi.org/10.1167/4.12.1)
- Howard IP, Rogers BJ (2002) Seeing in depth. I Porteous
- Jacobs R (2002) What determines visual cue reliability? *Trends in Cogn Sci* 6(8):345–350
- James KH, Humphrey GK, Goodale MA (2001) Manipulating and recognizing virtual objects: where the action is. *Can J Exp Psychol* 55(2):111–120
- Jones DG, Malik J (1992) Determining three-dimensional shape from orientation and spatial frequency disparities. In: European conference on computer vision, pp 662–669
- Julesz B (1971) Foundations of cyclopean perception. MIT Press, Cambridge
- Knill DC (2007) Robust cue integration: a Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *J Vis* 7(7):5.1-524. doi:[10.1167/7.7.5](https://doi.org/10.1167/7.7.5)
- Konen CS, Mruczek REB, Montoya JL, Kastner S (2013) Functional organization of human posterior parietal cortex: grasping- and reaching-related activations relative to topographically organized cortex. *J Neurophysiol* 109(12):2897–2908. doi:[10.1152/jn.00657.2012](https://doi.org/10.1152/jn.00657.2012)
- Kragic D, Christensen HI (2001) Cue integration for visual servoing. *IEEE J Rob Autom* 17(1):18–27. doi:[10.1109/70.917079](https://doi.org/10.1109/70.917079)
- Landy MS, Maloney LT, Johnston EB, Young M (1995) Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res* 35(3):389–412
- Lehky SR, Pouget A, Sejnowski TJ (1990) Neural models of binocular depth perception. *Cold Spring Harb Symp Quant Biol* 55:765–777
- Lehky SR, Sejnowski TJ (1990) Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity. *J Neurosci* 10(7):2281–2299
- Lippiello V, Siciliano B, Villani L (2001) Position and orientation estimation based on Kalman filtering of stereo images. In: IEEE Int Conf Control Appl 702–707. doi:[10.1109/CCA.2001.973950](https://doi.org/10.1109/CCA.2001.973950)
- Lippiello V, Siciliano B, Villani L (2006) 3D pose estimation for robotic applications based on a multi-camera hybrid visual system. In: IEEE International conference on robotics and automation, pp 2732–2737
- Loftus A, Servos P, Goodale MA, Mendarozqueta N, Mon-Williams M (2004) When two eyes are better than one in prehension: monocular viewing and end-point variance. *Exp Brain Res* 158(3):317–327. doi:[10.1007/s00221-004-1905-2](https://doi.org/10.1007/s00221-004-1905-2)

- Marotta JJ, Perrot TS, Nicolle D, Servos P, Goodale MA (1995) Adapting to monocular vision: grasping with one eye. *Exp Brain Res* 104(1):107–114
- Marr D (1982) *Vision: a computational investigation into the human representation and processing of visual information*. Freeman WH
- Mon-Williams M, Tresilian JR, Roberts A (2000) Vergence provides veridical depth perception from horizontal retinal image disparities. *Exp Brain Res* 133(3):407–413
- Mon-Williams M, Tresilian JR (1999) Some recent studies on the extraretinal contribution to distance perception. *Perception* 28(2):167–181
- Naganuma T, Nose I, Inoue K, Takemoto A, Katsuyama N, Taira M (2005) Information processing of geometrical features of a surface based on binocular disparity cues: an fMRI study. *Neurosci Res* 51(2):147–155. doi:[10.1016/j.neures.2004.10.009](https://doi.org/10.1016/j.neures.2004.10.009)
- Norman JF, Todd JT, Phillips F (1995) The perception of surface orientation from multiple sources of optical information. *Percept Psychophys* 57(5):629–636
- Orban GA, Janssen P, Vogels R (2006) Extracting 3D structure from disparity. *Trends Neurosci* 29(8):466–473. doi:[10.1016/j.tins.2006.06.012](https://doi.org/10.1016/j.tins.2006.06.012)
- O'Reilly RC, Munakata Y (2000) *Computational explorations in cognitive neuroscience—understanding the mind by simulating the brain*. MIT Press, Cambridge
- Parker AJ (2004) From binocular disparity to the perception of stereoscopic depth. In: Chalupa LM, Werner JS (eds) *The visual neurosciences*, MIT Press, Cambridge, MA, chapter 49, pp 779–792
- Perry CJ, Tahiri A, Fallah M (2014) Feature integration within and across visual streams occurs at different visual processing stages. *J Vis* 14(2). doi:[10.1167/14.2.10](https://doi.org/10.1167/14.2.10)
- Peters G (2004) Efficient pose estimation using view-based object representations. *Mach Vis Appl* 16(1):59–63
- Poggio GF, Gonzalez F, Krause F (1988) Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *J Neurosci* 8(12):4531–4550
- Pouget S, Sejnowski A (1997) Spatial transformations in the parietal cortex using basis functions. *J Cogn Neurosci* 9(2):222–237
- Ray B, Ray K (1995) A new split-and-merge technique for polygonal-approximation of chain coded curves. *Pattern Recogn Lett* 16(2):161–169
- Read J (2005) Early computational processing in binocular vision and depth perception. *Prog Biophys Mol Biol* 87(1):77–108. doi:[10.1016/j.pbiomolbio.2004.06.005](https://doi.org/10.1016/j.pbiomolbio.2004.06.005)
- Riesenhuber M, Poggio T (2000) Models of object recognition. *Nat Neurosci* 3:1199–1204. doi:[10.1038/81479](https://doi.org/10.1038/81479)
- Rolls ET, Webb TJ (2014) Finding and recognizing objects in natural scenes: complementary computations in the dorsal and ventral visual systems. *Front Comput Neurosci* 8:85. doi:[10.3389/fncom.2014.00085](https://doi.org/10.3389/fncom.2014.00085)
- Rosenhahn B, Perwass C, Sommer G (2004) Pose estimation of 3D free-form contours. *Int J Comput Vis* 62:267–289. doi:[10.1007/s11263-005-4883-3](https://doi.org/10.1007/s11263-005-4883-3)
- Rosner B (1975) On the detection of many outliers. *Technometrics* 17(2):221–227
- Rousseeuw PJ, Leroy AM (1987) *Robust regression and outlier detection*. Wiley, New York
- Rutschmann RM, Greenlee MW (2004) Bold response in dorsal areas varies with relative disparity level. *Neuroreport* 15(4):615–619
- Sakata H, Taira M, Kusunoki M, Murata A, Tanaka Y, Tsutsui K (1998) Neural coding of 3D features of objects for hand action in the parietal cortex of the monkey. *Philos Trans R Soc B: Biol Sci* 353(1373):1363–1373
- Sakata H, Taira M, Kusunoki M, Murata A, Tsutsui K, Tanaka Y, Shein WN, Miyashita Y (1999) Neural representation of three-dimensional features of manipulation objects with stereopsis. *Exp Brain Res* 128(1–2):160–169
- Salinas E, Thier P (2000) Gain modulation: a major computational principle of the central nervous system. *Neuron* 27(1):15–21
- Saxena A, Schulte J, Ng AY (2007) Depth estimation using monocular and stereo cues. In: *International joint conferences on artificial intelligence*, pp 2197–2203

- Shikata E, Hamzei F, Glauche V, Knab R, Dettmers C, Weiller C, Büchel C (2001) Surface orientation discrimination activates caudal and anterior intraparietal sulcus in humans: an event-related fMRI study. *J Neurophysiol* 85(3):1309–1314
- Taira M, Tsutsui KI, Jiang M, Yara K, Sakata H (2000) Parietal neurons represent surface orientation from the gradient of binocular disparity. *J Neurophysiol* 83(5):3140–3146
- Taira M, Nose I, Inoue K, Tsutsui K (2001) Cortical areas related to attention to 3D surface structures based on shading: an fMRI study. *Neuroimage* 14(5):959–966. doi:[10.1006/nimg.2001.0895](https://doi.org/10.1006/nimg.2001.0895)
- Taylor G, Kleeman L (2003) Fusion of multimodal visual cues for model-based object tracking. In: Australasian conference on robotics and automation, Brisbane, Australia
- Teh CH, Chin R (1989) On the detection of dominant points on digital curves. *IEEE Trans Pattern Anal Mach Intell* 11(8):859–872. doi:[10.1109/34.31447](https://doi.org/10.1109/34.31447)
- Thoma V, Henson RN (2011) Object representations in ventral and dorsal visual streams: fMRI repetition effects depend on attention and part-whole configuration. *Neuroimage* 57(2):513–525. doi:[10.1016/j.neuroimage.2011.04.035](https://doi.org/10.1016/j.neuroimage.2011.04.035)
- Thomas OM, Cumming BG, Parker AJ (2002) A specialization for relative disparity in V2. *Nat Neurosci* 5(5):472–478. doi:[10.1038/nn837](https://doi.org/10.1038/nn837)
- Todd JT (2004) The visual perception of 3D shape. *Trends Cogn Sci* 8(3):115–121. doi:[10.1016/j.tics.2004.01.006](https://doi.org/10.1016/j.tics.2004.01.006)
- Tresilian JR, Mon-Williams M, Kelly BM (1999) Increasing confidence in vergence as a cue to distance. *Proc R Soc B: Biol Sci* 266(1414):39–44
- Tresilian JR, Mon-Williams M (2000) Getting the measure of vergence weight in nearness perception. *Exp Brain Res* 132(3):362–368
- Trotter Y, Celebrini S, Stricanne B, Thorpe S, Imbert M (1996) Neural processing of stereopsis as a function of viewing distance in primate visual cortical area V1. *J Neurophysiol* 76(5):2872–2885
- Trotter Y, Celebrini S, Durand JB (2004) Evidence for implication of primate area V1 in neural 3-D spatial localization processing. *J Physiol Paris* 98(1–3):125–134. doi:[10.1016/j.jphysparis.2004.03.004](https://doi.org/10.1016/j.jphysparis.2004.03.004)
- Trucco E, Verri A (1998) Introductory techniques for 3-D computer vision. Prentice Hall
- Tsao DY, Vanduffel W, Sasaki Y, Fize D, Knutsen TA, Mandeville JB, Wald LL, Dale AM, Rosen BR, Essen DCV, Livingstone MS, Orban GA, Tootell RBH (2003) Stereopsis activates V3A and caudal intraparietal areas in macaques and humans. *Neuron* 39(3):555–568
- Tsutsui KI, Taira M, Sakata H (2005) Neural mechanisms of three-dimensional vision. *Neurosci Res* 51(3):221–229. doi:[10.1016/j.neures.2004.11.006](https://doi.org/10.1016/j.neures.2004.11.006)
- Tsutsui K, Jiang M, Yara K, Sakata H, Taira M (2001) Integration of perspective and disparity cues in surface-orientation-selective neurons of area CIP. *J Neurophysiol* 86(6):2856–2867
- Ullman S (1996) High-level vision. Object recognition and visual cognition. MIT Press, Cambridge
- van Ee R, Banks MS, Backus BT (1999) An analysis of binocular slant contrast. *Perception* 28(9):1121–1145
- von der Heydt R, Zhou H, Friedman HS (2000) Representation of stereoscopic edges in monkey visual cortex. *Vision Res* 40(15):1955–1967
- Wandell BA (1995) Foundations of vision. Sinauer Associates
- Watt SJ, Bradshaw MF (2003) The visual control of reaching and grasping: binocular disparity and motion parallax. *J Exp Psychol Hum Percept Perform* 29(2):404–415
- Weigl A, Hohm K, Seitz M (1995) Processing sensor images for grasping disassembly objects with a parallel-jaw gripper. In: TELEMAN Telerobotics conference
- Weiss K, Wörn H (2004) Tactile sensor system for an anthropomorphic robot hand. In: IEEE International conference on manipulation and grasping, Genova, Italy
- Welchman AE, Deubelius A, Conrad V, Bühlhoff HH, Kourtzi Z (2005) 3D shape perception from combined depth cues in human visual cortex. *Nat Neurosci* 8(6):820–827. doi:[10.1038/nm1461](https://doi.org/10.1038/nm1461)