Chapter 8

Object perception and recognition

In chapter 2, based on neuroscience experiments that studied how humans physically interact with objects at sensorimotor level, the building blocks for a complete autonomous sensor-based manipulation system were settled. In chapters 3, 4 and 5 we have presented the implementation and validation of each of the building blocks. However, the identified blocks focus on physical interaction and do not include any visual perception or recognition of objects, which is necessary for object manipulation.

That gap is addressed in this chapter. We present an object detection and recognition component, that can be integrated into the presented framework in order to provide information about the objects. First, we have taken inspiration from human and primate studies to propose a theoretical scheme for hierarchical object recognition based on a three step process: classification, recognition and recall. Finally, the proposal is implemented and tested on a real robot. The integration on the system, can be done using the concept of perceptual primitives already introduced in Chapter 3. This primitives allow us to include specific perceptual actions in the task definition.

8.1 Introduction

The visual cortex of humans and other primates is composed of two main information pathways, called ventral stream and dorsal stream in relation to their location in the brain, depicted in Fig. 8.1 [Goodale and Milner, 1992]. The dorsal, "where/how", stream is concerned with providing the subject the ability of interacting with its environment in a fast, effective and reliable way, such as in limb movements. The dorsal stream includes areas especially dedicated to extract and encode 3D features of objects in a format suitable to be used for planning and executing reaching and grasping actions toward them. The ventral, "what", stream is instead devoted to perceptual analysis of the visual input, such as in recognition, categorization and assessment tasks.

The streams dissociation has been supported, but also criticized, by the neuroscientific community, and the original theory is constantly being revised and updated. The trend is towards a more integrated view, according to which, the two streams have complementary tasks and often interact with each other [Goodale, 2004].

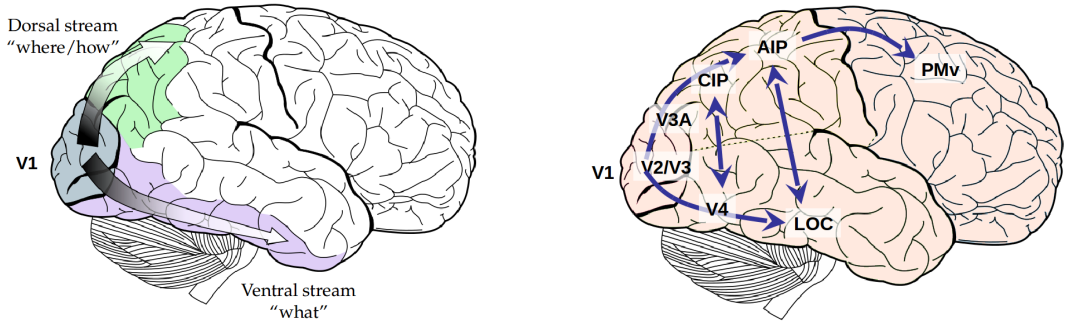


Figure 8.1: Left: Dorsal and ventral streams of the human brain. Right: Human brain areas of the dorsal and ventral streams.

In previous, related works, Chinellato et. al. modeled the visuomotor processing performed by dorsal stream areas in reaching and grasping actions [Chinellato and Del Pobil, 2008], [Chinellato and del Pobil, 2016]. That work devoted special attention to the area of the primate brain dedicated to grasping, Anterior Intraparietal Sulcus (AIP), but taking also into account possible interactions between the ventral and the dorsal streams [Chinellato and Del Pobil, 2009]. Their modeling efforts were validated by the implementation of the fundamental concepts on a real robotic setup, resulting in a skilled vision-based grasping behavior [Chinellato et al., 2008], [Grzyb et al., 2009]. The whole model framework is represented in Fig. 8.2, for a full detailed description of the model see [Chinellato and del Pobil, 2016].

In this chapter, we extend the work from [Chinellato and Del Pobil, 2009] implementing the ventral stream part of the model (i.e. object recognition) as presumably executed by the primate visual brain (light blue modules from Fig. 8.2). We offer a hierarchical interpretation of the incremental identification capabilities of a subject presented with geometrical 3D objects.

According to the proposed framework, ventral stream processing consists of 1) identifying the object class; 2) recognizing a single object within a class; 3) identifying a previously encountered object even among completely similar candidates. The first two steps of our object identification model have been implemented on a robot setup. The system is able to classify target objects in one of a given number of classes, and subsequently recognize a certain object among objects of the same class, taking advantage also from the estimation of object weight.

8.2 Neuroscience background

In humans, the visual information is processed through the dorsal and ventral streams in a sequential manner. Each of the streams goes through different brain areas that are

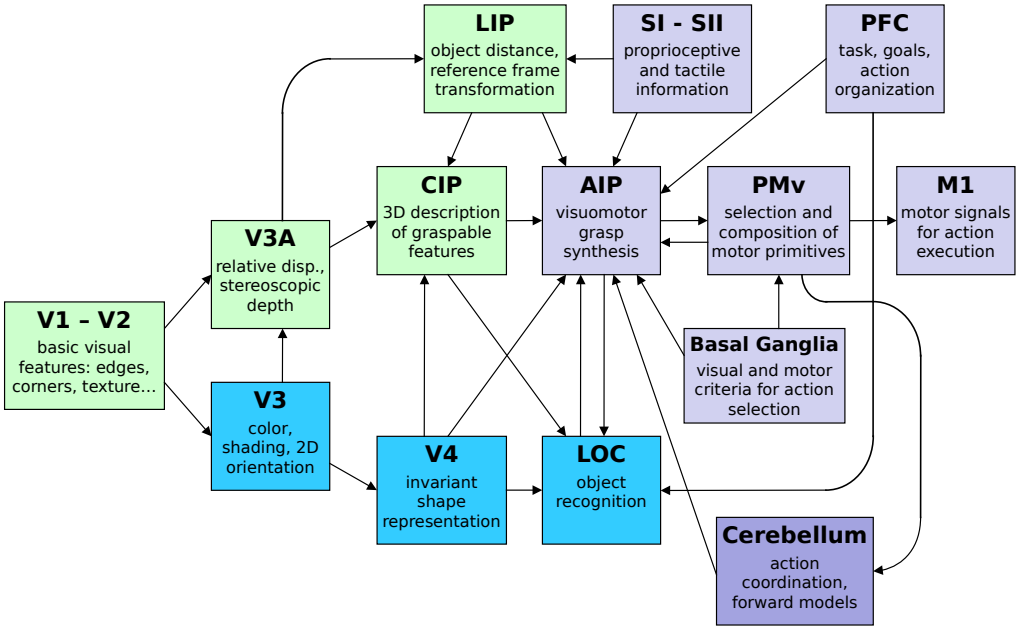


Figure 8.2: Model framework depicting all the principal areas involved in the planning and execution of vision-based grasping actions. Turquoise areas correspond to the dorsal stream. Light blue areas are considered to be ventral stream areas. Violet areas correspond to pre-motor cortex, motor cortex and the end of the dorsal stream.

separated from each other depending on their function (see Fig. 8.1). In the ventral stream, starting in the Primary Visual Cortex (V1), visual information is processed in a pipeline-like sequence until the objects in the scene are recognized and their identity is recalled in the Lateral-Occipital Complex (LOC). Although the information flows from V1 to LOC there are many feedback connections that connect the different areas to each other. In this chapter have used the functional model proposed by [Chinellato and del Pobil, 2016] and extended the already implemented dorsal stream with the implementation of the ventral pathway (see Fig. 8.2).

At the beginning of the visual processing, ventral and dorsal streams are not separated. Starting at V1 area, neurons are mainly sensitive to edges but also to the more global organisation of the scene. As information is further relayed to subsequent visual areas,

it is coded as increasingly non-local frequency/phase signals [Hubel and Wiesel, 1977]. The mathematical modelling of this function has been compared to Gabor transforms. Neurons in Secondary Visual Cortex (V2) are tuned to simple properties such as orientation, spatial frequency, and color [Hegdé and Van Essen, 2000]. Third Visual Complex (V3) area is where the division of the visual pathways begins, dorsal and ventral V3 have distinct connections with other parts of the brain, and contain neurons that respond to different combinations of visual stimulus. Colour-selective neurons are more common in the ventral V3 also known as VP. Visual Area V4 (V4) exhibits long-term plasticity, encodes stimulus salience in an invariant shape representation and is sensitive to attention [Sereno et al., 1995].

The last area is the LOC, in this area is where the results from the other ventral areas are integrated and the final steps of object recognition are performed. Object representation in LOC is highly invariant with respect to the stimulus type, showing equally good performances with either 3D or silhouette images, different color maps, lighting and so on. This suggests a higher level, conceptual representation of objects, independent of the actual stimulus that allowed recognition [Kourtzi and Kanwisher, 2001]. Object recognition in the ventral stream is very likely a "faded" process rather than a binary one. In fact, activation in the anterior part of the LOC is modulated by the actual level of recognition, and not by the nature of the stimulus [Bar et al., 2001]. In any case, geometric data are integrated with additional information, regarding for example color and texture of objects, to speed up and make object recognition more reliable [Grill-Spector et al., 1999].

Object recognition is performed gradually and hierarchically [Grill-Spector et al., 1998], [Bar et al., 2001]. Other findings indicate that the identification process is composed of at least two sequential 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, requiring just few milliseconds. 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. The second stage of object recognition is proper identification, performed by LOC, in which object identity is recognized within its category.

Regarding possible connections of ventral stream areas with the dorsal stream, a direct link has been found in the macaque brain between the most 3D responsive ventral inferior temporal area (the lower bank of the superior temporal sulcus) with the Caudal Intraparietal Sulcus (CIP) [Janssen et al., 2000]. This link could indicate both a ventral contribution to pose estimation, which we have previously modeled [Chinellato and Del Pobil, 2009] and a dorsal effect in object recognition.

8.3 Computational model of V4 and LOC areas

This section presents the description of the ventral stream modules from the brain functional model shown in Fig. 8.2 and described in [Chinellato and del Pobil, 2016].

Visual processing in the ventral stream is based on the production of increasingly invariant representations aimed at object recognition. In the functional model of the brain we follow, the ventral stream starts at V3 area (Fig. 8.2), region V4 codes at the same time shape, color and texture of features, which are then composed in the LOC to form more complex representations recognizable as objects. Output from area V3 is thus used by V4 to build a viewpoint invariant simple coding of the object, that can be used to classify it as belonging to one of a number of known object classes.

Downstream from V4, the LOC compares spatial and color data with stored information about previously observed objects, to finally recognize the target as a single, already encountered object. Object identification is thus performed in a hierarchical fashion, where the target is first classified into a given class and, only later, exactly identified as a concrete object. In each of these steps, recognition is not a true/false decision, but rather a probabilistic process, in which an object is classified or identified only up to a given confidence level. Thus, confidence values should be provided by the classification and identification procedures, so that ventral information can be given more or less credit. If recognition confidence is high, visual analysis can be simplified, as most required information regarding the target object is already available in memory. If recognition is instead considered unreliable, more importance is given to the on-line visual analysis performed by the dorsal stream. An 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., 1991]. 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 [Blanz et al., 1999], [James et al., 2001]. For this work, a viewpoint invariant classification procedure was implemented, based on basic 2D global object representations.

Considering the output of the V3 area as a segmented 2D contour of the object. Possible computational representations of 2D object contours are, for example, chain codes (e.g. Freeman Chain Code of Eight Directions [Freeman, 1961]) or 2D shape indexes (e.g. curvedness index).

Regarding possible dorsal contributions to ventral stream processing, various researchers pointed out that action-related information maintained in the dorsal stream is likely to

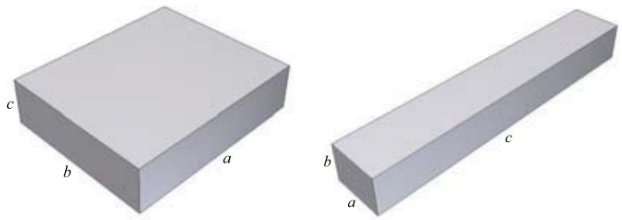


Figure 8.3: Examples of SOS (left) and AOS (right) dominant objects.

play an important role in the object recognition process. A set of possible affordances constitutes an additional way of identifying an object [Sugio et al., 1999], [Shmuelof and Zohary, 2005]. In this chapter we have modelled the dorsal-ventral interaction as a connection from the dorsal area CIP to the ventral area LOC as found in the macaque by [Janssen et al., 2000].

Two main neuronal populations have been distinguished in CIP. They both code for object orientation in space, but are selective for different object types. Surface Orientation Selective (SOS) neurons [Shikata et al., 1996] preferentially respond to flat stimuli of the kind shown on the left in Fig. 8.3. The second class of CIP neurons, Axis Orientation Selective (AOS) neurons [Sakata et al., 1998], represent the 3D orientation of the longitudinal axes of elongated objects (Fig. 8.3, right). The activation of SOS and AOS neurons according to different stimuli was previously modelled with the purpose of providing area AIP with information useful for grasp planning [Chinellato and Del Pobil, 2008]. Here, we employ this same information to aid LOC in object classification.

The SOS and AOS responsiveness found for the target object could be one possible format used by the dorsal stream to help the ventral areas in the recognition task. It is in fact very unlikely that two objects share the same SOS and AOS activations. CIP projections would thus provide the ventral stream with additional information for improving the reliability and speed of object recognition. For what concerns the representation of known objects, in their first years of development, human beings accumulate experience on properties such as color, texture, material, object identity, learning the likelihood of different relations among them. A working model of this recognition and generalization capacity should rely on a knowledge base founded on these properties (see e.g. the proposal of [Metzinger and Gallese, 2003]). In this chapter we use both SOS and AOS activations to aid object recognition and build a very simple knowledge base of geometrical shapes to use it for object identification purposes, reduced to basic features such as dimensions, color and weight.

In summary, emulating the mechanisms suggested by neuroscience studies, the approach to object classification proposed in the model is composed of a three stage process.

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 or 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. This process is performed in the V4 area of the brain.

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. This process occurs in the LOC area of the human brain.

3) *Object Recall.* In this final stage, that also happens in the LOC area, 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 confidence 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. The last stage of the process, Object recall, requires a higher level memory of the agent interactions with nearby objects, involving some sort of awareness regarding the nature of his behavior and his relation with the environment, and is thus beyond our goals and current modeling skills. The robotic implementation of the first two stages is described in the next section.

8.4 Implementation

The recognition system follows a hierarchical scheme, starting from categorization, then recognition and finally object recall. In this section the two first steps of the object recognition system described above are implemented on the robot setup of the Robotic Intelligence Lab. The implementation presented in this section takes into consideration the robotic setup used and the reduced universe of possible objects. It is a platform dependent implementation that intends to replicate the functional brain model for a further validation under certain known conditions.

8.4.1 Robotic setup

The robotic setup consisted of one PA10-7C 7DOF manipulator with a force-torque sensor and a barrett hand. A stereo camera Videre Design was mounted on the arm of the robot with an eye-in-hand configuration, see Fig. 8.5. This implementation was performed before the Tombatossals torso, described in Appendix A.1, was available. In fact, the system used for this implementation was later used to compose Tombatossals left side.

8.4.2 V4 area: Shape classification

The shape 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. In the reduced world of the robot, 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. As explained in the previous section, it is possible that the V4 area of the human brain encodes the objects using an invariant shape representation. Given that the input of the V4 area is the object contour, two different approaches were tested: a chain code representation and a curvedness index.

Chain code representation

The first tested object representation 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. However, after the preliminary experiments, this solution did not provide the required behaviour. In fact, results on training objects from different viewpoints 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 at recognizing known objects, but not at 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%.

Curvedness index object representation

This representation is based on only one index for each object, the curved fraction of its contour. The curvedness of a contour is calculated as the 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.

8.4.3 LOC area: Object Recognition

After object class has been identified, the second step in the recognition process is to distinguish among objects of the same class. A number of fundamental features can be defined for object recognition purposes in order to perform this second step. For the experiments we have only used box-like objects, we exploit this assumption to select the features that will form our object representation in the LOC area. We considered the estimated size of the three sides (D_1, D_2, D_3 ordered from larger to smaller such that $D_1 > D_2 > D_3$), color (C), weight (W) and the activation of SOS and AOS neurons (SOS, AOS).

As discussed in Sec. 8.2, dorsal information is likely forwarded to the ventral stream, SOS and AOS activation is a sort of information that is very likely forwarded to the ventral stream by dorsal areas. The implementation of SOS and AOS activation are non-linear combinations of the estimated principal object dimensions, defined according to neurophysiological data and potential use in vision-based grasping actions, the transfer functions of both AOS and SOS were modelled by [Chinellato and Del Pobil, 2008]. SOS activation is defined by equation 8.1 and AOS activation is defined by 8.2

$$R_{SOS} = 1 - \left(\frac{D_1 - D_2}{D_1 + D_2}\right)^2 - 0.03\frac{D_3}{D_1 + D_2} - 0.5\frac{1}{1 + e^{-0.04(D_3 - H)}} \tag{8.1}$$

$$R_{AOS} = 1 - \frac{D_1 - D_2}{D_1 + D_2} - 0.37\frac{D_1}{D_3} - 0.5\frac{1}{1 + e^{-0.04(D_1 - H)}} \tag{8.2}$$

Where D_1, D_2 and D_3 are the dimensions in millimetres of the object's bounding box and $D_1 > D_2 > D_3$. H corresponds to the comfortable hand opening parameter which generally is 150mm. The constant values were tuned to match the real response obtained from real experiments. The details about the computational modelling of SOS and AOS neurons is provided in [Chinellato and Del Pobil, 2008].

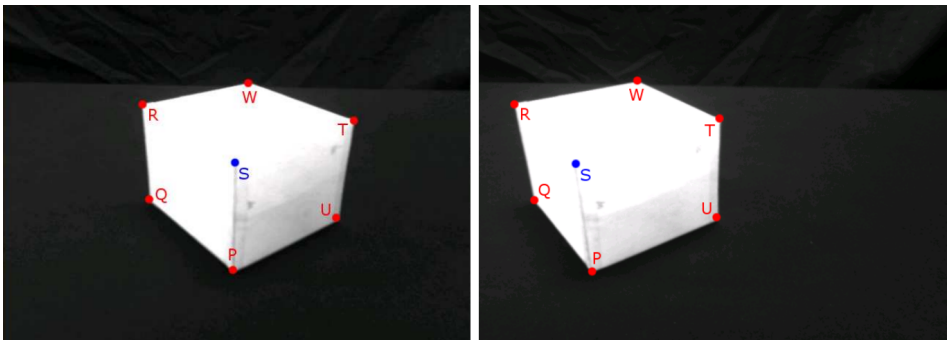


Figure 8.4: The corners of the box shaped object are used to determine its dimensions (D_1, D_2, D_3) ordered from larger to smaller. The dimensions are used for object recognition and approach vector computation in order to perform grasping actions on it.

The principal dimensions (D_1, D_2, D_3) are calculated exploiting the assumption of box-like objects. The principal corners of the object are detected in both images of the stereo pair (see Fig 8.4) and the main dimensions are calculated using the 3D position of the detected object points.

Given that the position and orientation of the object can be detected, a grasping action is performed on it, and using the force-torque sensor on the wrist, the weight of the grasped object can be estimated and used for recognition purposes.

8.5 Experiments

In order to validate the proposed implementation of the ventral stream and the functional model of the brain; two experiments have been carried out. The first one to test the implementation of the object classification module (V4 area), the second one to validate the implementation of the object recognition module (LOC area).

8.5.1 Scenario and assumptions

To ease the segmentation process in both experiments, objects are presented with light colors over a black background. A less restrictive object segmentation system that could be used for further experiments was detailed in Section 6.3.3. Objects are placed on top of a table in front of the robot inside its workspace.

8.5.2 Shape classification

For the implementation of the shape classification module (V4 area), two possible representations were proposed: chain code and curvedness index. However, during the preliminary experiments, the chain code was found to be not suitable and was dis-

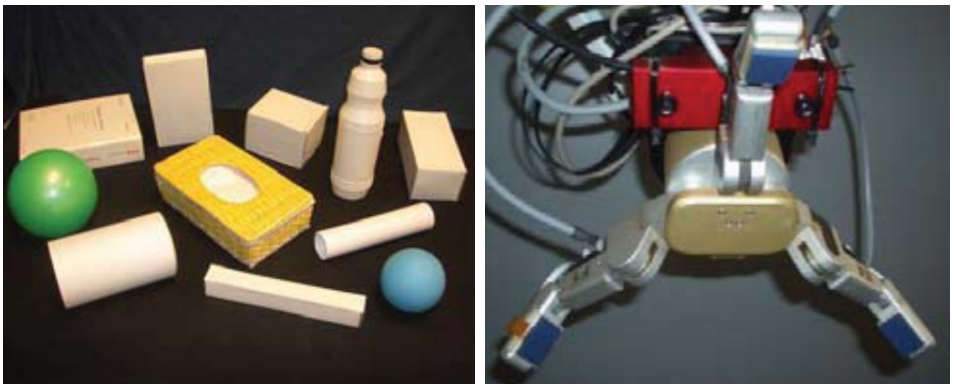


Figure 8.5: Left: Objects used for the experiments on black background. Right: Robotic setup consisting of a PA10-7C manipulator with a Barrett Hand and a Videre Stereo vision system.

carded. In this subsection the curvedness index representation is experimentally tested and validated.

Test objects

For the categorization, we divided the objects into three main classes: box-like, cylinders and spheres, see Fig. 8.5 Left. The objects used for training the system are pure basic shapes while some of the objects used for testing are regular objects.

Curvedness index object representation

To validate the use of the curvedness index as a shape category descriptor, we have performed an experiment that consisted of two phases, training and testing.

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). Images are taken again from 19 viewpoints distributed along a 90° range in azimuth, with elevation kept at about 40° to grant a clear 3D view of objects. 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 (i.e. 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:

$$m_{iK} = \frac{\sigma_K}{|c_i - \mu_K|} \tag{8.3}$$

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

$$p_{iK} = \frac{m_{iK}}{m_{iB} + m_{iC} + m_{iS}} \tag{8.4}$$

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 reaches this value, the object is not classified, and an exploratory movement aimed at providing the robot with images taken from a different viewpoint is required. An exception is the case of uncertainty between boxes and cylinders. If $p_{iB} + p_{iC} \geq 70\%$, then the object is classified in the less restrictive class, i.e., as a cylinder. This is because in our biologically-inspired pose estimation system [Chinellato and Del Pobil, 2009] boxes provide more useful information for orientation estimation than cylinders. Thus, a misclassification of a box as a cylinder would just imply that some available information is not used, whilst a misclassification of a cylinder as a box would very likely cause a wrong interpretation of the available data.

Results and discussion

After performing experiments using the two proposed object representations, the results obtained are presented and discussed in the next subsections.

Curvedness index object representation

Classification results for objects in the training set are provided in Table 8.1 Left. Cases of misclassification are highlighted in **bold red** whilst uncertain cases are underlined. 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 8.1 Right. Most cases of missing classification regard the same 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: 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.

8.5.3 Object Recognition

In the training phase, the system is provided with a number of labelled objects, and uses visual perception to associate detected features to object identity. For the recognition engine we have used a probabilistic linear estimator. A feature matches a given object

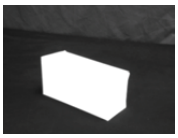
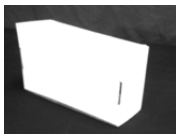
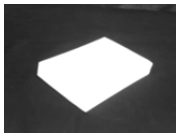
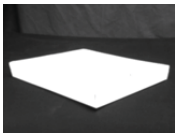
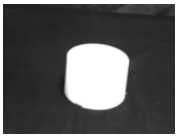
#	Object	0°	30°	60°	90°	#	Object	0°	30°	60°	90°
1		98.2 1.6 0.2	86.6 11.6 1.8	84.8 13.1 2.1	94.9 4.4 0.7	10		98.8 1.1 0.1	85.8 12.5 1.7	85.1 13.1 1.8	85.6 12.7 1.7
2		93.0 1.6 0.9	85.9 11.6 1.9	84.8 13.1 2.1	91.2 4.4 1.2	11		94.6 4.8 0.6	90.0 8.9 1.1	85.1 13.1 1.8	91.1 7.9 1.0
3		93.9 5.3 0.8	84.8 13.1 2.1	84.8 13.1 2.1	87.3 11.0 1.7	12		80.5 17.7 1.8	96.2 3.4 0.4	86.0 12.3 1.7	91.7 7.6 0.7
4		99.9 0.1 0.0	86.9 11.4 1.7	84.8 13.1 0.1	99.2 0.7 1.7	13		94.5 5.0 0.5	95.6 3.9 0.5	90.6 8.3 1.1	99.4 0.5 0.1
5		**86.2** 12.4 1.4	0.6 98.6 0.8	0.3 97.9 1.8	0.8 92.9 6.3	14		0.6 30.8 68.6	5.9 91.4 2.7	0.3 96.6 3.1	0.5 51.9 47.6
6		58.1 38.7 3.2	1.7 96.8 1.5	20.8 75.0 4.2	0.4 97.9 1.7	15		60.3 36.2 3.5	61.7 35.0 3.3	35.9 59.4 4.7	0.3 99.2 0.5
7		2.7 95.2 2.1	2.4 95.7 1.9	0.6 94.6 4.8	9.0 88.5 2.5	16		57.9 38.6 3.5	84.9 13.0 2.1	93.3 5.8 0.9	98.1 1.7 0.2
8		0.5 25.4 74.1				17		0.8 98.3 0.9	1.0 87.4 11.6	0.4 95.1 4.5	0.7 90.3 9.0
9		0.4 24.2 75.4				18		17.9 77.3 4.8	0.2 97.4 2.4	3.7 94.3 2.0	3.8 93.8 2.4

Table 8.1: Object classification percentages for different slants. Left: Training shapes (objects 1 to 9). Right: Test shapes (objects 10 to 18). Percentages of each class shown row-wise (B,C,S) for each object.

identity i if the set of features of the sample x is in the variability range of that object identity, expressed by the multidimensional mean μ_i and variance σ_i of its feature space:

$$\mu_i - n\sigma \leq x \leq \mu_i + n\sigma \tag{8.5}$$

where parameter n defines the tolerance of the classifier. We tested our classifier with a "risky" setting, $n = 3$, which should grant higher recognition rates but also more errors, and a more prudential $n = 2$, that should increase the number of unclassified samples. Statistically, $n = 3$ corresponds to about a 99% confidence interval, and $n = 2$ to approximately 95%. The mean and standard deviation vectors identifying the feature space of an object class are computed on a training set including samples of all available objects.

During the first phases of our experimental tests, we realized that the color feature is dominant, and no other features would be required if objects had different colors. Recognition tests in which shapes were distinguishable by color gave us more than 99% correct identification rate, showing that the problem was indeed too easy. For making our classifier more robust and test the importance of other features, in particular the SOS and AOS representations, we employed objects of the same material and the same color, and omitted color information in the computation. Thus our representation of an object is formed by 6 features: the three main object dimensions D_1, D_2, D_3, SOS and AOS activation, and weight W.

Test objects

Object recognition tests were performed on nine objects of the Box class, i.e. objects 1, 2, 3, 4, 10, 11, 12 and 13 of Table 8.1, plus one additional object. The weight of the target object was estimated upon grasping and lifting it, performed according to a multimodal visual/tactile procedure [Grzyb et al., 2009], [Felip and Morales, 2009].

Results and discussion

We checked the behavior of our probabilistic linear classifier including different subsets of the features, for $n = 3$ and $n = 2$, as shown in Table 8.2. We are especially interested in two aspects: the significance and usefulness of the SOS and AOS features and the advantages of multimodal integration offered by the use of object weight. Comparing the first three lines of Table 8.2 we notice that the pair SOS, AOS is nearly as informative as the entire set of dimensions D_1, D_2, D_3, being their performances nearly equal (only about 1% difference in correct answers). This indicates that the way we modeled neural activation of CIP neurons is not only suitable to represent object features for action planning, but captures also the global shape of the objects employed in recognition. Nevertheless, not all visual information regarding target objects is contained in the SOS, AOS pair, as can be seen by the increased performance obtained adding one of

	n=3			n=2		
Feature set	**C**	**W**	**U**	**C**	**W**	**U**
SOS, AOS	72.3	26.8	0.9	65.9	25.7	8.4
D_1, D_2, D_3	73.3	23.0	3.7	66.7	17.6	15.7
SOS, AOS, D_1	77.1	21.9	1.0	68.7	15.7	15.6
SOS, AOS, W	70.1	16.4	13.5	59.2	11.5	29.3
SOS, AOS, D_1, D_2, D_3, W	78.7	0.8	20.5	57.6	0.0	42.4

Table 8.2: Classification results of probabilistic linear classifier. Percentages correct (**C**) and wrong (**W**), and of uncertain cases (**U**)

Feature set	**MLS**	**MN**	**ND**
SOS, AOS	42.0	72.7	78.2
D_1, D_2, D_3	42.0	77.1	80.7
SOS, AOS, D_1	42.8	50.7	77.9
SOS, AOS, W	58.0	75.7	97.9
SOS, AOS, D_1, D_2, D_3, W	70.2	73.1	98.0

Table 8.3: Classification results of Minimum Least Square (MLS), Nearest Mean (NM) and Normal Density (ND) classifiers. Percentages of correct classifications.

the dimensions, e.g. D_1, as in line 3 of Table 8.2. It is significant though that the triplet SOS, AOS, D_1 performs better than the simple set of dimensions, again reinforcing the idea that our modeled expressions do capture significant visual characteristics of objects. On the other hand, object recognition in the ventral stream is substantially size-invariant, so it is reasonable that information on absolute size offers only little additional advantage.

The above considerations can be confirmed looking at Fig. 8.6 , in which all the set of samples for the nine target objects is depicted on an SOS/AOS space. Again, while for some objects the two features are very informative and nearly enough to recognition, it is apparent that other objects require additional information to be resolved from each other. The graph shows also that there is a large variability in the representation of some objects, due to visual imprecisions. It is worth reminding on this regard that the samples were taken observing all objects from different canonical viewpoints, and this constitutes an important additional complexity in the recognition process.

Regarding multimodal recognition aided by object weight estimation, lines 4 and 5 of Table 8.2 show that the performance in term of correct (C) answers does not really improve. On the other hand, the number of wrong (W) answers is now much smaller (less than 1% for the whole feature set, line 5), and many more samples are classified as uncertain (U). The introduction of the W feature seems to provide the system with the

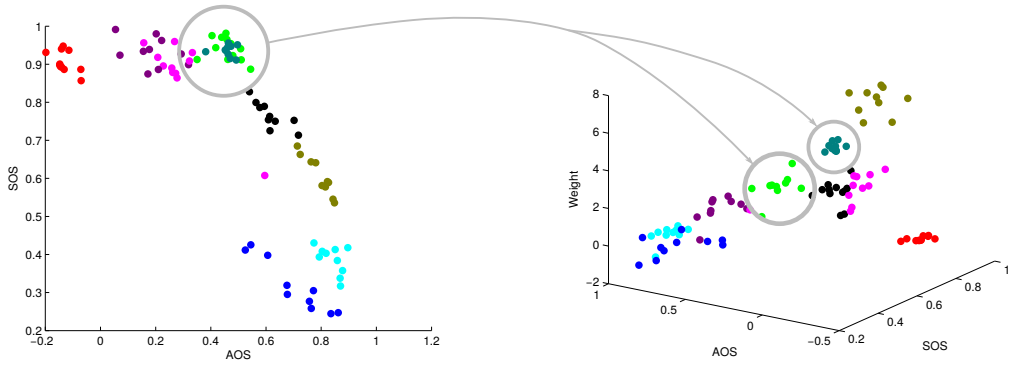


Figure 8.6: Distribution of object samples. Each color corresponds to samples of a different object. Right: Distribution of object samples plotted on a SOS/AOS feature space. Left: Distribution of object samples plotted on a SOS/AOS/W feature space. The grey circles and arrows show how adding the weight to the feature space it is possible to separate dark green and light green classes.

ability of detecting potentially problematic situations, in which the wiser decision is to avoid inserting the sample in any given class. The comparison in performance between the different values of n confirms the hypothesized behaviour, showing higher correct recognition values, but also more wrong answers, for $n = 3$, and more unclassified samples for $n = 2$. These results suggested us a new experimental scheme, in which n changes dynamically with the number of classified samples, starting from lower, more conservative values and growing ideally up to a value that grants no uncertain cases, once the set of objects has been fully learnt.

As the overall performance of the recognition system is not extremely good, we wanted to check whether this was due to the limits of our simple classifier or to the properties of the feature set. The graph in Fig. 8.6 also indicates that the triplet of dimensions SOS, AOS and W provide a high separability of the classes. In order to solve this issue, we applied other three classifiers to our set of features: Minimum Least Square (MLS), Nearest Mean (NM) and Normal Density (ND), from the Matlab PRTools4 Toolbox for pattern recognition [van der Heijden et al., 2004].

We did not consider classifiers that require to maintain a full memory of all encountered samples, such as k-nearest neighbours, for their lack of biological plausibility. The results of Table 8.3 show that at least one of the classifiers, ND, grant very high recognition rates, for all feature subsets. The performance of NM are comparable with our linear classifier, whilst MLS is definitely worse. Comparing again the different feature subsets, the pair SOS, AOS and the triplets D_1, D_2, D_3 and SOS, AOS, D_1 are approximately equivalent. The inclusion of W provides much better results (apart for the MN

classifier), and it is interesting to observe that for ND the subset SOS, AOS, W gives practically the same, extremely good performance that the whole set of features (97.9% against 98.0%). On the one hand, this confirms the appropriateness of the SOS-AOS representation to tackle the recognition problem, and the edge offered by multimodal processing and the use of object weight. On the other hand, the difference between our linear classifier and ND is not very large for the purely visual subset, but rises up to almost 28% in multimodal classification, suggesting that more complex tools are required to take full advantage of its potentialities.

We have also implemented an incremental version of the learning algorithm, in which, if a sample is classified, it is directly added to the classifier memory to be used in subsequent tests, and mean and variance are immediately recalculated. Unclassified instances are ignored, unless a human supervisor is available. In this case, he/she is asked to label unidentified samples so they can be added to the memory. Thus, in this implementation the module keeps learning while recognizing objects, and the more samples the system can include in its "knowledge" of the world, the more robust its classification becomes, and the approximation of the average to the real value of the feature set improves.

8.6 Conclusion

In this chapter we proposed and implemented a theoretical scheme for hierarchical object recognition inspired by primate brain mechanisms. The scheme proposed is based on a three step process. We implemented the two first steps, shape classification and object recognition on a real robot setup, achieving good results in both tasks.

Distinguishing features of our approach are: 1) the use of typical dorsal processing information, such as SOS and AOS activations, in a ventral visual task, implementing a possible link between the cortical visual streams; 2) multimodal integration by including object weight in the recognition process.

However, we have considered a reduced universe of objects and the representation used to encode them (i.e. contour curvedness) should be tested on a broader set of objects to validate its suitability for a real world scenario.

Current and future research include: 1) a dynamical learning framework for the object recognition step, in which the agent gradually increases its confidence in classifying new samples, and thus increasingly improve its knowledge of the world; 2) a return projection from the ventral stream to dorsal areas, in which remember object identity contributes in properly configuring the hand during grasping actions; 3) enhancement of initial visual processing to get rid of the color and background assumptions; 4) integration into the contact based manipulation framework.

The experiments and implementation presented in this chapter were performed before the development of the architecture presented through this thesis. However, the integration of the work presented in this chapter into the system was taken into consideration and it can be done through the use of perceptual primitives.

The research leading to the results presented in this chapter was published in [Chinellato et al., 2011]. The work presented in this chapter is the result of the collaboration with Eris Chinellato. The computational model of the brain shown in Figure 8.2, the implementation of the AOS and SOS activation and the object categorization using the shape curvedness descriptors are part of his PhD. Thesis [Chinellato, 2008].