Toward an Integrated Visuomotor Representation of the Peripersonal Space

Eris Chinellato[1], Beata J. Grzyb[1], Patrizia Fattori[2] and Angel P. del Pobil[1]

[1] Robotic Intelligence Lab
Universitat Jaume I, Castellón de la Plana, Spain
{eris,grzyb,pobil}@uji.es
[2] Dipartimento di Fisiologia Umana e Generale
Universita' di Bologna, Italy
patrizia.fattori@unibo.it

Abstract. The purpose of this work is the creation of a description of objects in the peripersonal space of a subject that includes two kinds of concepts, related to on-line, action-related features and memorized, conceptual ones, respectively. The inspiration of such description comes from the distinction between sensorimotor and perceptual visual processing as performed by the two visual pathways of the primate cortex. A model of such distinction, and of a further subdivision of the dorsal stream, is advanced with the purpose of applying it to a robotic setup. The model constitutes the computational basis for a robotic system able to achieve advanced skills in the interaction with its peripersonal space.

1 Introduction

Humans and other primates possess a superior ability in dealing with objects in their peripersonal space. Neuroscience research showed that they make use of a bi-fold visual and visuomotor process in order to analyze and interact with objects surrounding them. Indeed, the primate visual cortex is composed of two main information pathways, called *ventral stream* and *dorsal stream* in relation to their location in the brain, depicted in Figure 1. The traditional distinction [1] talks about the ventral "what" and the dorsal "where/how" visual pathways. In fact, the ventral stream is devoted to perceptual analysis of the visual input, such as in recognition, categorization, assessment tasks. The dorsal stream is instead concerned with providing the subject the ability of interacting with its environment in a fast, effective and reliable way. This second stream is directly involved in estimating distance, direction, shape and orientation of target objects for reaching and grasping purposes. The tasks performed by the two streams, their duality and interaction, constitute the neuroscientific basis of this work.

The research presented here is the first step toward the goal of improving the skills of autonomous robotic systems in their exploration of the nearby space and interaction with surrounding objects. We propose the outline of a model toward the achievement of an integrated object representation which includes on-line, action-oriented visual information (dorsal stream) with knowledge about nearby

object and memories of previous interaction experiences (ventral stream). Particular importance has been given to the use of binocular data and proprioceptive information regarding eye position, critical in the transformation of sensory data into appropriate motor signals.

The paper includes a synthetic bibliographic review of the neuroscience findings related to the task of vision-based reaching and grasping (Section 2). Neuroscience concepts are discussed and interpreted in order to build a coherent and comprehensive model of the integration between the two sorts of visual data, outlined in Section 3. Section 4 finally details those concepts that are directly useful for the generation of the integrated representation, starting from a real situation of an agent facing an environment within which it is expected to interact.

2 Neuroscience of Vision-Based Reaching and Grasping

The dualism between "vision for action" and "vision for perception" had been hypothesized long time before neuroimaging research [1]. Evidence for two distinct visual pathways having different roles and processing mechanisms has been provided during the last two decades by plenty of studies following different research approaches and techniques [2]. The **ventral stream** is dedicated to object recognition and classification, and works on a "scene-based" reference frame, in which size and location of an object are represented contextually with the size and location of nearby objects. The **dorsal stream** elaborates visual data in order to directly control object-directed actions, and thus follows an "actor-based" frame of reference, in which object location and size are represented with respect to the subject body, and especially to hand and arm.

The two streams hypothesis has been confirmed, but also criticized, by the neuroscientific community, and the original theory is constantly being revised and updated [3]. The trend is toward a more integrated view of the functioning of the two streams, that have in many cases complementary tasks, and the interaction between them seems to be extremely important for allowing both of them to function properly [2]. In this work we will deal especially with the more "pragmatic", action-oriented on-line processing of the dorsal stream, focused on the actual situation of the environment rather than on objects' implicit quality.

The brain areas more directly involved when a subject is interacting with his peripersonal space are briefly described below (refer to Figure 1). Visual data in primates flows from the retina to the lateral geniculate nucleus (LGN) of the thalamus, and then mainly to the primary visual cortex (V1) in the occipital lobe. The two main visual pathways go from V1 and the neighbor area V2 to the posterior parietal cortex (PPC) and the inferior temporal (IT) cortex. Object information flowing through the ventral pathway passes through V3 and V4 to the lateral occipital complex (LOC), that is in charge of object recognition. The dorsal pathway can be further subdivided in two parallel streams concerned respectively with movement of proximal (reaching) and distal joints (grasping). The dorso-medial pathway dedicated to reaching movements includes visual area V6, visuomotor area V6A and the medial intraparietal area (MIP). The two

latter areas project to the dorsal premotor cortex PMd [4]. For what concerns grasping, object related visual information flows through a dorso-lateral pathway including area V3A and the caudal intraparietal area (CIP), and then reaches the anterior intraparietal sulcus (AIP), the grasping area of the primate brain, which projects mainly to the ventral premotor area (PMv) [5]. Motor plans devised by PMd and PMv are sent to the primary motor cortex (M1) which release proper action execution signals.

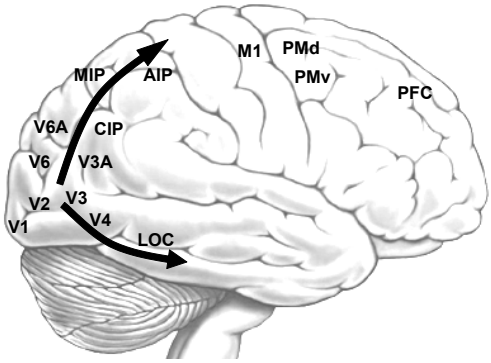

Fig. 1. The 2 visual pathways in the human brain (top arrow: dorsal; bottom arrow: ventral) with the areas involved in reaching and grasping actions.

3 Modeling the Interaction Between the Dorsal and Ventral Streams in Reaching and Grasping Actions

Comparing biologically-inspired robotic literature with computational models of vision-based reaching and grasping, it looks as they work on different assumptions and with different goals. On the one hand, biological or neuroscientific inspiration in robotics is often too superficial and conditioned by pragmatic goals and technological constraints. On the other hand, computational models are usually focused on specific issues and simulate low-level processes that are hard to scale in order to produce more complex behaviors.

Recent neuropsychological and neuroimaging research has shed a new light on how visuomotor coordination is organized and performed in the human brain. Thanks to such research, a model of vision-based arm movements which integrates knowledge coming from both monkey and human studies can now be developed. A previous model we developed [6] dealt mainly with grasping issues and the planning of suitable hand configurations and contacts on target objects, leaving mostly apart the transport component of the action. In this section we present an extended framework in which the process of reaching toward a visual target is thoroughly taken into account. The model we propose aims at an intermediate and really interdisciplinary solution that – while maintaining biological

plausibility, and the focus on neuroscience data, for the implementation of different visuomotor functions – provides the robot with the ability of performing purposeful, flexible and reliable vision-based reaching toward, and eventually grasping, nearby objects.

3.1 Complementary Roles of the Streams

Two kinds of properties have to be considered for a potential target object. Spatial properties related to its current situation, such as distance and pose, can only be assessed through actual estimation. Implicit properties like its size, weight and compliance are instead obtained through the integration of on-line, instantaneous visual information with memory of previously acquired knowledge about the object. These two sorts of properties are dealt with by the dorsal and the ventral streams, respectively. The complementary contribution of the two streams to the process of reaching and grasping is summarized in Table 1.

Table 1. Complementary tasks of the two streams.

Ventral stream	**Dorsal stream**
Object recognition	Visuomotor control
Global, invariant analysis	Local, feature analysis
Object weight, roughness, compliance	Object local shape, size
Object meaning	Object location
Previous experiences	Actual working conditions
Scene-based frame of reference	Effector-based frame of reference
Long-term representation	On-line computation

Many aspects affect the quantity and quality of tasks assigned to each stream in a given condition. In most cases the work partition between the streams is gradual, depending for example on action delay or on object familiarity [7, 8]. An explanation for this last case is that contribution of the ventral stream on action selection is modulated by the confidence achieved in the recognition of the target object. A higher confidence in object recognition reflects in a stronger influence of ventral stream data, such as knowledge of object weight and compliance. On the opposite, a more uncertain recognition leads to a more exploratory behavior, giving more importance to actual observation and dorsal analysis.

For identifying contact areas on the object surface in the case we want to act on the object (such as in grasping, pushing or pulling actions), additional constraints have to be taken into account. Usually, an estimation of the object center of mass affects the action plan. Such estimation relies on data coming from the ventral pathway, as the expected object composition and density. Similarly, surface texture and thus the expected contact friction, which affect the required grasping force, are ventral stream information. Extraction and integration of different kinds of object properties is a central issue in the present model.

3.2 Model Framework: a Subdivision Within the Dorsal Stream

Figure 2 shows the graphical schema of the whole model we propose. The fundamental data flow is the following. After the extraction of basic visual information in V1/V2, higher level features are generated in V3 and sent to the two streams. Along the ventral stream, an increasingly invariant representation of object shape is generated in order to perform a gradual recognition of the object (areas V4 and LOC [9]). In the dorsal stream, both object shape and location have to be processed. For what concerns shape, area CIP integrates stereoptic and perspective data in order to detect pose and proportion of the target object, using also information regarding object classification. Areas V6 and V6A estimate object location and distance, integrating retinal data with proprioceptive information about eye position. Both V3A and CIP project to AIP, which transforms object visual data in hand configurations suitable for grasping. At the same time, areas V6A and MIP determine the reaching direction and collaborate with AIP and PMd in order to execute the arm movement suitable to get to the target object. Grasping plans are devised by AIP in coordination with PMv, considering also the information on object identity coming from the ventral stream, and task requirements. Dorsal areas are supported by proprioceptive information coming from somatosensory areas SI/SII. The signals for action execution are sent to the motor cortex M1, and an AIP-PMv-Cerebellum loop is in charge of monitoring action execution in accordance to the plan.

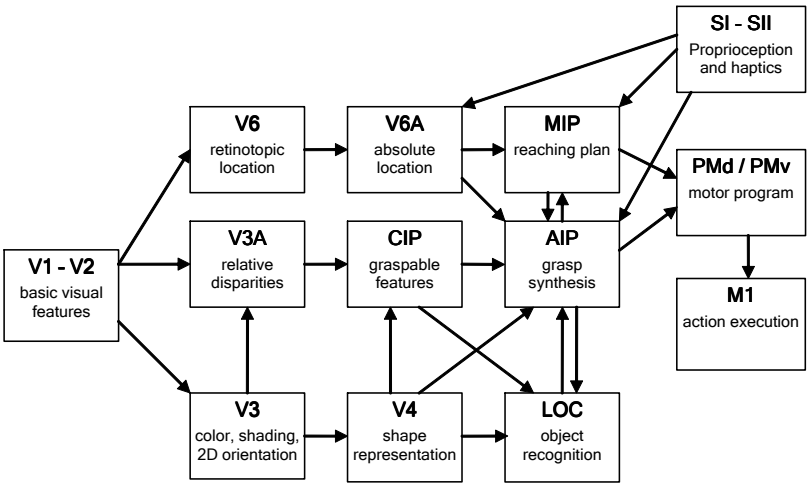

Fig. 2. Global model framework. The different information streams can be observed: the ventral stream V3-V4-LOC, the dorso-medial stream V6-V6A-MIP and the dorso-lateral stream V3A-CIP-AIP. Many more feedback connections are present, but not visualized for clarity reasons.

4 Obtaining an Integrated Representation of Reachable Objects

This section describes with more detail the sort of processing performed by the two streams and how an integrated representation of nearby objects, including perception-based and action-based aspects can be obtained. The following exposition includes neuroscience concepts, computational aspects and practical considerations, in order to gradually move from a purely theoretical to a prevalently applicative stance.

4.1 Processing of Basic Visual Information

Assumed that an object has been detected in the visual field, the first processing step is the extraction of fundamental visual data regarding the object. Starting from visual acquisition, an attentional mechanism is needed to focus on it, for isolating it from the background and from possible other objects. As in primates, vergence and version movements are executed in order to foveate the object, i.e. center it in the field of view so that its image is processed by the most sensitive section of the retina. Once the object is unambiguously identified and centered, visual elaboration can begin.

Visual areas V1 and V2 receive images and provide as output basic features, such as edges, corners, and absolute disparities. These features are used by downstream areas to build more complex ones. The most advanced visual representation common to both streams is a basic binocular description of the target object, composed for both eyes of its contour as a 2D silhouette and the retinal position of salient features, such as sharp corners. After this stage, the visual analysis is performed in parallel concurrent ways by the two pathways.

The ventral stream performs a gradual classification and identification of objects, probably through the integration of volumetric descriptions with 2D ones. On the other hand, the action-oriented dorsal processing is better done on descriptions of objects represented by 2D surfaces disposed in the 3D space. Color information, processed mainly by areas V3 and V4, can be used by the ventral stream to recognize objects more easily, and by the dorsal stream to track objects, but also to extract surface properties through shading and textures.

4.2 Dorsal Stream Processing

The description of visual object features relevant for reaching and grasping purposes is the next processing stage. The posterior parietal cortex, in charge of this task, does not construct any model or global representation of the object and the environment, but rather extract properties of visual features that are suitable for potential actions. In order to elaborate a proper action on an external target, two main inputs are required, the object shape and pose and its location with respect to the eyes and thus to the hand. These inputs are obtained by integrating retinal information regarding the object with proprioceptive data referred to

eyes, head and hand. All this information is managed contextually by the dorsal stream, through its two parallel sub-streams, dorso-medial and dorso-lateral.

Information regarding eye position and gaze is employed by V6A in order to estimate the position of surrounding objects and guide reaching movements toward them. Areas V6A and MIP seem to have a critical role in the gradual transformation from a retinotopic to an effector-centered frame of reference and in the modulation between visual data and proprioceptive information regarding gaze direction and arm position [10]. For what concerns the dorso-lateral stream and the control of distal joints, 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 and proportion of visual stimuli, represented in a viewer-centered way [11]. The evidence suggests that CIP integrates stereoptic and perspective cues for obtaining better estimates of visual targets. A possible interpretation of the job of CIP neurons is provided in a related work [12]. The 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 [13]. Despite the consolidated distinction between the cortical pathways for reaching and grasping, their tight interconnection is being proved by mutual projections between MIP and AIP, and by findings regarding V6A neurons involved in the execution of distal movements [14]. Indeed, the accomplishment of a complex visuomotor task such as graping requires a perfect coordination between proximal and distal joints and thus between the cortical areas that guide them.

Important for reaching and grasping movements is the estimation of object distance. Several areas of the dorsal stream are sensitive to the distance of a potential target, between others V6A, CIP and the lateral intraparietal sulcus LIP. Cues to distance estimation are retinal data, accommodation and vergence, this last being probably more influent in the dorsal stream, especially for grasping distances [2]. Psychophysiological experiments [15] 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. Computational modeling supports the hypothesis that such measure is more precise for close distances, and thus especially suitable for dealing with objects in the peripersonal space [16]. In the intraparietal sulcus, distance and disparity are processed together, the former acting as a gain modulation variable on the latter.This mechanism allows to properly interpret stereoscopic visual information [15].

4.3 Interactions Between the Streams

Visual processing in the ventral stream is based on the production of increasingly invariant representations aimed at object recognition. During grasping actions, ventral visual areas are in charge of identifying the object, and facilitating access to memorized properties which can be useful for the oncoming action. 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. Basic computational representations for this purpose are for example chain codes or 2D shape indexes.

Information on the basic shape of the object is probably forwarded to the dorsal stream, to CIP or AIP or both, to facilitate the feature extraction process. For example, if the object is recognized as roughly box-like, it can be assumed that its edges are parallel. Such assumption would facilitate the process of size and pose estimation, because reliable perspective estimation can be used in this case in addition to stereopsis.

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. In this way, 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.

Final output of the object recognition process are its identity and composition, which in turn allows to estimate its weight distribution and the roughness of its surface, that are valuable information at the moment of planning the action. Moreover, besides the recovery of memorized object properties, recognition allows to access stored knowledge regarding previous grasping experiences. Old actions on that object can be recalled and used to bias grasp selection, giving preference to learnt hand configurations which ended in successful action executions. Similarly to the classification confidence, the number and outcome of previous encounters with the same object will determine the reliability of the stored information.

5 Summary and Conclusions

Summarizing, a global, integrated representation of objects in the peripersonal space that takes into account both action-oriented and perception-oriented aspects should include all elements described in Table 2.

Computational models of the human visual system are largely available, especially for the first stages of visual processing, before the splitting of the two streams. At the same time, research on object recognition keeps involving a large part of the computer vision community. Nevertheless, few resources have been dedicated to the exploration of the mechanisms underlying the functioning of

Table 2. Elements of the integrated representation.

Ventral stream	
Object contour features	V2
Color/Texture	V3
Global contour representation	V2/V4
Global shape/Color	V4
Object class	V4
Object identity	LOC
Object meaning	LOC/PFC
Dorsal stream	
Absolute disparities	V1
Object contour features	V2
Relative disparities	V3
Local features	V3
Second order disparities	V3A
Features in 3D	V3A
Retinal location	V6
Absolute spatial location	V6A
Object distance	V6A/LIP
Object grasping features	CIP
Grasp synthesis	AIP
Motor program	PM

the action-related visual cortex, and the integration between the contributions of the two visual pathways is nearly unexplored at the computational level and even more in robotics. Thanks to recent neuroscience findings, the outline of a model of the brain mechanisms upon which vision-based reach and grasp planning relies could be drawn in this work. With respect to the available models, the proposed framework has been conceived to be applied on a robotic setup, and the analysis of the functions of each brain area has been performed taking into account not only biological plausibility, but also practical issues related to engineering constraints.

Previous works related to this model focused especially on the job of areas CIP and AIP. The next step in this research is to further develop and implement, first computationally and then on a robotic setup, the integration between stereoptic retinal data with somatosensory information about object and arm state, in order to estimate object position and devise a reaching action plan as performed by area V6A in the dorsal stream.

Acknowledgments

Support for this research has been provided in part by the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement 217077 (EYESHOTS project), by Ministerio de Ciencia y Innovación (FPU grant AP2007-02565) and by Fundació Caixa-Castello (project P1-1B2005-28).

References

1. A. D. Milner and M. A. Goodale. *The visual brain in action*. Oxford University Press, 1995.
2. M. A. Goodale and A. D. Milner. *Sight Unseen*. Oxford University Press, 2004.
3. G. Rizzolatti and M. Matelli. Two different streams form the dorsal visual system: anatomy and functions. *Experimental Brain Research*, 153(2):146–157, November 2003.
4. C. Galletti, D. F. Kutz, M. Gamberini, R. Breveglieri, and P. Fattori. Role of the medial parieto-occipital cortex in the control of reaching and grasping movements. *Experimental Brain Research*, 153(2):158–170, November 2003.
5. J. C. Culham, C. Cavina-Pratesi, and A. Singhal. The role of parietal cortex in visuomotor control: what have we learned from neuroimaging? *Neuropsychologia*, 44(13):2668–2684, 2006.
6. E. Chinellato, Y. Demiris, and A. P. del Pobil. Studying the human visual cortex for achieving action-perception coordination with robots. In A.P. del Pobil, editor, *Artificial Intelligence and Soft Computing*, pages 184–189. Acta Press, Anaheim, CF, USA, 2006.
7. M. Himmelbach and H.-O. Karnath. Dorsal and ventral stream interaction: contributions from optic ataxia. *The Journal of Cognitive Neuroscience*, 17(4):632–640, April 2005.
8. T. Sugio, K. Ogawa, and T. Inui. Neural correlates of semantic effects on grasping familiar objects. *Neuroreport*, 14(18):2297–2301, December 2003.
9. E. Chinellato, B. J. Grzyb, and A. P. del Pobil. Brain mechanisms for robotic object pose estimation. In *Intl. Joint Conf. on Neural Networks*, 2008.
10. N. Marzocchi, R. Breveglieri, C. Galletti, and P. Fattori. Reaching activity in parietal area V6A of macaque: eye influence on arm activity or retinocentric coding of reaching movements? *Eur J Neurosci*, 27(3):775–789, Feb 2008.
11. H. Sakata, M. Taira, M. Kusunoki, A. Murata, Y. Tanaka, and K. Tsutsui. Neural coding of 3D features of objects for hand action in the parietal cortex of the monkey. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1373):1363–1373, August 1998.
12. E. Chinellato and A. P. del Pobil. Neural coding in the dorsal visual stream. In *Intl. Conf. on the Simulation of Adaptive Behavior*, 2008.
13. E. Shikata, F. Hamzei, V. Glauche, R. Knab, C. Dettmers, C. Weiller, and C. Büchel. Surface orientation discrimination activates caudal and anterior intraparietal sulcus in humans: an event-related fMRI study. *Journal of Neurophysiology*, 85(3):1309–1314, 2001.
14. P. Fattori, R. Breveglieri, N. Marzocchi, D. Filippini, A. Bosco, and C. Galletti. Hand orientation during reach-to-grasp movements modulates neuronal activity in the medial posterior parietal area V6A. *J Neurosci*, 29(6):1928–1936, Feb 2009.
15. J. R. Tresilian and M. Mon-Williams. Getting the measure of vergence weight in nearness perception. *Experimental Brain Research*, 132(3):362–368, June 2000.
16. E. Chinellato and A. P. del Pobil. Distance and orientation estimation of graspable objects in natural and artificial systems. *Neurocomputing*, 72:879–886, 2008.