Project no.: FP7-ICT-217077
Project full title: Heterogeneous 3-D Perception across Visual Fragments
Project Acronym: EYESHOTS
Deliverable no: D4.1
Title of the deliverable: Merging perception-related and action-related visual information: description of integrated representation

Date of Delivery:	02 December 2008	
Organization name of lead contractor for this deliverable:	UJI	
Author(s):	E. Chinellato, E. Cervera, B. J. Grzyb, A. P. del Pobil	
Participant(s):	UJI, UG, UNIBO	
Workpackage contributing to the deliverable:	WP4	
Nature:	Report	
Version:	1.0	
Total number of pages:	18	
Responsible person:	Angel P. del Pobil	
Revised by:	P. Fattori, C. Galletti, S. Sabatini	
Start date of project:	1 March 2008	**Duration:** 36 months

Project Co-funded by the European Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Abstract:

This deliverable describes the research work performed by Eyeshots partners in order to achieve the main goal of Task 4.1: the description of an integrated representation of the peripersonal space, which includes both action-related and perception-related aspects. This constitutes the first step toward a robotic system highly-skilled in its capacity of exploring the nearby space. The current knowledge on the neuroscience of vision-based reaching and grasping in humans and other primates was described and analyzed, in order to establishing the computational bases for a robotic system able to achieve advanced skills in the interaction with close objects. The outcome is a representation of objects that includes on-line, action-oriented visual information with perceptual knowledge about objects and memories of previous interactions.

Contents

Executive summary

This deliverable describes the research work performed by the Robotic Intelligence Lab of Universitat Jaume I and its partners with regard to the first task of Work Package 4 of the Eyeshots project. The main goal of Task 4.1 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 form the distinction between sensorimotor and perceptual visual processing as performed by the two visual pathway of the primate cortex. The deliverable summarizes the work carried on so far by the consortium for achieving the goals of WP4.

The report 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 (Task 4.2), starting from a real situation of an agent facing an environment within which it is expected to interact. Summarizing, this report proposes a conceptual frame on which more detail modeling will be performed by related tasks in Eyeshots.

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 in the nearby world. Indeed, 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 Figure 1. The traditional distinction [33] 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 robotic systems in their exploration of the nearby space. The current knowledge on the neuroscience of vision-based reaching and grasping in humans and other primates was analyzed, in order to establishing the computational bases for a robotic system able to achieve advanced skills in the interaction with close objects. 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 outcome is a description 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).

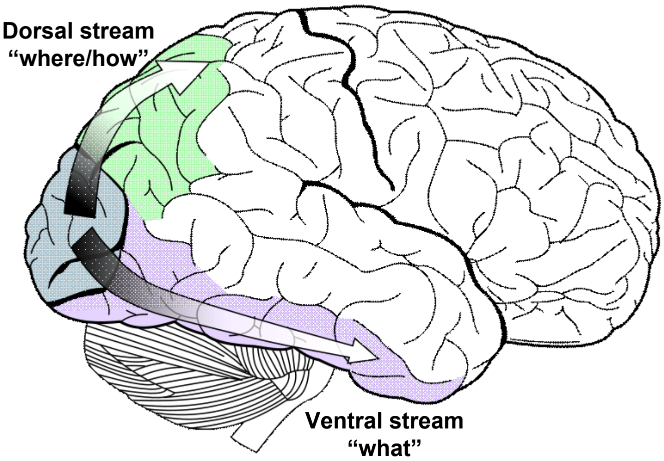

Figure 1: Dorsal and ventral visual streams.

2 Background

2.1 The 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 [33]. Studies with neurally-impaired people, especially on two categories of brain damages, *visual agnosia* and *optic ataxia*, suggested such dualism. Evidence for the different roles and processing mechanisms of the two pathways has been provided during the last two decades by plenty of studies following different research approaches and techniques. fMRI research showed the complementary responsiveness of the two streams in identification and spatial analysis of visual stimuli [57]. Behavioral studies based on optical illusions, distractor stimuli and concurrent tasks suggest that visual information is analyzed and processed differently by the streams. Explicit object perception in the ventral stream is "scene-based" and the size and location of an object is represented contextually with the size and location of nearby objects [60]. The control of object-directed actions by the dorsal stream follows instead 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. Dorsal visual analysis is driven by the absolute dimension and location of the target object, and only one object at a time can be the focus of attention for the dorsal stream, such that other objects in the environment are likely to be considered and hence taken into account only as potential obstacles [2].

Several studies demonstrated that ventral stream areas show adaptation for different views of the same object, denoting viewpoint invariance [26]. On the contrary, areas of the intraparietal sulcus do not exhibit such invariance, and respond to different views as they were different objects. This suggests a more "pragmatic", action-oriented on-line processing along the dorsal stream, focused on the actual situation of the environment rather than on objects' implicit quality. The streams dissociation has thus been confirmed, but also criticized, by the neuroscientific community, and the original theory is constantly being revised and updated [43, 18]. 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 [20].

The anatomy of the visual and motor cortices of human and closer superior primates is well known. Although the knowledge regarding associative regions of the brain, such as the posterior parietal or the inferior temporal cortices, is less established, it is possible to outline a simplified schema of the brain areas more directly involved when a subject is interacting with his peripersonal space. Those areas more thoroughly considered in this work are depicted in Figure 2 and very briefly described next.

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 sulcus (MIP). The two latter areas project to the dorsal premotor cortex PMd. 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). Motor plans devised by PMd and PMv are sent to the primary motor cortex (M1) which release proper action execution signals.

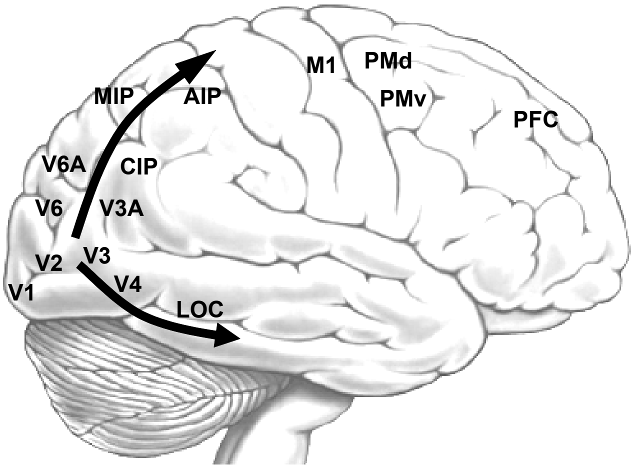

Figure 2: Brain areas involved in reaching and grasping actions.

2.2 Related approaches

Previous models of vision-based reaching and grasping have built so far mainly, when not exclusively, on monkey data. 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 single-cell monkey studies with human data can be developed. A previous model we developed [10] dealt mainly with grasping issues and the planning of suitable hand configurations and contacts on target objects, leaving apart the transport component of the action. We present here an extended framework in which the process of reaching toward a visual target is thoroughly taken into account.

Computational modeling of visual mechanisms in mammals is a wide and developed area [44] and strongly biologically inspired models of visual areas have been implemented, focusing mainly on the primary visual cortex [45], or on increasingly invariant object representations for recognition as performed along the ventral stream [37]. Complex agent-object interactions – such as in reaching and grasping actions – are not a usual target of computational models. Even more, the integration between the contributions of the two visual pathways is a subject virtually unexplored, and only a few studies explicitly deal with dorsal stream processing.

The most complete attempt to computationally describe the sensorimotor mechanisms of vision-based grasping in primates is the FARS model [15], which focuses on the interaction between AIP and the premotor cortex, and is oriented to the action execution part of the process. Other models distinguish between vision for action and vision for perception (VFP) and suggest a bidirectional interaction between them [29]. Cisek [11] proposes a computational model for action selection between optional reaching movements in which the decision is taken through the interaction of posterior parietal, prefrontal and premotor cortices.

Overall, comparing biologically-inspired robotic literature with the computational models regarding 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 complex behaviors, such as those we pursue in this work. 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 and grasping toward nearby objects.

3 Modeling the interaction between the dorsal and ventral streams in reaching and grasping actions

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 consistence 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 visual targets 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 rather progressive, depending for example on action delay or on object familiarity [23]. 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 Reaching and grasping

The correct coupling between the reaching and grasping movements, often neglected in robotic applications, is instead a fundamental and largely studied aspect in human grasping, and various plausible models on the relation between reaching and preshaping movements have been developed [48]. The hypothesis of parallel visuomotor channels for the transport and the preshaping components of the reach-to-grasp action is well recognized [27]. Anatomically, these two channels fall both inside the dorsal stream, and are sometimes named dorso-medial and dorso-lateral visuomotor channels. Cortical areas nomenclature is still controversial, and the correspondence between human and macaque studies not completely solved, but new studies confirm the duality of the reaching-grasping process [12]. For humans, the reaching area have been named PRR (parietal reach region) or SPOC (superior parietal-occipital cortex), but in this work we will stick to the more established monkey nomenclature, according to which the most important reach-related cortical areas are V6A and MIP, both receiving their main input from V6 and projecting to the dorsal premotor cortex[18, 16, 51].

Alternative models for the coupling between reaching and grasping, such as the *multiple finger reaching* idea [52], are not given much credit, due to the quantity and quality of evidence supporting the mainstream hypothesis [58]. In any case, the coupling between the two subsystems (and others) linking parietal and premotor cortex is tight, and they must share a common mechanism for coordinating with each other [27, 42]. Various computational models of reaching and grasping coordination that might be suited to robotic implementation are already available [34, 24].

3.3 Model framework

Figure 3 shows the graphical schema of the whole model. 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 invariant representation of object shape is generated in order to perform a gradual recognition of the object. Area V4 is in charge of classifying the object as pertaining to a given class, and only later full recognition is performed in LOC. 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 given by the prefrontal cortex, PFC. Dorsal areas are supported by proprioceptive information coming from somatosensory areas SI/SII, and action selection in AIP and PMv is modulated by the basal ganglia (BG). 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.

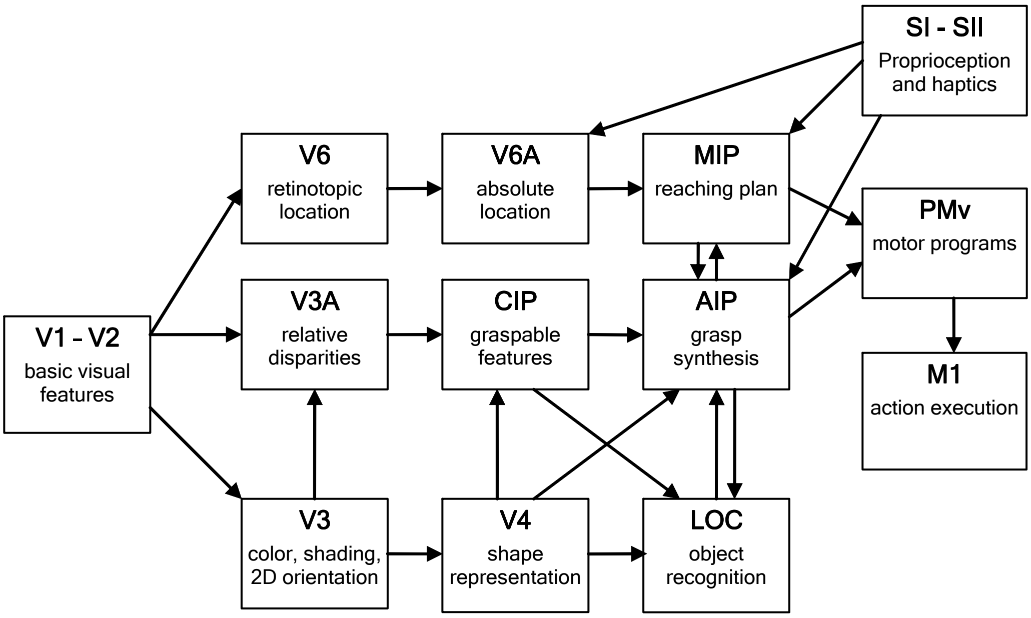

Figure 3: Global model framework. The different information streams can be observed: the ventral stream V3-V4-LOC, the dorso-medial stream V6-V6A-MIP and the dorso-lateral stream V3A-CIP-AIP. Many more feedback connections are present, but not visualized for clarity reasons.

Blocks in the diagram correspond to brain areas, according to neuroscience concepts, but also to implementable functional modules. In the following sections, basic concepts regarding the functioning, the role and the interactions of the brain areas which appear in the model are given. The basic interest is directed toward the integration of different types of information as it is thought to happen through the connections between the dorsal and the ventral streams.

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 by area V3, 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.

A fundamental aspect of dorsal elaboration is the use of binocular vision and stereopsis [5, 30]. 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. The basic mechanisms of stereoscopic vision have been studied for long time [28, 31], and neuronal responses to disparity stimuli in cortical visual areas have also been throughly investigated [38, 13]. Disparity detection is a fundamental aspect of visual processing that begins already in V1 and V2 [40]. 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 [1, 55] and in humans [3, 59]. For what concerns the processing of higher order disparities, there is a general consensus regarding a prominent role of V3A in representing relative disparities [3, 55, 6].

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. Two types of neurons have been found in V6A that allow to sustain this hypothesis [18]. The receptive fields of neurons of the first type are organized in retinotopic coordinates, but they can encode spatial locations thanks to gaze modulation. The receptive fields of the second type of neurons are organized according to the real, absolute distribution of the subject peripersonal space. In addition, V6A contains neurons that arguably represent the target of reaching retinocentrically, and others that use a spatial representation [32]. This strongly suggests a critical role of V6A in the gradual transformation from a retinotopic to an effector-centered frame of reference. Moreover, some V6A neurons appear to be directly involved in the execution of reaching movements [16, 17], indicating that this area is in charge (probably together with MIP) of performing the visuomotor transformations required for the purposive control of proximal arm joints, integrating visual, somatosensory and somatomotor signals in order to reach a given target in the 3D space.

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 [46, 53]. The evidence suggests that CIP integrates stereoptic and perspective cues for obtaining better estimates of visual targets [56, 59]. Features extracted by CIP are coded using the response properties of surface orientation selective (SOS) and axis orientation selective (AOS) neurons, which are selective for the orientation and relative proportion of flat and elongated parts respectively [46, 50]. Some responsiveness to actual object size has also been observed, suggesting that at least a part of AOS and SOS neurons code for absolute object dimensions. A possible interpretation of the job of these CIP neurons is provided in [9]. 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 [47, 49].

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 [17]. 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 th 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 [20]. Psychophysiological experiments [54] 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. Such measure is more precise for close distances, and thus especially suitable for dealing with objects in the peripersonal space (for a model of this process, see[8]). In the intraparietal sulcus, distance and disparity are processed together, the former acting as a gain modulation variable on the latter [19]. This mechanism allows to properly interpret stereoscopic visual information [35].

4.3 Ventral stream processing

Object recognition in the ventral stream is performed gradually and hierarchically [22, 4]. Recent findings indicate that object recognition is composed of at least two subsequent stages, categorization and identification [21]. In the first stage, an object is classified as belonging to a given class or family of objects. Such process is strikingly fast, probably ensuring that there is time for feeding 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 Figure 3. Anatomical and functional evidence supports this early integration between the streams. 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. At least for the first classification stage, visual input is very likely compared to memorized 2D representations [7, 36]. A classification based on 3D representations would require mental rotation, and this can hardly be performed with the quickness observed in neuroscience experiments [21]. 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 [25].

Various biologically inspired methods for object recognition have been developed in computer vision, and different models of ventral stream processing are available [37, 41]. Some of them are strongly inspired by neuroscience findings, and use plausible approaches such as radial basis function networks [39, 14]. For the purposes of this work, object recognition is functional to reaching and grasping actions, and the interest is not in detailed modeling of ventral stream mechanisms.

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. Moreover, 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.

4.4 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.

4.5 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.

Table 2: Elements of the integrated representation.

Ventral stream	
Object contour features	V2
Global contour representation	V2/V4
Global shape	V4
Color/Texture	V3
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
Action plan	AIP

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 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.

The next step in this research work is to further develop and implement, first computationally and then on a robotic setup, some parts of the model. Special focus will be initially put on 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.

Bibliography

[1] D. L. Adams and S. Zeki. Functional organization of macaque V3 for stereoscopic depth. *Journal of Neurophysiology*, 86(5):2195–2203, November 2001.

[2] C. Ansuini, V. Tognin, L. Turella, and U. Castiello. Distractor objects affect fingers' angular distances but not fingers' shaping during grasping. *Experimental Brain Research*, 178(2):194–205, April 2007.

[3] B. T. Backus, D. J. Fleet, A. J. Parker, and D. J. Heeger. Human cortical activity correlates with stereoscopic depth perception. *Journal of Neurophysiology*, 86(4):2054–2068, October 2001.

[4] M. Bar, R. B. Tootell, D. L. Schacter, D. N. Greve, B. Fischl, J. D. Mendola, B. R. Rosen, and A. M. Dale. Cortical mechanisms specific to explicit visual object recognition. *Neuron*, 29(2):529–535, February 2001.

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

[6] G. J. Brouwer, R. van Ee, and J. Schwarzbach. Activation in visual cortex correlates with the awareness of stereoscopic depth. *Journal of Neuroscience*, 25(45):10403–10413, November 2005.

[7] H. H. Bülthoff, S. Y. Edelman, and M. J. Tarr. How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5(3):247–260, 1995.

[8] E. Chinellato and A. P. del Pobil. Distance and orientation estimation of graspable objects in natural and artificial systems. *Neurocomputing*, In Press, 2008.

[9] E. Chinellato and A. P. del Pobil. Neural coding in the dorsal visual stream. In *International Conference on the Simulation of Adaptive Behavior*, 2008.

[10] E. Chinellato, Y. Demiris, and A. P. del Pobil. Studying the human visual cortex for achieving action-perception coordination with robots. In *IASTED International Conference on Artificial Intelligence and Soft Computing*, 2006.

[11] P. Cisek. A computational model of reach decisions in the primate cerebral cortex. In *Modeling Natural Action Selection*, 2005.

[12] J. C. Culham, J. P. Gallivan, C. Cavina-Pratesi, and D. J. Quinlan. fMRI investigations of reaching and ego space in human superior parieto-occipital cortex. In M. Behrmann & B. MacWhinney R. Klatzky, editor, *In Press*. Lawrence Erlbaum Associates, 2008.

[13] B. G. Cumming and G. C. DeAngelis. The physiology of stereopsis. *Annual Review of Neuroscience*, 24:203–238, 2001.

[14] S. Deneve and A. Pouget. Basis functions for object-centered representations. *Neuron*, 37(2):347–359, January 2003.

[15] A. H. Fagg and M. A. Arbib. Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks*, 11(7-8):1277–1303, October 1998.

[16] P. Fattori, M. Gamberini, D. F. Kutz, and C. Galletti. 'arm-reaching' neurons in the parietal area v6a of the macaque monkey. *Eur J Neurosci*, 13(12):2309–2313, Jun 2001.

[17] Patrizia Fattori, Rossella Breveglieri, Katia Amoroso, and Claudio Galletti. Evidence for both reaching and grasping activity in the medial parieto-occipital cortex of the macaque. *European Journal of Neuroscience*, 20(9):2457–2466, November 2004.

[18] 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.

[19] A. Genovesio and S. Ferraina. Integration of retinal disparity and fixation-distance related signals toward an egocentric coding of distance in the posterior parietal cortex of primates. *Journal of Neurophysiology*, 91(6):2670–2684, June 2004.

[20] M. A. Goodale and A. D. Milner. *Sight Unseen.* Oxford University Press, 2004.

[21] K. Grill-Spector and N. Kanwisher. Visual recognition: as soon as you know it is there, you know what it is. *Psychological Science*, 16(2):152–160, February 2005.

[22] K. Grill-Spector, T. Kushnir, T. Hendler, S. Edelman, Y. Itzchak, and R. Malach. A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human Brain Mapping*, 6(4):316–328, 1998.

[23] 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.

[24] Y. Hu, R. Osu, M. Okada, M. A. Goodale, and M. Kawato. A model of the coupling between grip aperture and hand transport during human prehension. *Experimental Brain Research*, 167(2):301–304, November 2005.

[25] K. H. James, G. K. Humphrey, and M. A. Goodale. Manipulating and recognizing virtual objects: where the action is. *Canadian Journal of Experimental Psychology*, 55(2):111–120, June 2001.

[26] T.W. James, G.K. Humphrey, J.S. Gati, R.S. Menon, and M.A. Goodale. Differential effects of viewpoint on object-driven activation in dorsal and ventral streams. *Neuron*, 35(4):793–801, August 2002.

[27] M. Jeannerod. Visuomotor channels: Their integration in goal-directed prehension. *Human Movement Science*, 18(2):201–218, June 1999.

[28] B. Julesz. *Foundations of cyclopean perception.* MIT Press, 1971.

[29] M. A. Lebedev and S. P. Wise. Insights into seeing and grasping: Distinguishing the neural correlates of perception and action. *Behavioral and Cognitive Neuroscience Reviews*, 1(2):108–129, 2002.

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

[31] D. Marr. *Vision: a computational investigation into the human representation and processing of visual information.* W. H. Freeman, 1982.

[32] 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.

[33] A. D. Milner and M. A. Goodale. *The visual brain in action.* Oxford University Press, 1995.

[34] M. Mon-Williams and J. R. Tresilian. A simple rule of thumb for elegant prehension. *Current Biology*, 11(13):1058–1061, July 2001.

[35] M. Mon-Williams, J. R. Tresilian, and A. Roberts. Vergence provides veridical depth perception from horizontal retinal image disparities. *Experimental Brain Research*, 133(3):407–413, August 2000.

[36] G. A. Orban, P. Janssen, and R. Vogels. Extracting 3D structure from disparity. *Trends in Neurosciences*, 29(8):466–473, August 2006.

[37] R. C. O'Reilly and Y. Munakata. *Computational Explorations in Cognitive Neuroscience - Understanding the Mind by Simulating the Brain.* MIT Press, 2000.

[38] G. F. Poggio, F. Gonzalez, and F. Krause. Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *Journal of Neuroscience*, 8(12):4531–4550, December 1988.

[39] S. Pouget and A. Sejnowski. Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, 9(2):222–237, 1997.

[40] J. Read. Early computational processing in binocular vision and depth perception. *Progress in Biophysics & Molecular Biology*, 87(1):77–108, January 2005.

[41] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3:1199–1204, November 2000.

[42] G. Rizzolatti and G. Luppino. The cortical motor system. *Neuron*, 31(6):889–901, September 2001.

[43] Giacomo Rizzolatti and Massimo Matelli. Two different streams form the dorsal visual system: anatomy and functions. *Experimental Brain Research*, 153(2):146–157, November 2003.

[44] E.T. Rolls and G. Deco. *Computational Neuroscience of Vision.* Oxford University Press, Oxford, UK, 2002.

[45] S. P. Sabatini, F. Solari, G. Andreani, C. Bartolozzi, and G. M. Bisio. A hierarchical model of complex cells in visual cortex for the binocular perception of motion-in-depth. In *Advances in Neural Information Processing Systems*, pages 1271–1278, 2001.

[46] 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.

[47] H. Sakata, M. Taira, M. Kusunoki, A. Murata, K. Tsutsui, Y. Tanaka, W. N. Shein, and Y. Miyashita. Neural representation of three-dimensional features of manipulation objects with stereopsis. *Experimental Brain Research*, 128(1-2):160–169, September 1999.

[48] R. Shadmehr and S. P. Wise. *The computational neurobiology of reaching and pointing: A foundation for motor learning.* MIT Press, 2005.

[49] 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.

[50] E. Shikata, Y. Tanaka, H. Nakamura, M. Taira, and H. Sakata. Selectivity of the parietal visual neurones in 3D orientation of surface of stereoscopic stimuli. *Neuroreport*, 7(14):2389–2394, October 1996.

[51] S. Shipp, M. Blanton, and S. Zeki. A visuo-somatomotor pathway through superior parietal cortex in the macaque monkey: cortical connections of areas v6 and v6a. *Eur J Neurosci*, 10(10):3171–3193, Oct 1998.

[52] J. B. J. Smeets, E. Brenner, and M. Biegstraaten. Independent control of the digits predicts an apparent hierarchy of visuomotor channels in grasping. *Behavioural Brain Research*, 136(2):427–432, November 2002.

[53] M. Taira, K. I. Tsutsui, M. Jiang, K. Yara, and H. Sakata. Parietal neurons represent surface orientation from the gradient of binocular disparity. *Journal of Neurophysiology*, 83(5):3140–3146, May 2000.

[54] 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.

[55] D. Y. Tsao, W. Vanduffel, Y. Sasaki, D. Fize, T. A. Knutsen, J. B. Mandeville, L. L. Wald, A. M. Dale, B. R. Rosen, D. C. Van Essen, M. S. Livingstone, G. A. Orban, and R. B. H. Tootell. Stereopsis activates V3A and caudal intraparietal areas in macaques and humans. *Neuron*, 39(3):555–568, July 2003.

[56] K.-I. Tsutsui, M. Taira, and H. Sakata. Neural mechanisms of three-dimensional vision. *Neuroscience Research*, 51(3):221–229, March 2005.

[57] K. F. Valyear, J. C. Culham, N. Sharif, D. Westwood, and M. A. Goodale. A double dissociation between sensitivity to changes in object identity and object orientation in the ventral and dorsal visual streams: A human fMRI study. *Neuropsychologia*, 44(2):218–228, 2006.

[58] C. van de Kamp and F. T. J. M. Zaal. Prehension is really reaching and grasping. *Experimental Brain Research*, 182(1):27–34, September 2007.

[59] A. E. Welchman, A. Deubelius, V. Conrad, H. H. Bülthoff, and Z. Kourtzi. 3D shape perception from combined depth cues in human visual cortex. *Nature Neuroscience*, 8(6):820–827, June 2005.

[60] D. A. Westwood and M. A. Goodale. A haptic size-contrast illusion affects size perception but not grasping. *Experimental Brain Research*, 153(2):253–259, November 2003.