UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Visuomotor Coordination in Reach–to–Grasp Tasks: From Humans to Humanoids and Vice Versa

Luka Lukic

Supervisor: Doctor José Alberto Rosado dos Santos-Victor

Co-Supervisor: Doctor Aude Billard

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury
Chairperson: Chairman of the IST Scientific Board

Members of the Committee:
Doctor David Vernon
Doctor Giulio Sandini
Doctor Aude Billard
Doctor José Alberto Rosado dos Santos-Victor
Doctor Estela Guerreiro da Silva Bicho Erlhagen
Doctor Alexandre José Malheiro Bernardino

2015

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Visuomotor Coordination in Reach–to–Grasp Tasks: From Humans to Humanoids and Vice Versa

LUKA LUKIC

Supervisor: Doctor José Alberto Rosado dos Santos-Victor

Co-Supervisor: Doctor Aude Billard

Thesis approved in public session to obtain the PhD Degree in
Electrical and Computer Engineering

Jury final classification: Pass with Distinction

Jury

Chairperson: Chairman of the IST Scientific Board

Members of the Committee:

Doctor David Vernon, Professor, School of Informatics, University of Skövde, Sweden
Doctor Giulio Sandini, Professore Ordinario, Università degli Studi di Genova, Italy
Doctor Aude Billard, Full Professor, École Polytechnique Fédérale de Lausanne, Switzerland
Doctor José Alberto Rosado dos Santos-Victor, Professor Catedrático do Instituto Superior Técnico, da Universidade de Lisboa
Doctor Estela Guerreiro da Silva Bicho Erlhagen, Professora Associada da Escola de Engenharia da Universidade do Minho
Doctor Alexandre José Malheiro Bernardino, Professor Associado do Instituto Superior Técnico, da Universidade de Lisboa

2015

Abstract

Understanding the principles involved in visually-based coordinated motor control is one of the most fundamental and most intriguing research problems across a number of areas, including psychology, neuroscience, computer vision and robotics. Humans perform visually driven actions such looking at, reaching, and grasping a morning cup of coffee on a daily basis, without much effort and still very reliably. Yet, not very much is known regarding computational functions that the central nervous system performs in order to provide a set of requirements for visually-driven reaching and grasping. Additionally, in spite of several decades of advances in the field, the abilities of humanoids to perform similar tasks are by far modest when needed to operate in unstructured, unpredictable and dynamically changing environments.

In this thesis, we are interested in studying the principles behind the transformations from the retinotopic target encoding to the representations that are used to generate eye-head and arm movements. Next, we study how the movements of the eyes, arm and hand are generated and coordinated in reach-to-grasp tasks. In addition to this, we investigate the tailoring of visual resources with respect to spatio-temporal requirements of the motor system. We start from studying the visuomotor principles in humans and monkeys and further proceed with investigating how they can be useful to robotic applications. Once we create our computational models, we are able to go in the *backward direction*, from robotics to neuroscience, by providing some hypotheses and predictions regarding the functions of the central nervous system.

More specifically, our first focus is understanding the principles involved in human visuomotor coordination. Not many behavioral studies considered visuomotor coordination in natural, unrestricted, head-free movements in complex scenarios such as obstacle avoidance. To fill this gap, we provide an assessment of visuomotor coordination when humans perform prehensile tasks with obstacle avoidance, an issue that has received far less attention. Namely, we quantify the relationships between the gaze and arm-hand systems, so as to inform robotic models, and we investigate how the presence of an obstacle modulates this pattern of correlations.

Second, to complement these observations, we provide a robotic model of visuomotor coordination, with and without the presence of obstacles in the workspace. The parameters of the controller are solely estimated by using the human motion capture data from our human study. This controller has a number of interesting properties. It provides an efficient way to control the gaze, arm and hand movements in a stable and coordinated manner. When facing perturbations while reaching and grasping, our controller adapts its behavior almost instantly, while preserving coordination between the gaze, arm, and hand.

Furthermore, in the third part of the thesis, we study the neuroscientific literature of the primates (including humans). We here stress the view that the cerebellum uses the cortical reference frame representation. The cerebellum by taking into account this representation performs closed-loop programming of multi-joint, compound movements and movement synchronization between the eye-head system, arm and hand. Based on this investigation, we propose a functional architecture of the cerebellar-cortical involvement. Based on our theoretical work, we derive a number of improvements of our visuomotor controller for obstacle-free reaching and grasping. Because this model is devised by carefully taking into account the neuroscientific evidence, we are able to provide a number of testable predictions about the functions of the central nervous system in visuomotor coordination.

Finally, in the last part of the thesis, we tackle the flow of the visuomotor coordination in the direction from the arm-hand system to the visual system. We develop two models of motor-primed attention for humanoid robots. Motor-priming of attention is a mechanism that implements prioritizing of visual processing with respect to motor-relevant parts of the visual field. Recent studies in humans and monkeys have shown that visual attention supporting natural behavior is not exclusively defined in terms of visual saliency in color or texture cues (which is a predominant premise of the majority of attentional models), rather the reachable space and motor plans present the predominant source of this attentional modulation. In this thesis, we show that motor-priming of visual attention can be used to very efficiently distribute robot's computational resources devoted to visual processing.

We have validated our models with the humanoid robot iCub, in simulation and with the real-world robot platform. We believe that the work presented in this thesis represents a contribution relevant to both robotics and cognitive science.

Keywords: active vision, coupled dynamical systems, gaze, humanoid robot, learning, motor control, motor-primed visual attention, neuroscience, reaching and grasping, visuomotor coordination

Resumo

Compreender os princípios envolvidos no controlo motor baseado na visão é um dos problemas de investigação mais fundamentais e intrigantes num conjunto de áreas que inclui a psicologia, a neurociência, a visão computacional e a robótica. Os seres humanos executam ações guiadas visualmente, como olhar, alcançar, e agarrar uma chávena de café diariamente, sem muito esforço e com grande fiabilidade. No entanto, não se sabe muito sobre as funções computacionais que o sistema nervoso central realiza a fim de proporcionar um conjunto de requisitos para alcançar e agarrar objetos recorrendo à visão. Além disso, apesar de várias décadas de avanços na área, a capacidade actual dos humanóides para executar tarefas semelhantes é de longe bastante modesta quando estas tarefas são realizadas em ambientes não estruturados, imprevisíveis e dinâmicos.

Nesta tese, estamos interessados em estudar os princípios por trás das transformações da codificação retinotópica do alvo para as representações que são utilizadas para gerar os movimentos dos olhos, cabeça e braço. Em seguida, vamos estudar como os movimentos dos olhos, do braço e da mão são gerados e coordenados em tarefas de alcançar-para-agarrar. Além disso, investigamos a adaptação de recursos visuais no que diz respeito aos requisitos espaciotemporais do sistema motor, partindo do estudo dos princípios visuomotores em humanos e macacos e continuando com a investigação de como estes podem ser úteis para aplicações robóticas. Uma vez criados os nossos modelos computacionais, somos capazes de ir na direção inversa, a partir da robótica para a neurociência, fornecendo algumas hipóteses e previsões sobre as funções do sistema nervoso central.

Mais especificamente, o nosso primeiro foco é entender os princípios envolvidos na coordenação visuomotora humana. Não são muitos os estudos comportamentais que consideraram a coordenação visuomotora em movimentos naturais, livres, com a cabeça liberta, em cenários complexos, tais como o desvio de obstáculos. Para preencher essa lacuna, fornecemos uma avaliação da coordenação visuomotora quando os seres humanos executam tarefas preênseis como o desvio de obstáculos, uma questão que tem recebido muito menos atenção. Nomeadamente, quantificamos as relações entre os sistemas de olhar e braço-mão, de modo a construir modelos robóticos e investigar como a presença de um obstáculo modula esse padrão de correlações.

Em segundo lugar, para complementar estas observações, fornecemos um modelo robótico de coordenação visuomotora, com e sem a presença de obstáculos na área de trabalho. Os parâmetros do controlador são apenas estimados usando os dados de captura de movimentos humanos do nosso estudo humano. Este controlador tem uma série de propriedades interessantes, sendo capaz de fornecer uma maneira eficiente de controlar os movimentos do olhar, do braço e da mão de uma forma estável e coordenada. Na presença de perturbações durante o alcance, o controlador adapta o seu comportamento quase instantaneamente, preservando ao mesmo tempo a coordenação entre olhar, braço e mão.

Além disso, na terceira parte da tese, estudamos a literatura neurocientífica sobre os primatas (incluindo os humanos). Aqui frisamos a visão de que o cerebelo usa a representação cortical referencial. O cerebelo, levando em conta essa representação realiza uma programação em circuito fechado de movimentos multi-articulares, movimentos compostos e sincronização de movimentos entre o sistema de olho-cabeça, braço e mão. Com base nessa investigação, propomos uma arquitetura funcional do envolvimento cerebelar-cortical. Com base no trabalho teórico, derivamos uma série de melhorias do nosso controlador visuomotor para alcançar e agarrar na ausência de obstáculos. Como este modelo é concebido com uma cuidada aderência a resultados da neurociência, somos capazes de fornecer um número de predições testáveis sobre as funções do sistema nervoso central no que diz respeito à coordenação visuomotora.

Finalmente, na última parte da tese, abordamos o fluxo da coordenação visuomotora na direção do sistema braço-mão para o sistema visual. Desenvolvemos dois modelos de preparação motora de atenção para robôs humanóides. A preparação motora de atenção é um mecanismo que implementa a priorização do processamento visual em relação a partes relevantes do ponto de vista motor do campo visual. Estudos recentes em humanos e macacos demonstraram que a atenção visual suportando comportamentos naturais não é exclusivamente definida em termos de saliência visual, cor ou textura (o que é uma premissa predominante da maioria dos modelos de atenção), pelo contrário, o espaço acessível e planos motores são a fonte predominante desta modulação de atenção. Nesta tese, mostramos que a preparação motora da atenção visual pode ser usada para distribuir de forma muito eficiente os recursos computacionais de um robô dedicados ao processamento visual.

Validámos os nossos modelos com o robô humanóide iCub, em simulação e com a plataforma física do robô. Acreditamos que o trabalho apresentado nesta tese representa uma contribuição relevante tanto para a robótica como para a ciência cognitiva.

Palavras-chave: visão activa, sistemas dinâmicos acoplados, olhar, robô humanóide, aprendizagem, controlo motor, preparação motora da atenção visual, neurociência, alcançar e agarrar, coordenação visuomotora

Riassunto

Comprendere i principi applicabili nel controllo motorio basato sulla visione è uno dei problemi di ricerca fondamentali e tra i più affascinanti nell'ambito di diverse discipline, tra cui psicologia, neuroscienza, visione artificiale e robotica. Gli esseri umani sono in grado di compiere azioni guidate dalla visione come guardare, raggiungere e afferrare una tazzina di caffè quotidinamente, senza sforzo, ma comunque in modo alquanto preciso. Tuttavia, non si sa molto riguardo alle funzioni computazionali che il sistema nervoso centrale svolge per fornire un insieme di requisiti ai compiti di raggiungimento e afferramento guidati dalla visione. Inoltre, nonostante molti anni di progressi sul campo, l'abilità degli umanoidi nell'effettuare compiti analoghi è finora modesta, quando occorre che operino in ambienti non strutturati, imprevedibili e che evolvono dinamicamente.

In questa tesi ci interessiamo allo studio dei principi che regolano le trasformazioni dalla codifica retinotopica dell'obiettivo visuale alle rappresentazioni impiegate per generare movimenti occhio-testa e occhio-braccio. Inoltre studiamo in che modo i movimenti di occhi, braccia e mani sono generati e coordinati durante compiti di raggiungimento e afferramento. Inoltre, esaminiamo come le risorse visuali vengano adattate ai requisiti spazio-temporali del sistema motorio. Iniziando dallo studio dei principi visuo-motori negli esseri umani e nelle scimmie, proseguiamo analizzando come tali principi possano essere utili per applicazioni robotiche. Dopo aver costruito dei modelli computazionali, possiamo procedere nella direzione opposta, dalla robotica alla neuroscienza, formulando ipotesi e previsioni riguardanti le funzioni del sistema nervoso centrale.

Nello specifico, per prima cosa ci siamo concentrati sui principi applicabili alla coordinazione visuo-motoria umana. Pochi sono gli studi comportamentali che hanno considerato la coordinazione visuo-motoria in movimenti naturali, senza restrizioni e in cui la testa è libera di muoversi in scenari complessi, come l'aggiramento di ostacoli. Al fine di colmare questa lacuna, forniamo una valutazione della coordinazione visuo-motoria da parte di esseri umani durante compiti di afferramento con aggiramento di ostacoli: questo è un tema che ha ricevuto assai meno attenzione in letteratura. In particolare, quantifichiamo le relazioni tra direzione dello sguardo e sistemi comprensivi di braccio e mano, così da fornire informazioni a modelli robotici, e approfondiamo in che modo la presenza di ostacoli influenza questo tipo di correlazioni.

In secondo luogo, per completare le osservazioni di cui sopra, forniamo un modello robotico di coordinazione visuo-motoria, con e senza la presenza di ostacoli nello spazio di lavoro. I parametri del controllore vengono stimati usando solamente dati di *motion capture* derivanti dal nostro studio effettuato con esseri umani. Questo controllore mostra delle proprietà interessanti: consente di controllare in maniera efficiente i movimenti di sguardo, braccia e mani in modo stabile e coordinato. Quando si imbatte in perturbazioni durante il raggiungimento e l'afferramento di oggetti, il nostro controllore adatta il suo comportamento quasi istantaneamente, mantenendo tuttavia la coordinazione tra sguardo, braccia e mani.

Inoltre, nella terza parte della tesi, studiamo la letteratura di neuroscienza riguardante i primati (compresi gli esseri umani). Insistiamo sulla prospettiva secondo cui il cervelletto usa la rappresentazione del sistema di riferimento corticale. Nel considerare questa rappresentazione, il cervelletto compie un controllo ad anello chiuso di movimenti multi-giunto e composti, oltre a occuparsi della sincronizzazione dei movimenti tra il sistema occhi-testa, le braccia e le mani. In base a questa analisi, proponiamo un'architettura funzionale del coinvolgimento tra cervelletto e corteccia. Partendo dal nostro lavoro teorico, apportiamo una serie di migliorie al nostro controllore visuo-motorio durante i compiti di raggiungimento e afferramento privi di ostacoli. Dal momento che questo modello è sviluppato da un'attenta considerazione delle prove neuroscientifiche, possiamo fornire una serie di predizioni verificabili riguardo ai ruoli che il sistema nervoso centrale ricopre nella coordinazione visuo-motoria.

Infine, nell'ultima parte della tesi, affrontiamo il problema del flusso di coordinazione visuo-motoria percorrendo il verso che va dal sistema braccia-mani verso il sistema visuale. Sviluppiamo due modelli di attenzione stimolata dai movimenti (*motor-primed attention*) per robot umanoidi. La stimolazione di attenzione basata sui movimenti è un meccanismo che fa attribuire un ordine di priorità dell'elaborazione visuale in relazione alle parti del campo visivo con rilevanza in termini di movimento. Recenti studi su esseri umani e scimmie hanno mostrato che l'attenzione visuale supportata da comportamenti naturali non è esclusivamente definita in termini di salienza visuale di segnali legati al colore o al materiale (questi sono i presupposti della maggior parte dei modelli di attenzione in letteratura), mentre invece la fonte predominante di questa modulazione dell'attenzione è costituitita dallo spazio di lavoro raggiungibile e dai piani motòri. In questa tesi, mostriamo che la stimolazione motoria dell'attenzione visuale può essere usata per distribuire in modo efficiente quelle risorse computazionali che nei robot sono dedicate al processamento visuale.

Abbiamo convalidato i nostri modelli sul robot umanoide iCub, in simulazione e con la piattaforma robotica reale. Riteniamo che il lavoro presentato in questa tesi rappresenti un contributo rilevante sia per la robotica che per le scienze cognitive.

Parole chiave: visione attiva, sistemi dinamici accoppiati, direzione dello sguardo, robot umanoidi, apprendimento, controllo motorio, stimolazione di attenzione basata sui movimenti (*motor-primed attention*), neuroscienza, raggiungimento e afferramento, coordinazione visuo-motoria.

To all people who have supported me,
to my mother and my brother
and to Ranka

Acknowledgments

I would like to express my gratitude to Prof. Aude Billard, my EPFL thesis advisor. I sincerely appreciate her scientific advices and her patience during the time when I was still "chasing" challenges of my research at the beginning of my Ph.D. studies. She has provided numerous valuable comments on the research I have done and the papers I have written. Next, I would like to thank Prof. José Santos-Victor, my IST thesis advisor, for giving me a lot of encouragement to explore scientific questions that have interested me, for providing a constant supply of optimism, and for sharing the same appreciation of biologically-inspired techniques in robotics. I should thank him for regularly updating me about Serbian football players in S.L. Benfica, as well. I am grateful to both of my advisors for making this thesis possible and for keeping high standards of research, which has greatly stimulated my personal development.

Enlisting all people from EPFL and IST that have deserved to be mentioned in this acknowledgment would make a very long list, and I am afraid that I would forget somebody. At both places, I have been privileged to be around outstandingly brilliant, funny and interesting people coming from diverse geographical origins. I truly thank them for their friendship, many interesting discussions about various topics and many funny moments on various occasions.

However, I must particularly thank a selected group people who have directly helped me during my thesis. I would like to thank Basilio Noris, for his machine learning and gaze tracking advices, to Ashwini Shukla, for discussions on the iCub and CDS, to Lorenzo Jamone, for providing several insightful comments on some papers I wrote, to Bruno Damas, for his input on the IMLE, to Prof. Alexandre Bernardino and Prof. José Gaspar, for their useful advices.

I would like to thank to my mother Milica and my brother Marko, for being my "core team" and my fiercest supporters. They have always unconditionally stood by my side in many challenging moments.

Finally, but most warmly, I would like to thank to my wife Ranka, for her love, patience, support, encouragement, many nice moments... for everything.

This work was funded by a Fundação para a Ciência e a Tecnologia (FCT) doctoral grant (SFRH/BD/51072/2010) under the IST-EPFL Joint Doctoral Initiative.

Table of Contents

1 Introduction 1
1.1 Motivation 1
1.2 Thesis outline 4

2 Human Motion Study of Reaching and Grasping with Obstacle Avoidance . . 7
2.1 Background research 8
2.1.1 The general role of visual information in guiding reaching and grasping 8
2.1.2 Visuomotor coordination in reaching and grasping 9
2.2 Human Motion Study of Reaching and Grasping with Obstacle Avoidance 10
2.2.1 Experimental setup 11
2.2.2 Task 13
2.2.3 Apparatus 13
2.2.4 Calibration and data processing 14
2.2.5 Analysis of recordings from human trials 15
2.3 Summary and discussion 25

3 Robotic Visuomotor Controller Based on the Human Motion Capture Study . 27
3.1 Background research 28
3.1.1 Visually-aided robotic reaching and grasping 28
3.1.2 Robotic obstacle avoidance 29
3.2 Computational approach and system architecture 29
3.2.1 A single DS and GMM/GMR 30
3.2.2 Coupled Dynamical Systems 32
3.2.3 Eye-arm-hand coordination for obstacle avoidance 34
3.2.4 Robot vision system 38
3.3 Results 41
3.3.1 Model learning 41
3.3.2 Model validation for robot control 41
3.4 Summary and discussion 45

4 Improvements of the Robotic Visuomotor Controller Based on the Lessons from Neuroscience 49
4.1 Background research 50
4.1.1 Neuroscientific models of gaze control and visuomotor coordination 50
4.2 Schematic model of the central nervous system for visuomotor control 52
4.2.1 Cortical reference frames for target encoding 53
4.2.2 Cerebellar dynamical control and motor coupling 56
4.3 Computational model for visuomotor control 60
4.3.1 Model premises 60

4.3.2 Modeling reference frame transformations 61
4.3.3 Modeling coupled motor control 64
4.4 Results 65
4.4.1 Improvements with respect to the previous controller 65
4.4.2 Replicating the observations from Rand et al. study 68
4.4.3 Double-step saccade adaptation experiment 70
4.4.4 Predictions of the model: new experiment with primates 71
4.5 Summary and discussion 72

5 Models of Motor-primed Visual Attention for Humanoid Robots 75
5.1 Background research 76
5.1.1 Computational modeling of attention and robotic attention 76
5.1.2 Current shortcomings of attention-based models for robot vision and their biological solutions 78
5.2 Peripersonal space-primed and motor plans-primed attention 82
5.2.1 Mapping of the workspace to the image plane 82
5.2.2 Peripersonal space-primed attention 85
5.2.3 Motor plans-primed attention 89
5.3 Attention-driven visual processing 90
5.3.1 Thresholding and sampling 90
5.3.2 Closing the loop: from covert attentional landscape to overt eye movements and manipulation 92
5.4 Results 93
5.5 Summary and discussion 97

6 Conclusion and Future Work 99
6.1 Thesis contributions 99
6.2 Limitations and Future Work 100
6.2.1 Gaze fixation pattern at the target 100
6.2.2 Safety margin for obstacle avoidance 101
6.2.3 More complex human obstacle avoidance studies 101
6.2.4 Flow of visuomotor coordination 101
6.2.5 More complex visuomotor coupling 102

References 103

1 Introduction

1.1 Motivation

Humans execute visually driven actions, such as preparing a morning cup of coffee in the kitchen. Humans very reliably manipulate with the cup and kitchenware tools without disastrous consequences such as spilling the hot liquid or colliding with sharp obstacles on the way. Tasks like this appear to us as profoundly simple, straightforward and easy to do. However, beneath this easygoing appearance, resides a very powerful and sophisticated neural machinery that directs the orchestra of intermingled and complex neural computations needed to solve various computational functions.

On the other hand, in spite of the last several decades of theoretical and technological breakthroughs in the field, the abilities of autonomous humanoid robots to perform similar tasks are by far modest when compared to the human performance. This is particularly apparent when robots are needed to operate in unstructured, unpredictable and dynamically changing environments. For this reason, biologically inspired mechanisms have a tremendous potential in bridging this gap and for endowing robotic systems with a set of skills that are comparable to those found in humans. The second benefit of implementing biologically inspired visually driven mechanisms in robotic systems, but not less important than the former one, is the possibility to test the plausibility of various neuroscientific hypotheses and theories by implementing them on an artificial physical system (Sandini et al., 2004, 2007; Vernon et al., 2010; Metta et al., 2010).

Human motor control requires complex integration of multiple sensory modalities, such as visual, tactile and proprioceptive information (Prablanc et al., 1979; Jeannerod, 1984; Desmurget et al., 1998a; Purdy et al., 1999; Crawford et al., 2004). A sensorimotor system of an agent placed in the dynamic and unpredictable world must obey the real-time requirements for performing a set of tasks: visual scene analysis, sensorimotor reference frame transformations, motion replanning, calculating and issuing motor commands and synchronizing the movements of different limbs, while constantly monitoring execution of all movement stages. In the first stage of the process, the visual scene is projected on the retina and the targets are represented in retinal coordinates. In the last stage, in order to accomplish actions defined in space, body movements require activation of muscles that revolve proximal and distal limb segments around joints, constituting a kinematically redundant, high-dimensional and nonlinear system. Thus, the need for synergistic and coordinated actions of different effectors such as the eyes, head, arm, hand and torso, and, if the motor goal is beyond the peripersonal space, the whole-body action, imposes demands for: (a) a series of reference frame transformations and (b) adjutory synchronization commands to appropriately sequence and coordinate the motion of different effectors. The neural algorithms that provide us with these abilities are amazingly powerful and yet beautifully elegant: they are highly optimized for computational efficiency, modular and capable of performing computations simultaneously in both parallel and sequential fashion.

While humans and other primates have mastered gazing, reaching and grasping task to a great extent, modern humanoid robots are far from being able to autonomously and reliably accomplish the tasks we take for granted and do with ease. In robots, the visual and motor system remain largely independent modules. In this thesis, we exploit three paradigms (and the interplay between them) from the human visuomotor system that can endow robots with a higher degree of dexterity and autonomy: *active vision* that is *coupled and synchronized* with the motor system constituting a coherent, but still modular, mechanism, which can *rapidly react to perturbations* in the environment. Some computer vision problems that are inherently ill-posed when using passive vision become well-posed when employing an active vision strategy[1] (Gibson, 1950; Bajcsy, 1988; Bajcsy and Campos, 1992). Aloimonos et al. (1988) and Ballard (1991) have shown that an observer engaged in the active vision strategy gains a number of advantages over a passive observer, namely in terms of the cost of visual computation, the stability of algorithms and the uniqueness of solutions when determining shapes, determining structure from motion and computing depth. In active visual systems, visual servo control is computationally easier and more robust to errors in measurements as well (Ballard, 1991). Coupling mechanisms between different control modules play an important role for ensuring a proper coordinated execution of complex tasks, such as visually guided reaching where the torso, head (including the eyes), arm and hand are simultaneously engaged. A proper coordination pattern between modules is especially crucial when performing prehensile tasks in the face of perturbations (Shukla and Billard, 2011). Finally, a real-world environment can be rather highly dynamic and unpredictable. The agent must be able to re-plan and react in a time range of several milliseconds to changes that can happen unexpectedly. Not being able to rapidly and synchronously react to perturbations can cause fatal consequences for both the robot and its environment.

Vision is one of the most important functional modules, if not the most important one, to provide support to motor control in both artificial and biological systems. The evolutionary motive of vision, according to some authors, stems from the need for improved motor control (Churchland et al., 1994; Wilson, 2002). Yet, vision is one of the most computationally demanding modules. In spite of this fact, humans and non-human primates have the ability to rapidly and graciously perform complicated tasks with a limited amount of computational resources. One of the reasons for their superior performance in visuomotor tasks is an efficient distribution of the visual resources to select only relevant information for reaching and grasping among the plethora of visual information. Humans are able to efficiently and routinely manage this challenging task of selective information

[1] Active vision systems employ gaze control mechanisms to actively position the camera coordinate system in order to manipulate the visual constraints.

processing, in a seemingly effortless manner, by means of highly customized attentional mechanisms. When dynamically changing environmental conditions demand rapid motor reactions, there is no time to compute the full visual model of the world (Ballard, 1991; Wilson, 2002). The humans and non-human primates use attention to select important visual information, and cheaply compute only a relevant subset of them on the fly. Furthermore, visual attention (covert and overt) is tightly coupled with the motor system. Numerous findings from visual neuroscience and psychology provide evidence that visual attention is bound and actively tailored with respect to spatio-temporal requirements of manipulation tasks (Hayhoe et al., 2003; Baldauf et al., 2006; Baldauf and Deubel, 2008; Geisler, 2008; Baldauf and Deubel, 2009).

In most of the humanoid robots, the computational demands for processing stereo images represent very often a bottleneck for real-time manipulation, where replanning and computation of visuomotor actions are time-locked within a time range of only a few milliseconds. Most of the approaches in robot vision are based on the standard, "off the shelf", image processing techniques, ignoring most, if not all, the information regarding the current motor state and planned motor actions. This implies that the visual system and the arm-hand system are usually considered as two largely independent modules that communicate only in the direction from vision to manipulation, which implies that during visual processing the valuable information from the manipulation system is mostly ignored. This decoupling of visual processing from the motor information manifests itself in an inefficient, hence slow, visual processing.

In this thesis, we focus on the problem of visuomotor coordination in reaching and grasping tasks. We first study humans in visuomotor tasks and complement our behavioral experiment with the investigation of the neuroscientific literature in monkeys, to extract the fundamental principles of visuomotor coordination. Based on these principles we target to solve three complex problems in humanoid robotics:

- Computation of a sequence of transformations from the retinotopic encoding to reference frames suitable to generate eye-head and arm movements.
- Generating movements of the eyes, arm and hand and appropriately coordinating the movements of these effectors.
- Tailoring vision with respect to spatio-temporal requirements of the motor system.

Hand-tuning the parameters of block elements of a visuomotor coordination scheme of a humanoid robot with $\sim 40+$ actuated joints from the hips up and $\sim 2 \times 320 \times 240$ pixels in the stereo cameras would be a daunting, if rather impossible, task. For these reasons, in this thesis, we embrace babbling-like exploration and programming by demonstration learning paradigms. The recent advancement in machine learning, namely in the domain of non-linear regression, provides a very convenient means to deal with this problem by learning from a set of empirically obtained data. However, even with the powerful machine learning tools in our hands, the combinatorial explosion of straightforwardly learning visuomotor parameters would make the task of learning visuomotor coordination very difficult. To circumvent this problem, we introduce priors in our modeling by studying the human and monkey visuomotor principles. By doing this, we constrain our modeling and make learning feasible, and the size of the resulting set of parameters reasonably small to be able to efficiently run inference computations in real-time.

1.2 Thesis outline

The major contributions presented in this thesis have been published in peer-reviewed conferences and journals. Here we briefly present the topics presented in each chapter together their associated contributions.

In Chapter "Human Motion Study of Reaching and Grasping with Obstacle Avoidance", we investigate the role of obstacle avoidance in visually guided reaching and grasping movements. We report on a human study in which subjects performed prehensile movements with obstacle avoidance where the position of the obstacle was systematically varied across trials. These experiments suggest that reaching with obstacle avoidance is organized in a sequential manner, where the obstacle acts as an intermediary target. Furthermore, we demonstrate that the notion of workspace traveled by the hand is embedded explicitly in a forward planning scheme, which is actively involved in detecting obstacles on the way when performing reaching. We find that the gaze proactively coordinates the pattern of the arm-hand motion during obstacle avoidance. This study also provides a quantitative assessment of the coupling between the gaze-arm-hand motion. We show that the coupling follows regular phase dependencies, and that it is unaltered during obstacle avoidance. The human study from this chapter provides quantitative information about the eye-arm-hand organization to support the development of the robotic model of visuomotor coordination presented in the subsequent chapter.

Chapter "Robotic Visuomotor Controller Based on the Human Motion Capture Study" describes a robotic visuomotor controller developed based on the observations of our human study and by using the gaze-arm-hand data acquired in the human trials. Our controller extends the Coupled Dynamical Systems (CDS) framework and provides fast and synchronous control of the eyes, the arm and the hand within a single and compact framework, mimicking similar control system found in humans. The generalization abilities of the CDS framework ensure the coordinated behavior of the visuomotor controller, even when the motion is abruptly perturbed outside the region of the provided human demonstrations. Similar to classical visual servoing, it performs a closed-loop control, hence it ensures that the target can be reached under perturbations. We validate our model for visuomotor control of a humanoid robot. The observed forward planning mechanism for obstacle detection in our human study has motivated the development of a similar scheme for the robotic controller. The observation that the visuomotor system treats the obstacle as an intermediary target tremendously reduces the computational and architectural complexity of our visuomotor model for obstacle avoidance.

In "Improvements of the Robotic Visuomotor Controller Based on the Lessons from Neuroscience", we present improvements to the visuomotor model presented in Chapter 3. We derive this new model by investigating the main computational principles reported in the neuroscientific literature regarding the reference frames used for programming visuomotor movements, the cerebellar contribution to multivariate synchronization of motor control and the functional organization of these systems. We stress the view that the cerebellum uses the cortical reference frame representation, and, based on this representation, performs closed-loop programming of multi-joint, compound movements and movement synchronization between different effectors (i.e. the eye-head system, arm and hand). We then attempt to unify these considerations in our computational model. In order to complement our theoretical and modeling work, we validate the model's effectiveness in experiments with the humanoid robot iCub. Because this model is derived by carefully taking into account the neuroscientific computational principles, we are able to provide some complementary theoretical predictions to be tested in future work.

Chapter "Models of Motor-primed Visual Attention for Humanoid Robots" aims to complement the *vision-to-motor* direction of coordination, presented in the previous chapters, by modeling the flow of influence in the other direction, from the motor system to the vision system. This chapter presents a novel, bio-inspired, approach to an efficient allocation of visual resources for humanoid robots in the form of a motor-primed visual attentional landscape. The attentional landscape is a more general, dynamic and a more complex concept of an arrangement of spatial attention than the popular "attentional spotlight" or "zoom-lens" models of attention. Motor-priming of attention is a mechanism for prioritizing visual processing to motor-relevant parts of the visual field, in contrast to other, motor-irrelevant, parts. In particular, we present two techniques for constructing a visual "attentional landscape". The first, more general, technique, is to devote visual attention to the reachable space of a robot (peripersonal space-primed attention). The second, more specialized, technique is to allocate visual attention with respect to motor plans of the robot (motor plans-primed attention). Hence, in our model, visual attention is not exclusively defined in terms of visual saliency in color, texture or intensity cues, it is rather modulated by motor information. This computational model is inspired by recent findings in visual neuroscience and psychology. In addition to two approaches to constructing the attentional landscape, we present two methods for using the attentional landscape for driving visual processing. We show that motor-priming of visual attention can be used to very efficiently distribute limited computational resources devoted to the visual processing. The proposed model is validated in a series of experiments conducted with the iCub robot, both using the simulator and the real robot.

In "Conclusion and Future Work", we provide a discussion and summarize the thesis, its interdisciplinary contributions and the hypotheses regarding the human visuomotor system. We then propose several future directions for improvements of the presented work and some possible directions that could be natural extensions of the work.

2 Human Motion Study of Reaching and Grasping with Obstacle Avoidance

Manipulation and grasping skills are complex and rely on the conjunction of multiple sensing modalities, including vision, tactile and proprioceptive information (Prablanc et al., 1979; Jeannerod, 1984; Purdy et al., 1999). Vision provides important information in the early stages of motion planning (Prablanc et al., 1979; Abrams et al., 1990; Spijkers and Lochner, 1994; Rossetti et al., 1994). It is also used to perform closed-loop control to drive the hand in space unobstructed visually (Abrams et al., 1990; Paulignan et al., 1991b), while tactile information becomes crucial in the last stage of prehension and to compensate when vision cannot be used[1] (Jeannerod, 1984; Purdy et al., 1999). Vision is particularly useful to plan the motion so as to avoid obstacles without touching them (Johansson et al., 2001). It also enables to react rapidly in the face of a sudden perturbation, such as an obstacle entering the workspace (Aivar et al., 2008). There is a tight coupling between the visual and motor system when driving the prehensile motion (Prablanc et al., 1979; Land et al., 1999; Johansson et al., 2001). While this coupling has been documented at length in the literature on free space movements (Johansson et al., 2001; Hayhoe et al., 2003; Bowman et al., 2009), little is known about how this coupling is exploited to enable fast and reliable obstacle avoidance, and in particular when the obstacle appears after the onset of the motion. Such fast and on-line control of the hand motion in response to visual detection of an obstacle is crucial for humans, but also for robots. Indeed, in spite of impressive advances in robotics over the last decades, robots are still far from matching the human versatility in the control of their motion, even when performing the most simple reach and grasp motion.

In this chapter, we study behavioral principles of the visuomotor coupling between the eye-arm-hand systems, when this coupling is modulated by the presence of an obstacle. Identifying and modeling the mechanisms at the basis of human visuomotor control in the presence of the obstacle is relevant for understanding how the human visuomotor system is organized. It could provide a promising research direction to improve the design of similar controllers in robots, as well. Here, we hypothesize that the visuomotor system preserves a coordinated manner in gaze-arm-hand control in complex natural tasks with head-free movements, such as visually-aided obstacle avoidance. We hypothesize that the central nervous system (CNS) favors task segmenting when performing obstacle avoidance instead of holistic programming. The rationale for this is that the first strategy offers a simplified computational approach compared to the second one. Furthermore,

[1] Humans can perform prehensile actions without visual feedback, by relying on tactile and acoustic senses.

in order to identify obstacles, we expect to observe some sort of anticipatory visuomotor planning. If this forward planning exists, it should be observed in proactive gaze fixations of the object when it obstructs intended arm movements. In other words, if the object identified as an obstacle is the intermediary target for the visuomotor system, it is expected that it will be visually fixated during reaching. The opposite should be true, if the object is not identified as the obstacle, we do not expect that the gaze would fixate it, due to the demand to bind visual resources only to the parts of the visual field that are relevant to the requirements of the motor system. Under the assumption that visuomotor coordination remains preserved in obstacle avoidance tasks, motor segmenting and forward planning should be also observed in the arm kinematic parameters and in the pattern of the correlations between the gaze and arm parameters. If, indeed, the aforementioned task segmenting strategy exists, the profile of coordination of the gaze and arm with respect to the obstacle, under the hypothesis that the obstacle acts as an intermediary target, should be similar to the pattern observed with respect to the target. The gaze-arm correlations when approaching to avoid the obstacle (the first stage of the movement) should be very similar to the correlations in the second segment of the movement (when the obstacle is passed by and the eye-arm system aims for the target).

The human study provides quantifiable information about the eye-arm-hand coupling to support the design of the robotic model's parameters, presented in the subsequent chapter.

We next provide a short review on existing works, focusing on the role of visual information in guiding manipulation and visuomotor coordination mechanisms in humans.

2.1 Background research

2.1.1 The general role of visual information in guiding reaching and grasping

Vision provides a plethora of by far the most valuable and most reliable information about the state of the environment on which the planning and motor systems depend heavily. The object's extrinsic properties (spatial location and orientation) are used to control the reach component, whereas the object's intrinsic properties (shape, size, weight, centroid and mass distribution) are used in programming the grasp component (Jeannerod, 1984). The role of vision in manipulation is best shown in behavioral experiments where visual feedback is deprived by modulating experimental conditions.

Several studies have shown that manipulation without any visual feedback in highly structured, static scenarios can almost match the performance of the full-vision manipulation (Castiello et al., 1983; Purdy et al., 1999). After a number of practice trials, manipulation of subjects who did not have any visual feedback only slightly differed from full-vision manipulation in terms of the kinematic measures of both the reach and grasp components. However, if manipulation without visual feedback is performed in an unstructured environment, without previous kinesthetic assistance from a teacher or extensive trial-and-error learning, the performance (e.g. overall success rate, accuracy of reaching, speed of movement, etc.) drastically degrades compared to trials where vision was not deprived (Purdy et al., 1999).

Vision is used to guide every stage of prehensile movements, from pre-planning, initial reach, high-speed mid-section of the movement, to the deceleration and grasping phases. Prablanc et al. (1979) and Rossetti et al. (1994) showed that seeing the limb before the onset of movements improves the reaching accuracy. In addition to this, Pelisson et al. (1986) found that the initial information about the target affects the final reaching accuracy. Similarly, the sight of the current position of the limb and the goal in the later stage of the movements improves the end point accuracy (Prablanc et al., 1979; Pelisson et al., 1986). In studies of manipulation where no visual feedback on the moving limb (Gentilucci et al., 1994; Berthier et al., 1996) and the target (Jakobson and Goodale, 1991) is available, a dramatic increase in the overall movement time and the grip aperture was observed. Finally, visual information assists fine control of the arm and hand in the closing phase of grasping (Paillard, 1982). The gaze is driven to the grasping points on the target object during a prehensile task, for the purpose of planning reliable placement of the fingers (Brouwer et al., 2009). These studies suggest that vision is used for on-line control of both the reaching and grasping components of a prehensile movement.

A number of studies have shown that both peripheral and foveal vision contribute to reaching and grasping. Sivak and MacKenzie (1990) found that when central vision was blocked, it affected both the transport and grasp components (longer movement times, lower peak accelerations and peak velocities, larger maximum grip apertures and longer time after the maximum grip aperture). When peripheral vision was not available, however, they observed that it affected the transport component only, and the grasp component remains unaltered. In their follow-up study, González-Alvarez et al. (2007) found that peripheral and foveal visual cues jointly contribute to both reaching and grasping.

Further evidence that vision is used for on-line control of movements comes from perturbation studies. The studies of Paulignan et al. (Paulignan et al., 1991b,a) have shown that subjects were able to instantly modulate, by relying on visual feedback, the arm and hand movements with respect to on-line perturbations of the position and shape of the target object, with only minimal increase in the response time (∼100 ms) compared to the motion in the absence of perturbations. Aivar et al. (2008) studied adjustments of the hand movements with respect to abrupt online perturbations of obstacles and/or the target. They found similar latencies to those reported by Paulignan et al. (Paulignan et al., 1991b,a) for the responses to the perturbations of the target position and slightly longer adaptation latencies for the obstacles.

2.1.2 Visuomotor coordination in reaching and grasping

The human visual and motor systems are not independent, they operate in coordination and share control signals adapting to mutual demands, even when doing simple and well-practiced routines (Land et al., 1999; Hayhoe et al., 2003). A body of literature documented how the gaze precedes movements. The gaze shows an anticipatory strategy leading a whole body movement during navigation (Grasso et al., 1998; Hicheur and Berthoz, 2005; Rothkopf and Ballard, 2009). The gaze precedes the arm and the hand movement in manipulation tasks with a tool in the hand (Johansson et al., 2001). Similar pattern, the gaze leading the arm, is observed in a task where subjects contacted multiple target objects arranged in a sequence (Bowman et al., 2009). Abrams et al. (1990) found that the gaze leads limb movements in rapid tasks as well. Furthermore, it is also observed that the gaze leads the arm and the whole body movements in reach-for-grasp tasks (Land et al., 1999; Hayhoe et al., 2003; Hesse and Deubel, 2011). Physiological studies of reaching and grasping report that the arm transport and the hand preshape components are coordinated by the motor system in reach-for-grasp maneuvers, even in the presence of perturbations (Castiello et al., 1993; Haggard and Wing, 1995). Furthermore, there is a strong evidence that control signals also flow from the hand to the eyes, not only in the opposite direction (Fisk and Goodale, 1985; Neggers and Bekkering, 2000).

While we have emphasized until now the importance of active gaze control to drive the arm-hand motion, it is noteworthy that humans can also grasp an object without fixating it and even perform more complicated tasks such as obstacle avoidance by solely relying on peripheral vision (Prablanc et al., 1979; Abrams et al., 1990; Johansson et al., 2001). In spite of the fact that humans may reach without looking at the target, in natural and unrestricted tasks, the gaze seems to lead the arm-hand movement. This mechanism is likely a safeguard mechanism to ensure accurate reaching in the face of obstacles. Indeed, when saccades to the target and obstacle were prohibited, significantly decreased manipulation accuracy was observed (Abrams et al., 1990; Johansson et al., 2001), and manipulation resulted in frequent collisions with the obstacle (Johansson et al., 2001). These experiments provide further evidence that coupling between active vision and the motor system is an important and fundamental mechanism, synchronously orchestrated between different regions in the central nervous system (CNS).

2.2 Human Motion Study of Reaching and Grasping with Obstacle Avoidance

We start from the hypothesis that the eyes precede the arm motion, so as to guide the planning of the arm transport component. There is ample evidence of such saccadic eye movements toward the target during reaching; see e.g. (Land et al., 1999; Johansson et al., 2001; Hayhoe et al., 2003; Hayhoe and Ballard, 2005), however, few studies have analyzed visuomotor behavior in trials where the position of the obstacle was systematically varied. We assume that the obstacle acts as an intermediary target when performing obstacle avoidance. This movement-segmented strategy sub-

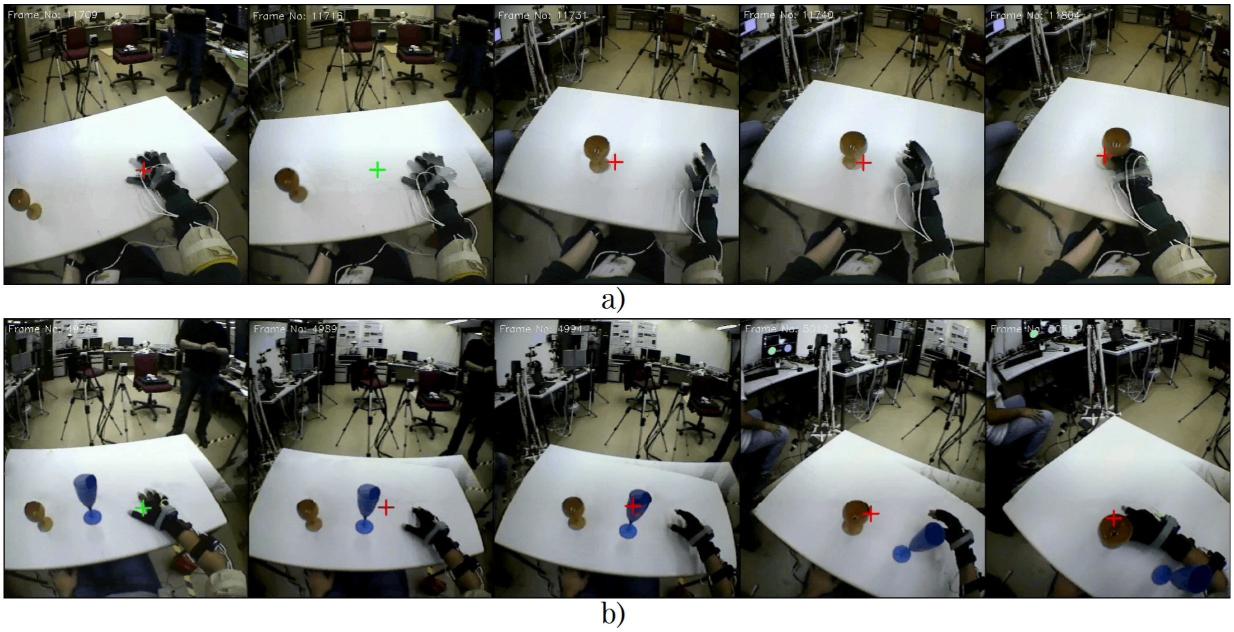


Figure 2.1: Snapshots from the WearCam video from the start of the task (left) until successful grasp completion (right), in (a) no-obstacle and (b) obstacle scenarios. The cross superposed on the video corresponds to the estimated gaze position. The color of the cross indicates whether the gaze is the fixation state (red) or the saccade state (green).

stantially reduces the complexity of motor control compared to the holistic control policy (Alberts et al., 2002; Johansson et al., 2009; Hesse and Deubel, 2010). Furthermore, we hypothesize that there exists a visuomotor forward control scheme in which the presence of the obstacle is used to modulate the path of the arm. This modulation depends on the distance of the original path to the target. We also assume that the obstacle avoidance maneuver consists in passing the obstacle on the side of the obstacle where the collision would have occurred. This choice participates in a minimum effort strategy with only a small modulation of the intended path. We report our analysis of the visuomotor obstacle avoidance scheme in the following sub-sections. Figure 2.1 shows snapshots taken from the WearCam video illustrating the mechanism of the gaze leading the arm motion and fixating the obstacle on the path when reaching the target.

The first part of this section describes the experimental procedure followed during our human motion study. In the second part, we analyze the results of this study and state our findings of visuomotor coordination that constitute a basis for developing our computational model.

2.2.1 Experimental setup

Eight unpaid subjects from the university staff participated in this experiment (5 males and 3 females; mean age 27.1 years and std. 3 years). Subjects were right-handed and did not have any neurological or ophthalmological abnormalities. Subjects were unaware of the purpose of the experiment.

Subjects sat in a height-adjustable chair facing a rectangular table with task-relevant objects placed on the surface of the table (Figure 2.2). Subjects sat in front of a table such that the sagittal plane "cut" the width of the table at approximately the midline, and the distance from the frontal part of the trunk to the edge of the table was ∼10 cm. The initial positions of the right hand, the target object and the obstacle object were predetermined and they were laid along a line parallel to the coronal plane of the body, 18 cm displaced from the edge of the table on the subject's side. The distance measured in the table plane from the initial hand position (hand centroid) to the obstacle was 25 cm, and from the obstacle to the target it was 20 cm (i.e. 45 cm from the starting hand position to the target). Starting positions were indicated by markers on the table. The two objects used for manipulation were IKEA glasses, color tinted to enable automatic color-based segmentation on video recordings. The wine glass (max. diameter 7.5 cm, height 13 cm) was the object to be grasped (target) and the champagne glass (max. diameter 5 cm, height 21 cm) was the object to be avoided (obstacle).

2.2.2 Task

Grasping during all trials was conducted with the right hand. The left hand remained on the table, to provide support for the trunk to reduce the movements of the trunk in the coronal plane. At the start of grasping, the subjects were instructed to look at the colored patch mounted on the data glove. A sound signal indicated the start of execution of grasping, instructing the subjects that they were free to unlock gaze from the colored patch, mounted on the data glove, and start a trial. Once the grasping motion was completed, the subject was instructed to go to the starting position.

Each subject performed 8 trials of reaching and grasping the target (wine glass). In all the trials, the obstacle (champagne glass) was present. The location of the champagne glass was changed at each trial. Starting from 6 cm from the edge of the table on the subject's side, we progressively displaced the champagne glass at each trial in increments of 4 cm along the midline of the desk (parallel to the sagittal plane of the subject's body in resting position). An alternative to this approach is to place the obstacle in a randomly indexed position for every trial. By incrementally displacing the obstacle in each trial, we implicitly force subjects to change their previous obstacle avoidance strategy, whereas with random displacements, the hand path which assured successful obstacle avoidance in the previous trial (e.g. obstacle in position 4) could be reused for a new trial (e.g. obstacle in position 2), without much adaptation.

For all trials, subjects were instructed to perform manipulation in a natural manner, without any additional instructions that could affect their visuomotor behavior. The subjects had one trial of practice before recording to ensure that they had understood the instructions. Subjects were unaware of the purpose of the experiment. Figure 2.2 illustrates our setup for this experiment.

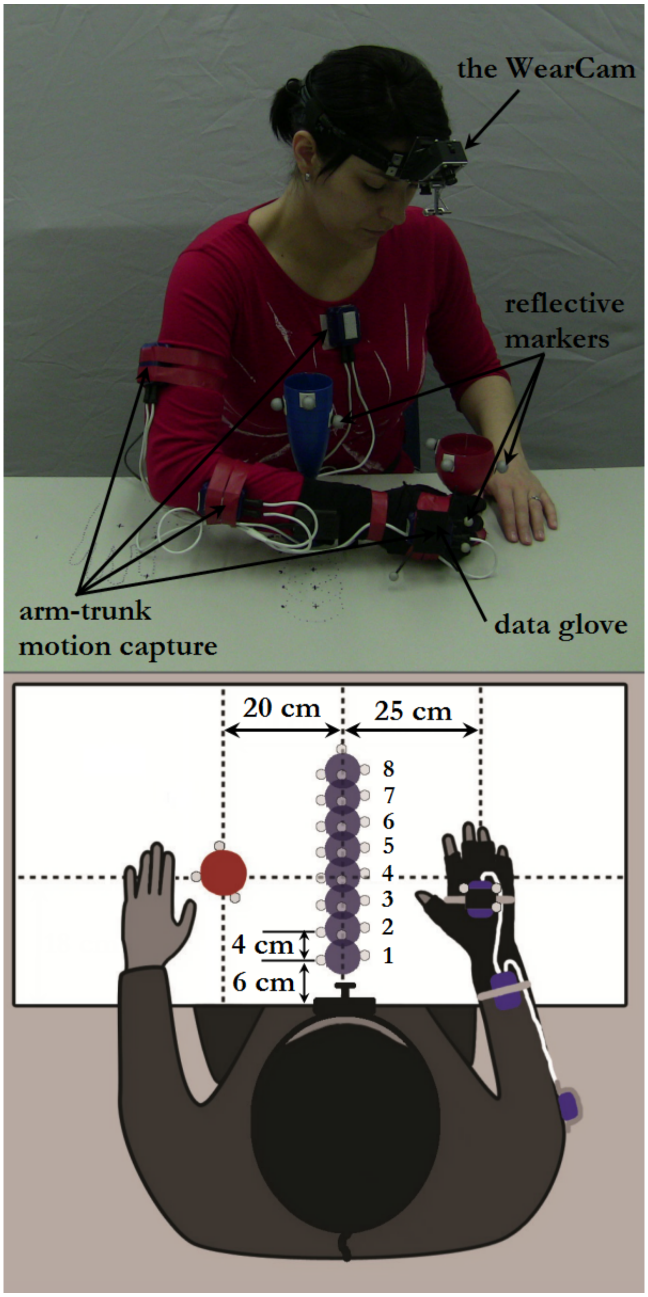


Figure 2.2: Experimental setup to record eye-arm-hand coordination from human demonstrations in grasping tasks where the obstacle (dark blue disk) is progressively displaced in each trial. Obstacle positions (superposed as transparent dark blue disks) are numbered from obs1 to obs8, numbered with respect to the increasing distance from the subject. obs1 is the starting position of the obstacle, 6 cm from the edge of the table. We progressively displaced the champagne glass for each trial in increments of 4 cm along the midline of the desk. obs8 is the farthest position of the obstacle (34 cm from the edge). In this trial, the human subject is grasping the target object (wine glass) avoiding the obstacle (champagne glass).

2.2.3 Apparatus

A head-mounted eye-tracker designed in our laboratory (LASA EPFL), the WearCam system (Noris et al., 2010), was used for gaze tracking and for recording the scene as viewed from the subject's standpoint. The system uses two CCD cameras to record a wide field of view (96°×96°). It uses Support Vector Regression to estimate the gaze direction from the appearance of the eyes. The system has an accuracy of 1.59°. The video and gaze position from the WearCam were recorded in 384×576 MJPEG format at 25 Hz. The WearCam video from our experiment can be seen in Figure 2.1. The XSens^{TM} inertial motion capture system was used for recording the trunk motion and arm motion. The sensors were mounted on the trunk, the upper arm, the forearm and the hand. The system provided information about three joints of the trunk motion (roll, pitch and jaw), three joints that model the shoulder (flexion-extension, abduction-adduction and circumduction), two joints in the elbow (flexion-extension and pronation-supination) and two wrist angles (abduction-adduction and flexion-extension). The 5DT^{TM} data glove, with flexure-sensors technology, was used for recording the finger joint angle motion. The data from the XSens^{TM} IMU motion capture sensors and the 5DT^{TM} data glove were recorded at 25 Hz.

The OptiTrack^{TM} multi-camera system was used for tracking the 3D positions of the hand and the objects in the scene. The speed of data recording from the multi-camera system was 150 Hz, and the accuracy was ∼2 mm.

2.2.4 Calibration and data processing

The WearCam system was calibrated at the beginning and the end of the task for each subject by using the procedure explained in Noris et al. (2010). The state of the WearCam was verified after each trial by checking its relative position with respect to the head and observing the video that was streamed. We checked the state of the multi-camera system by observing the performance of real-time detection of the objects in the workspace, and we recalibrated it when the accuracy was not satisfactory. The data glove and the motion capture sensors were calibrated after each trial by requesting the subject to adopt an upright straight posture of the torso and to perform a sequence of opening and closing fingers. The state of the data glove and the motion capture sensors was verified by using an in-house GUI tool that shows the body posture of the subject by using real-time readings from the sensors.

All recorded signals were filtered with a preprogrammed peak-removal technique that consisted of removing outliers from sensor misreadings and replacing them with linearly interpolated values between two closest valid readings. All signals were re-sampled at 25 Hz. Synchronization and parsing of signals were performed by using time-stamps for recorded signals and verified by observing recorded videos on a frame-by-frame basis. The signals were smoothed with a moving average filter. Piecewise spline fitting was done, which did additional smoothing as well. Finally, we visually assessed comparative plots of both raw signals, and synchronized and smoothed signals in order to make sure that filtering and smoothing did not distort general signal profiles.

We detected gaze fixations as all instances where the gaze remained steady for at least 80 ms with the gaze motion not exceeding 1° of the visual field (Inhoff and Radach, 1998; Jacob and Karn, 2003; Dalton et al., 2005). We say that a person is looking at either of the two objects (target or obstacle) if a gaze fixation is contained within the object blob, or it is within a 5-pixel radius around the object blob. This 5-pixel radius accounts for imprecision in the blob segmentation, and in the estimation of the gaze position. It also accounts for the fact that the "functional fovea" forms a 3-degree circular region around the center of the gaze, which means that the visual system can obtain high-quality visual information fixating very close to the edges of interesting objects (Rothkopf and Ballard, 2009). We empirically obtained this specific value of a 5-pixel tolerance by computing the average closest distance between the estimated gaze point detected in the fixation state (but outside the segmented blob) and the boundary of the blob. This was done for a number of sub-parts of the reach-for-grasp task for which it is well-known that motoric actions impose strong demands for foveal visual information about the object's state. One of the sub-parts of the task, when gaze fixations at the target object are expected with a high probability, is the moment just before the wine glass is grasped, as it is reported from previous studies that the gaze consistently fixates grasping parts before fingers touch the object (Brouwer et al., 2009).

2.2.5 Analysis of recordings from human trials

Visuomotor strategy and visuomotor coupling in obstacle avoidance

Figure 2.3(a) reveals the obstacle avoidance strategy that the subjects employed with respect to the position of the obstacle. It can be seen that the subjects preferred to avoid the obstacle from the anterior side if the obstacle was positioned between the subject's body and the line that is defined from the starting position of the hand to the target object (obs1-4). If the obstacle was positioned in the anterior direction from the line (obs5-8) then the preferred obstacle avoidance strategy was to veer from the posterior side when reaching to grasp the target object. It can be seen that the subjects are very consistent in their obstacle avoidance strategy, except for the obstacle position number 4 (obs4), for which 5 subjects avoided the obstacle from the anterior side, and 3 subjects veered from the posterior side. Post-hoc analysis of the recorded videos from the experiment revealed that 3 subjects who veered for obs4 from the posterior side kept the posture of the torso more upwards than other subjects during manipulation, hence for them veering from the posterior side was a choice that required less effort. The results presented here provide a basis for the computational model of our obstacle avoidance strategy regarding the choice of the preferred obstacle avoidance side, as discussed in the next chapter.

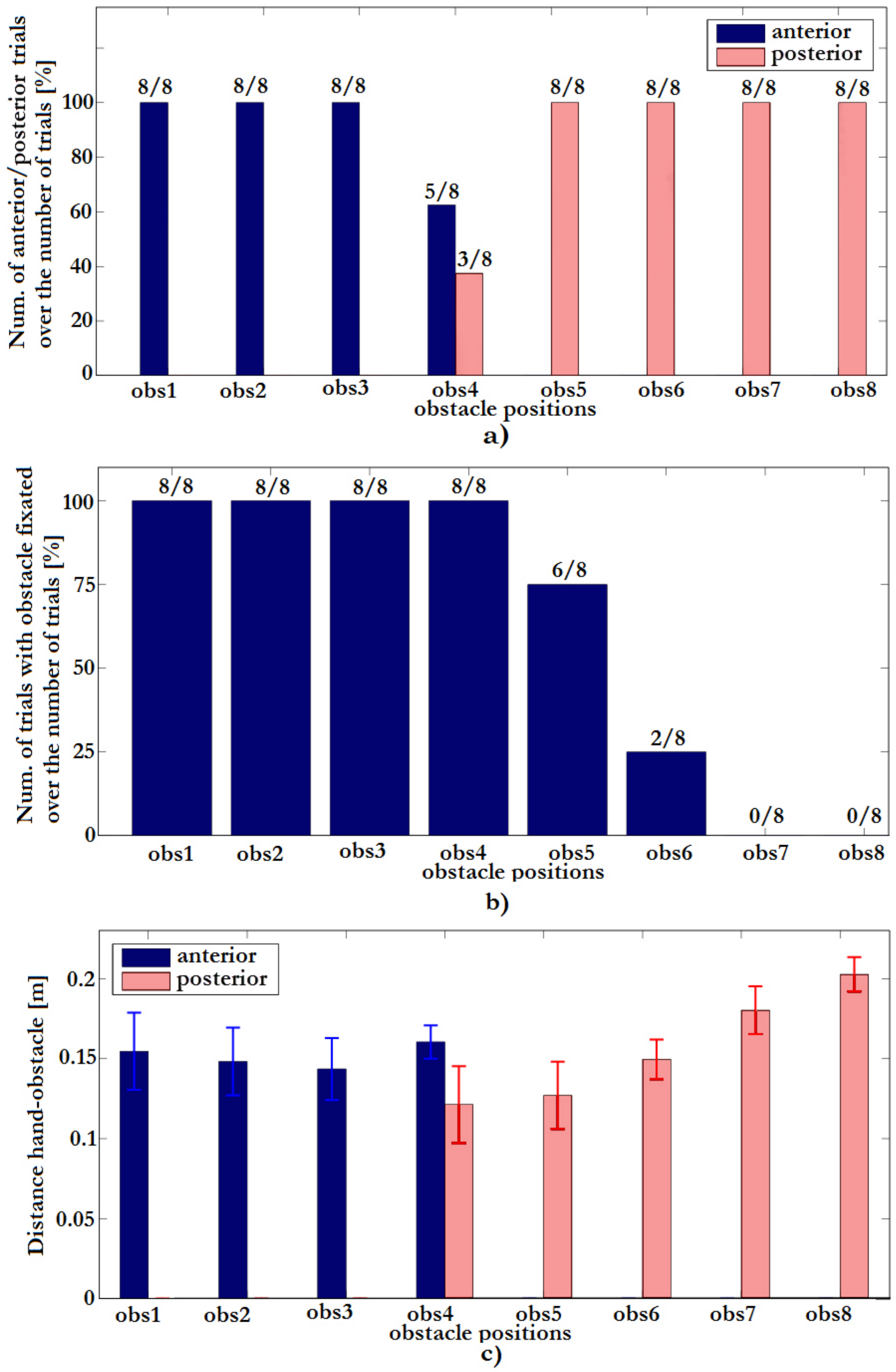


Figure 2.3: Results from the experiment with human subjects where the obstacle was progressively moved along the midline of the table: (a) Influence of the position of the obstacle on a strategy to avoid the obstacle from anterior/posterior side, (b) Influence of the obstacle position on gaze fixations at the obstacle during manipulation, and (c) Safety distances from the hand to the obstacle when avoiding it from anterior/posterior side.

An important part of the forward planning scheme is that an object in the workspace is tagged as an obstacle if it is estimated that the hand will collide with it. As the object identified as an obstacle is the intermediary target for the visuomotor system, it is expected that it will be visually fixated during reaching. Figure 2.3(b) shows the proportion of trials for each obstacle position in which the obstacle object was visually fixated. It can be seen that the champagne glass was always fixated when it was positioned at location 1 through 4 (obs1-4 in the figure). For position obs5, the obstacle was fixated in only 80% of the trials. The amount of fixation rapidly drops to 20% for position obs6, and to zero for positions obs7 and obs8. As expected, once the obstacle is sufficiently far, it is no longer of interest. These results are consistent with Tresilian (1998), who argued that objects treated as obstacles by the motor system are very likely to be visually fixated during manipulation. Thus, our results indicate that the most likely explanation of visual ignorance of the champagne glass when it is placed at obs6-8 is that the visuomotor planning scheme did not identify it as an obstacle.[2]

Based on the study by Dean and Brüwer (1994) and the results of our human experiment where the safety distance between the hand and obstacle was kept (Lukic et al., 2012), we hypothesized that the control system would keep the same safety margin of $\sim 0.14 \pm 0.01\,\mathrm{m}$ across all trials where the champagne glass was considered as an obstructing object (namely for position 1 through 6). In the other position, this safety margin would not be preserved as the obstacle would then be ignored.

In Figure 2.3(c), we plotted the minimum distance (the mean and the standard deviation) between the hand and the champagne glass for all positions of the champagne glass. It can be seen that the distance is quite consistent for obs1 to obs6, and starts increasing for obs7 and obs8. These results also indicate that an obstacle object positioned such that it does not obstruct the original prehensile motion is not identified as an obstacle, and it is not treated as the intermediary target.

A two-way ANOVA[3](factors: subjects and a binary variable that represents whether the obstacle was fixated/not fixated in a trial) on the distance hand-obstacle reveals a significant effect of the *obstacle fixations* factor ($F(1, 63) = 78.3$, $p < 0.001$), and no effect of the *subject* factor ($F(7, 63) = 0.47$) and no factor interaction ($F(7, 63) = 0.35$). These results reflect the fact that the distance between the hand and the obstacle is significantly different when the subjects visually fixate the obstacle, compared to the case without gaze fixations at the obstacle object in the trial. We interpret these results as a confirmation of the influence of forward planning on visuomotor coordination. When forward planning estimates that the object obstructs intended movement, the motor system treats the obstacle as an intermediary target. The gaze fixates the obstacle, and the hand keeps a consistent safety distance from the object. If the object is placed in a position where it does not obstruct movements (obs6-8), it is not "tagged" as an obstacle. The visuomotor system ignores

[2] At the end of all trials, we asked 2 subjects to try to reach the target when the champagne glass (obstacle) was present, but without modification of the path (as in the no-obstacle setup). Unsurprisingly, the arm/hand collided with the champagne glass always when it was positioned at obs2, obs3, obs4, in 6 out of 8 trials the hand collided for obs1 and obs5. The hand never collided when the obstacle was in positions obs6, obs7 and obs8.

[3] ANOVA (analysis of variance) is a statistical method which compares the variances around two or more means, to determine whether significant differences exist between distinct conditions of the experiment. See Montgomery and Runger (2010) for more.

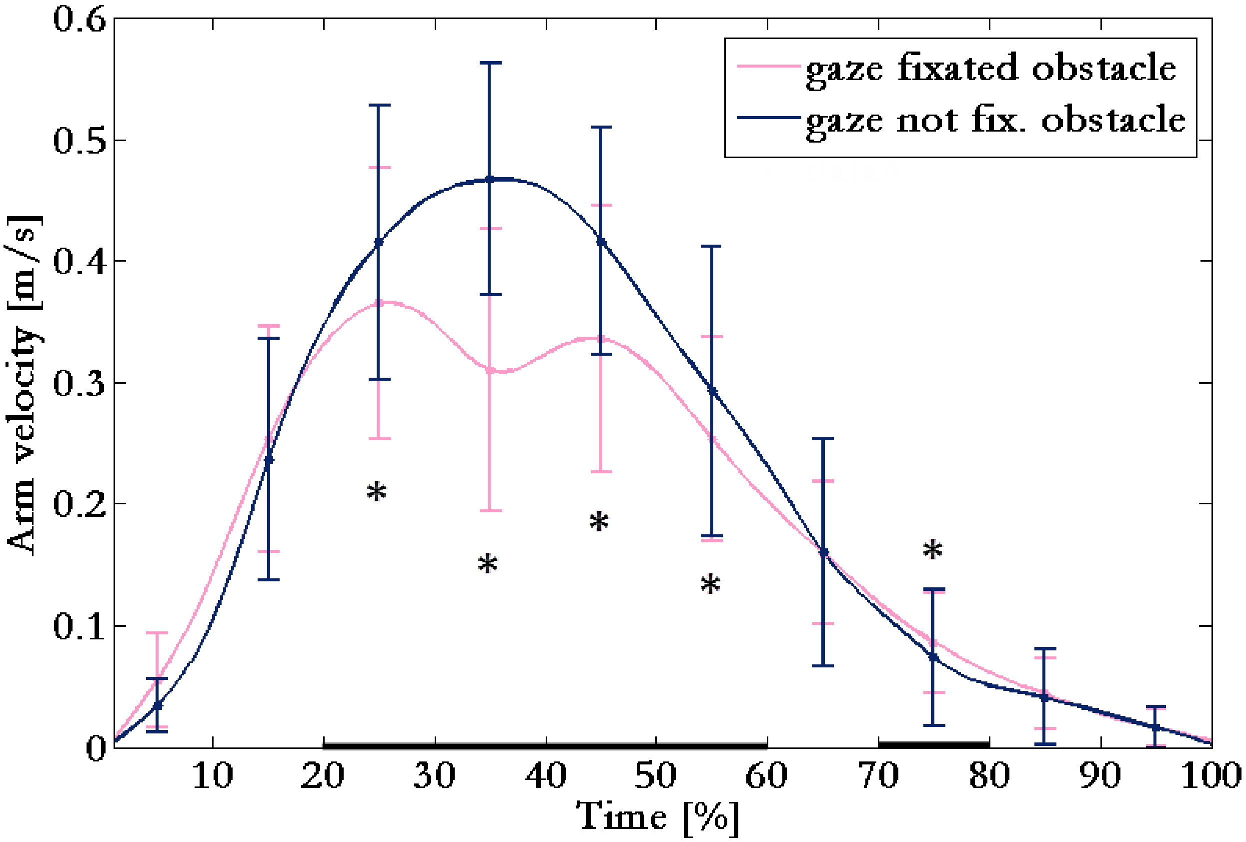


Figure 2.4: Arm velocity profiles, time normalized and averaged over all subjects for the two conditions (gaze fixated the obstacle or not). The stars represent the time bins for which a post-hoc t-test shows a significant difference between the fixation conditions ($p < 0.05$).

objects that are irrelevant to manipulation: they are not visually salient for the gaze (Land, 1999; Hayhoe et al., 2003; Rothkopf et al., 2007; Rothkopf and Ballard, 2009), and the hand is controlled without keeping some safety distance with respect to them.

We show that in the trials, where the location of the obstacle is varied, gaze fixations at the obstacle indicate that the arm keeps the safety distance from the obstacle. To further analyze the coupling between the gaze and the arm when performing obstacle avoidance, we investigated the influence of the gaze on the velocity profile of the arm. Alberts et al. (2002) and Hesse and Deubel (2010) showed that the velocity profile usually reaches a local minimum when the arm passes by the obstacle. In our experiment, the obstacle seems to influence the motion solely in trials when the gaze stops at the obstacle. Hence, we would expect that the arm would slow down at the obstacle only in these trials when the gaze fixates the obstacle. In the absence of the obstacle on the path toward the target, there should be no need to visually guide the arm to avoid it. Figure 2.4 compares the mean arm velocities across the trials in which the gaze fixated the obstacle versus the trials where the gaze did not fixate the obstacle. The observation of such a minimum velocity confirms the hypothesis that the obstacle acts as an intermediary target during movements (Alberts et al., 2002; Hesse and Deubel, 2010). In contrast, and as hypothesized, the velocity profile in obstacle-free trials follows a regular bell-shaped profile.

We apply a two-way ANOVA on the velocity profiles recorded during trials with two factors: a) an *obstacle fixations* factor representing the type of trial, coded as a binary variable, to distinguish between the conditions in which the obstacle was fixated versus not fixated; b) a *time bin* index (the total time of each trial is divided into 10 equal time bins) to determine when, during a trial, an influence of the presence/absence of the obstacle could be observed. We observe a strong effect of the *obstacle fixations* factor ($F(1, 6199) = 109.9$, $p < 0.001$). This confirms that the arm velocity profile is indeed significantly reduced when passing by the obstacle. There is also a significant effect of the *time bin* factor ($F(9, 6199) = 1849.44$, $p < 0.001$), indicating that during the progress of the task arm velocity changed. As expected, the interaction between the factors is significant ($F(9, 6199) = 41.44$, $p < 0.001$) showing that the velocity profiles in trials where the gaze fixates the obstacle changes differently as the task progresses from the trials where the obstacle is not fixated. We run post-hoc t-tests between the fixated and not fixated trials to determine time bins for which the velocity arm profiles differ between the two conditions (Figure 2.4).

The finding that the gaze fixations at the obstacle modulate the arm velocity profiles supports the hypothesis that the gaze-arm coupling exists when humans perform prehension with obstacle avoidance.

Gaze-arm correlations

To see whether the gaze-arm mechanism follows a quasi-constant lag, we analyze trial-by-trial correlations between the gaze and arm positions (computed as the Euclidean distance) with respect to the obstacle (in the first segment of the movement) and correlations between the gaze and arm distances with respect to the target (in the second segment of the movement) as the task progresses. We plot the histogram of the Pearson's correlation coefficient between the gaze and the arm distances computed on a trial-by-trial basis when approaching the obstacle (Figure 2.5(a) and (b)) and the target (Figure 2.5(c) and (d)). We see the prevalence of very high visuomotor correlations for both objects. The distribution of trial-by-trial correlation coefficient between the gaze and arm distances to the obstacle has a sample mean of 0.917, and the 25 %, 50 % (median) and 75 % percentile correspond to 0.876, 0.956 and 0.986, respectively. Similarly, the correlation coefficient between the gaze and arm distances to the target has the sample mean 0.799, and the 25 %, 50 % (median) and 75 % percentile correspond to 0.721, 0.847 and 0.921, respectively. A two-way ANOVA for the correlations to the obstacle (factors: subjects and obstacle position) does not reveal a statistical significance of the *subject* factor ($p = 0.186$) and no effect of the *obstacle position* factor ($p = 0.77$). A two-way ANOVA for the correlations to the target (factors: subjects and obstacle position) shows no statistical significance *subject* ($p = 0.164$) and no effect of the *obstacle position* ($p = 0.934$) as well.

The correlations between the gaze and arm trajectories when reaching to avoid the obstacle are

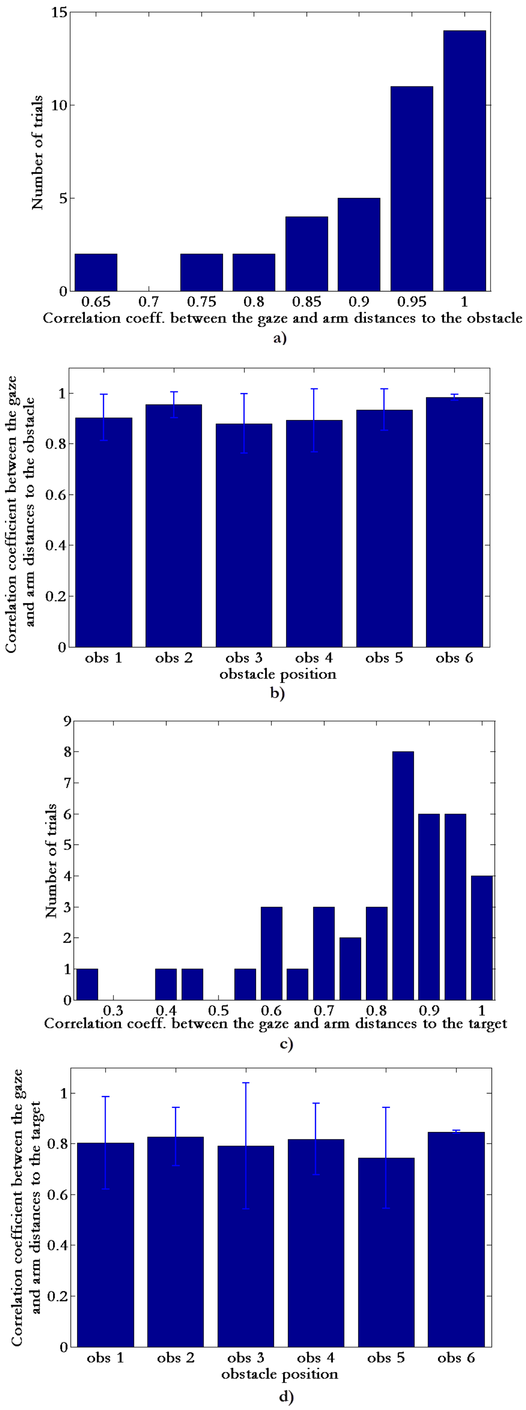


Figure 2.5: The correlation coefficient between the gaze and arm distances with respect the obstacle and the target computed on a trial-by-trial basis when avoiding the obstacle. The motion is segmented into two parts: from the starting position to the obstacle and from the obstacle to the target and we compute the correlations for the corresponding parts of the movements: (a) Histogram of the gaze-arm correlation coefficient when reaching to avoid the obstacle and (b) corresponding values for different fixated obstacle positions, (c) Histogram of the gaze-arm correlation coefficient when reaching the target and corresponding values for different fixated obstacle positions (d).

quasi-constant across trials and subjects, and they are almost the same as those observed for the target. These observations suggest that the eyes and the arm might be driven to both the obstacle and the target by the same mechanism of spatial coordination.

Fixation durations at the obstacle

We now present the results of our analysis of gaze fixation durations at the obstacle. It is well established that the gaze fixation durations, together with the position of the gaze, provide a measure of cognitive processing when performing an ongoing task, being positively correlated with cognitive load required for processing visual information (Rayner, 1998; Deubel et al., 2000; Jacob and Karn, 2003; Hayhoe and Ballard, 2005; Tatler et al., 2011). Gaze fixations in visually guided manipulation allow very specific task-dependent acquisition of visual information (Triesch et al., 2003). This selectivity in information processing is reflected in the duration of fixations (i.e. a variability in fixation duration corresponds to a variability in visual features being selectively acquired from the early visual structures and further processed in the higher cortical structures). Figure 2.6(a) shows the histogram of the fixation durations at the obstacle where the data are pooled from all subjects. The distribution is positively skewed with the sample mean fixation duration at 146.4 ms, where the 25 %, 50 % (median) and 75 % percentile correspond to 80 ms, 120 ms and 160 ms, respectively. The predominance of short fixations observed in our experiment is a common feature of a gaze fixation pattern in natural manipulation tasks (Land, 1999; Hayhoe et al., 2003; Hayhoe and Ballard, 2005), where the average durations of fixations are shorter compared to durations observed in picture viewing and reading (Rayner, 1998; Henderson and Hollingworth, 1999). In spite of the predominance of brief durations of fixations in prehension movements, it has been shown that they do support movement control. Several studies have shown that visual information necessary for movement control can be computed within a single fixation (Ballard et al., 1995; Land et al., 1999). This indicates quite efficient visual processing of some easy-to-compute visual features required for online arm movement control. A two-way ANOVA (factors: subjects and an index variable that represents a position of the obstacle) shows no significant effect of the *subject* factor ($p = 0.321$) and no effect of the *obstacle position* factor ($p = 0.564$, see Figure 2.6(b)) indicating that fixations times are consistent both across subjects and obstacle positions. These results are in agreement with the prior results of Johansson et al. (2001), who observed the predominance of brief fixations at the obstacle. An interesting result comes from one of their obstacle avoidance experiments. When active gaze movements were inhibited during obstacle avoidance, they observed a great variability in the minimum distance kept between the obstacle and the hand. We can speculate that the existence of these brief and quite consistent fixation times reflect the consistency in processing simple visual features of the obstacle in order to guide the arm and hand, because the existence of brief fixation periods does not allow to compute some complex features such as in reading (Rayner, 1998). Considering the predominance of brief fixation times and an increased variability in estimating the position of the obstacle, one of these features computed is

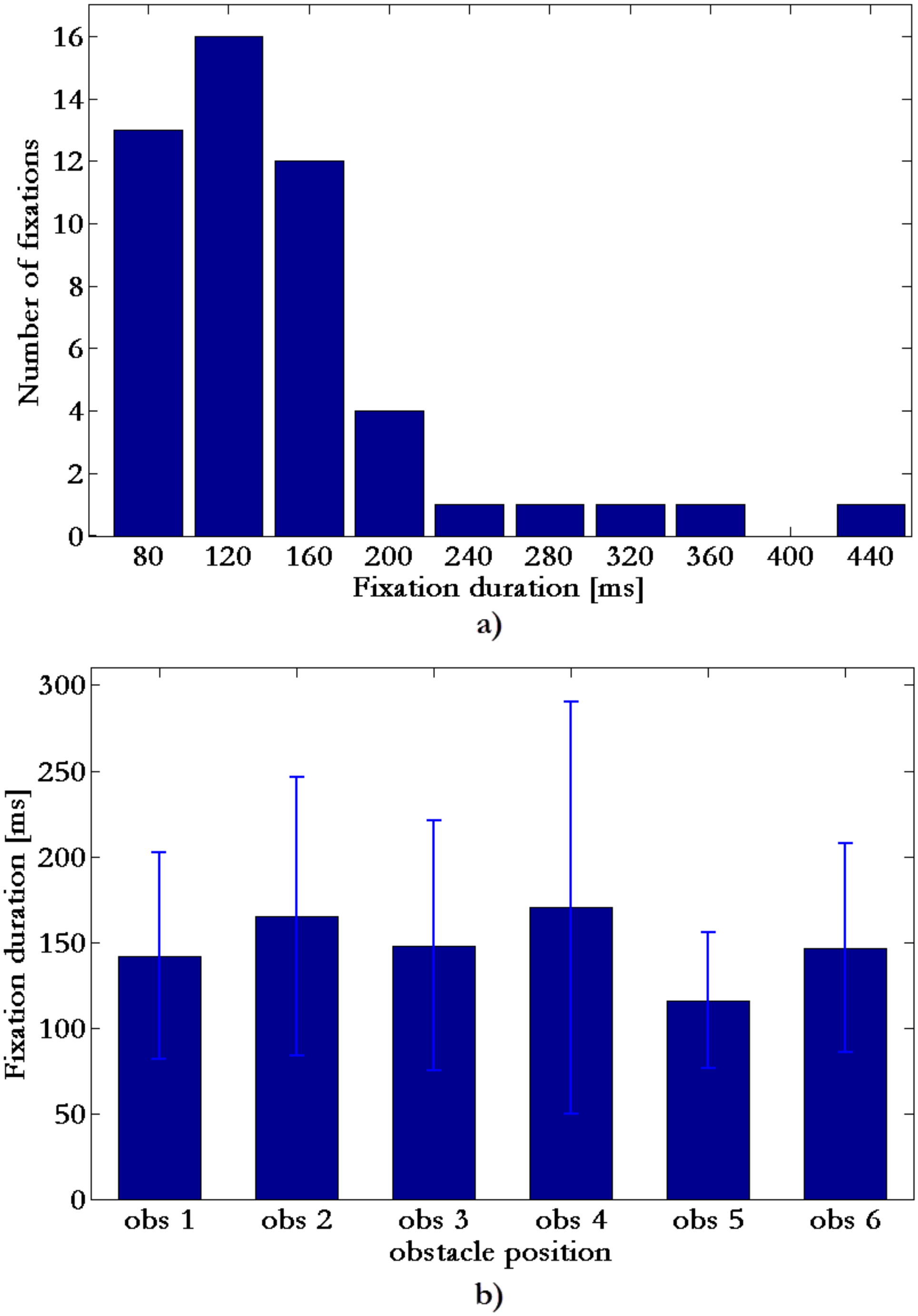


Figure 2.6: Distribution of gaze fixation durations at the obstacle: (a) Histogram of fixation durations pooled from all subjects across all fixated obstacle positions, (b) The mean and the standard deviations of times for different fixated obstacle positions. In this plot we show only fixations times and the standard deviations for positions at which the obstacle is fixated (obs1-6), positions obs7-obs8 are omitted from the figure because subjects never fixated the obstacle when it was placed at these positions.

most likely the spatial position of the obstacle. The spatial location of the obstacle can be rapidly computed from retinal (foveal and parafoveal visual information) and extraretinal information (the relative position of the eyes and the head) available at the moment of fixation by the specialized neural circuitry of the dorsal visual stream (Goodale and Haffenden, 1998; Goodale, 2011), and it is a necessary feature in order to safely guide the arm around the obstacle.

In summary, this analysis of the duration of the gaze fixations provides support to the view that the CNS computes simple features during fixations at the obstacle in order to aid obstacle avoidance. The spatial location of the obstacle is likely one of the main features computed during these gaze fixations on the obstacle.

Gaze and arm exit times from the obstacle

We provide a quantitative assessment of the relation between the gaze exit time and the arm exit time from the obstacle[4]. If some coordination exists between the gaze and the arm when performing obstacle avoidance, these two measures should be correlated. Moreover, the magnitude of the lag between them (i.e. the difference between the exit times of the gaze and arm from the zone of the obstacle) should be kept relatively tight compared to the overall time necessary to complete the movement. When plotting the onset time of the gaze versus the arm onset time from the obstacle we pooled data from all subjects, except for Subject 1[5]. We can see from Figure 2.7(a) that these two variables are linearly correlated (Pearson's correlation coefficient $r = 0.897, p < 0.001$). The slope of the fit indicates that, on average, the gaze exits the obstacle zone slightly earlier than the hand. Figure 2.7(b) shows the histogram of the difference between the gaze exit times and arm exit times, where positive values indicate that the gaze exists the obstacle first. The distribution has the sample mean at 220.78 ms, where the 25 %, 50 % (median) and 75 % percentile correspond to 120 ms, 200 ms and 280 ms, respectively. A two-way ANOVA (factors: subjects and an index variable that represents a position of the obstacle) shows no significant effect of the *subject* factor ($p = 0.18$) and no effect of the *obstacle position* factor ($p = 0.549$), indicating that the difference between gaze and arm exit times were consistent both across subjects and obstacle positions (Figure 2.7 (c)). The predominance of positive differences gives evidence that the gaze leaves the obstacle before the hand leaves it. However, the median time of this lag corresponds to only 8.3 % of the median time

[4] The gaze exit time from the obstacle is defined as the time from the beginning of the trial until the onset of a saccade away from the fixated obstacle. The arm exit time is defined as the time from the beginning of a trial until the moment when the arm reaches the closest distance to the obstacle and starts moving toward the target.

[5] The coordination of the gaze and arm exit times from the obstacle for Subject 1 substantially differed from the rest of the subjects. She has shown significantly different amount of the gaze-arm lag when exiting the zone of the obstacle (mean: 448 ms, std: 210.5 ms) compared to the rest of the subjects (mean: 220.78 ms, std: 135.75 ms) and this difference achieved statistical significance (one-way ANOVA: $F(1, 39) = 10.93, p = 0.002$). A careful analysis of the video from the eye tracker revealed her visuomotor strategy. Interestingly, her eye and arm movements were normal, and the gaze guided the arm in all trials. However, she mostly used the coordination strategy where the gaze first visits the obstacle and the moment when gaze switches toward the target she started to move the arm, i.e. start of her arm movement was significantly postponed. In all the other measures, she did not significantly differ from the rest of the subjects.

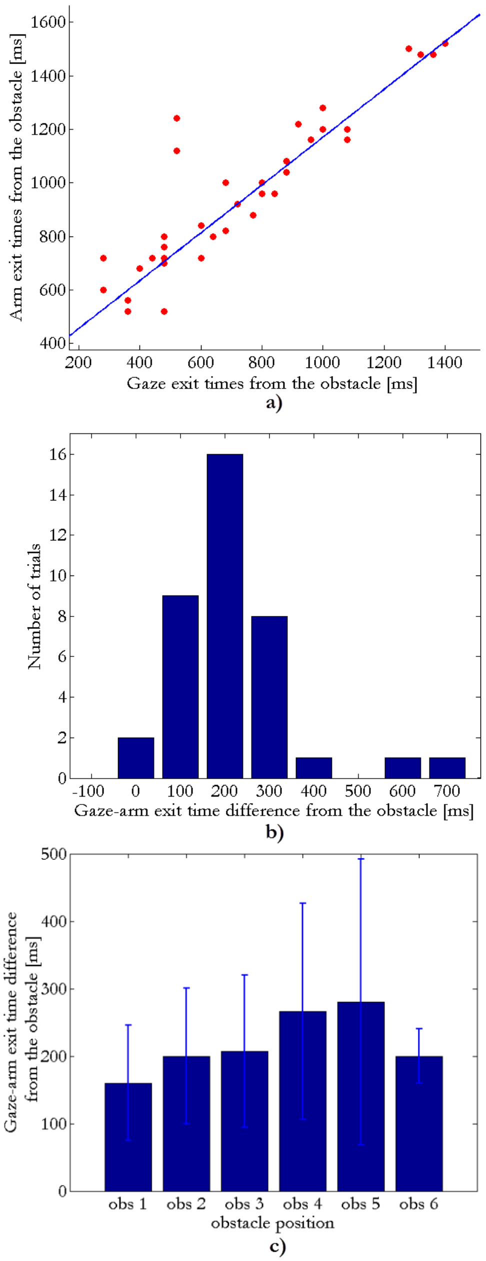


Figure 2.7: Gaze exit times vs arm exit times from the obstacle: (a) Scatter plot of gaze exit times vs arm exit times from the obstacle pooled from Subjects 2-8 across all fixated obstacle positions, (b) Histogram of gaze-arm exit time differences from Subjects 2-8 across all fixated obstacle positions, where positive values mean that the gaze exits the obstacle zone before the arm, (c) The mean and the standard deviations of gaze-arm exit time differences for different fixated obstacle positions. In this plot we show only fixations times and the standard deviations for positions at which the obstacle is fixated (obs1-6), positions obs7-obs8 are omitted from the figure because subjects never fixated the obstacle when it was placed at these positions.

(2.4 s) needed to complete the whole reaching movement with obstacle avoidance. This means that this period of apparent asynchrony after the gaze switched toward the target while the arm is in the obstacle zone takes only a small fraction of the overall movements. For the remaining 91.7 % of the task gaze and arm movements are synchronously driven to the same goal (to the obstacle during the first segment of the movement, and toward the target after the obstacle is passed). Land et al. (1999) observed in their tea-making experiment that the gaze and arm movements are highly coupled during execution of each subtask, but when it comes to a transition toward a new target the gaze switches approximately 0.5 s before the movement of the arm to the previous object is completed. Johansson et al. (2001) found that the difference between the gaze exit times and arm exit times was quite tight when executing sequential tasks, but the gaze starts moving toward the new target slightly before the hand does (∼100-200 ms), as well. The results were similar for a number of different movement sub-targets, including the obstacle[6].

From our results and from the two aforementioned studies, it is evident that the gaze and arm exit times, when completing one movement segment and switching to a new target, are tight compared to the average duration of movements. Nevertheless, it remains to be discussed why this lag is not exactly zero, meaning that the gaze and the arm switch to the next target at exactly the same time. We here provide two alternative explanations.

First, this lag may be due solely to the well-known delays in processing the visuomotor control loop. Such delays are of the order of 100-250 ms (Wolpert et al., 1998, 2001), which amounts to the time delays in our experiments. Although the dorsal visual stream is capable of performing fast visuomotor transformations, it is possible that switching toward the new target is easier for the gaze than for the arm, due to both the greater physiological complexity of the arm control system and increased delays resulting from longer neural pathways. However, one could state an alternative explanation that relates to the fundamental control strategy in the CNS. Because the arm avoids the obstacle at some safety distance, and the experimental task is designed such that obstacle position is kept constant during the trials, the “buffered” position of the obstacle from the last fixation at the obstacle is a very good reference point for the arm. Land and Furneaux (1997) have shown that information buffering of spatial coordinates acts as an adjutory mechanism when transitions between visuomotor sequential tasks occur. The arm is at the moment when the gaze leaves the obstacle displaced at some distance from to the obstacle and hence neither much adjustment is needed nor very precise visual information is needed to avoid the obstacle. This could be an efficient strategy in terms of the attentional resources considering that there is neither much

[6] It is important to note that Johansson et al. (2001) focused most of their analysis on gaze and arm timing with respect to entering or exiting the so-called “landmark zones”. They defined the landmark zone as an area with the radius 3 ° of visual angle (2 cm) in the work plane in all directions from the corresponding objects in the workspace, including the obstacle. They found that the gaze and arm have almost identical exit times from the obstacle landmark zone. Considering that an approximate overall vertical arm displacement in their experiment was 12 cm, these landmark zones established a coarse representation of the workspace. However, from the plots where precise spatio-temporal measures were presented (Figure 6A in their paper), it can be seen that the difference between the median gaze and arm exit times at the exact location of the obstacle differ approximately 200 ms in favor of gaze exiting first the obstacle. Similar measures of the gaze-arm exit lag hold for the other intermediary targets (e.g. support surface, target switch, bar tool, etc.).

surprise in the task, nor the extreme precision is required. This suggests that the CNS employs “loose” transition between the subtasks, saving valuable, limited attentional resources, whenever prior information about the task suggests that not much change in the workspace is expected and not much accuracy is needed. In the task where sequential movements had very high precision constraints by means of the requirements of precisely touching a target, the gaze exit times were almost always tightly synchronized with the arm exit times (Bowman et al., 2009). The experiment of Bowman et al. (2009) shows that the “tight” switching strategy holds as well.

This analysis shows that the gaze and arm exit times from the obstacle are highly correlated, suggesting strong visuomotor synchronization with respect to the obstacle. The time difference between the gaze and the arm times when switching from the obstacle is non-zero positive, but it remains small compared to the overall task duration.

2.3 Summary and discussion

In this chapter, we presented a human study in which 8 volunteers performed reach and grasp movements to a single target in the presence of an obstacle. We analyzed the kinematics of visuomotor coordination to provide quantitative measurements on the phase relationships across the effectors.

Our human study contributed a quantitative assessment of the eye-arm coordination when performing obstacle avoidance, an issue that has received little attention to date. Precisely, it demonstrated that obstacle avoidance is included in forward planning and modulates the coordinated pattern of the eye-arm motion in a distinctive way. The results of the study: a) quantified the phase relationship between the gaze and arm systems, so as to inform robotic models; and b) provided insights how the presence of an obstacle modulates this pattern of correlations. We showed that the notion of workspace traveled by the hand is embedded explicitly in a forward planning scheme that allows subjects to determine when and when not to pay attention to the obstacle.

We hypothesized that the visuomotor system treats the obstacle as an intermediary target. Our evidence of a systematic pattern that the gaze precedes and leads the motion of the arm through the different landmarks, defining the stages of a sequential task, supports this hypothesis (Johansson et al., 2001).

In summary, the mechanism of the eyes leading the arm was observed in all trials. This study corroborated other findings in the literature on a strong coupling between the arm and eye motion, where the eyes lead the arm in a systematic and coordinated pattern. Additionally, the study supported the hypothesis that the obstacle may act as an intermediary target.

The coordination between the gaze, arm and hand noticed in our human study is implemented in the robotic model that we will present next. The reported existence of the forward planning mechanism for obstacle detection has inspired us to implement the equivalent scheme for robotic obstacle avoidance. In addition to this, the observation that the visuomotor system treats the obstacle as the intermediary target tremendously reduces the computational and architectural complexity of our visuomotor model for obstacle avoidance scenarios. We should emphasize that this study was particularly instrumental in providing us with quantitative data onto which to ground the parameters of our computational model, as we describe next.

3 Robotic Visuomotor Controller Based on the Human Motion Capture Study

In the introductory chapter, we raised up that the problem of visuomotor coordination boils down to two main computational problems. The first fundamental problem of visuomotor coordination is the computation of a sequence of transformations from gaze-centered target encoding to coordinate representations suitable to generate arm and hand movements. The second fundamental problem, once the reference frames are computed, is how to: (a) generate movements of the eyes, arm and hand and (b) how to appropriately coordinate the movements of these effectors. The first problem has been extensively addressed in both neuroscience and robotic community (Hoffmann et al., 2005; Natale et al., 2005, 2007; Hulse et al., 2009; Jamone et al., 2012, 2013). On the other hand, the second problem, on which we focus in this chapter, has received far less attention. Similarly, robotic active gaze allocation to aid complex tasks, such as obstacle avoidance, has not been studied to the appropriate extent. In this chapter, we jointly tackle the problems of coordinated visuomotor control and the gaze integration in a complex prehensile task such as obstacle avoidance.

In this chapter, we present a novel computational model of the coordinated visuomotor control when performing reaching and grasping with and without the presence of obstacles. To guide our modeling, we used the human study described in the previous chapter, in which 8 volunteers performed reach and grasp movements to a single target in the presence of an obstacle. The human study corroborated the coordination pattern of the gaze, arm and hand noticed in previous studies and extended this by confirming that this pattern is present in more complex tasks when the obstacle is introduced in the workspace. We implement this visuomotor coordination pattern in our robotic model. In the human study, we have shown that the notion of workspace traveled by the hand is explicitly embedded in a forward planning scheme that allows subjects to determine when and when not to pay attention to the obstacle. This observation has inspired us to implement a scheme for our model for robotic obstacle avoidance. The results from humans provided significant evidence that the visuomotor system considers the obstacle as an intermediary target in prehensile tasks. Treating the obstacle as an intermediary target of the visuomotor system tremendously simplifies the computational model of the visuomotor controller, from the robotic viewpoint. Finally, in our human study, we found that humans keep a minimum safety distance between the hand and the obstacle when performing prehensile arm movements. We implement this observation in our robotic model, as well. The human study provided quantifiable information about the eye-arm-hand coupling to support the design of the model's parameters.

In our modeling, we extend the Coupled Dynamical Systems (CDS) framework, originally used for arm-hand coordination (Shukla and Billard, 2011), to model the eye-arm-hand coordinated pattern measured in the human study. The CDS framework provides fast and synchronous control of the eyes, the arm and the hand within a single and compact framework, mimicking similar control system found in humans. The parameters of our computational model are estimated based on the data recorded in the human study.

Particularly, we extend the CDS framework for visuomotor coordination to encapsulate: a) model of the eye-arm-hand coupling and b) modulation by an obstacle. In our work, we exploit a biologically inspired notion of forward models in motor control (Wolpert et al., 1998, 2001) and use a model of the dynamics of the reaching motion to predict collisions with objects in the workspace when reaching and grasping the target object. We use the observation from the human study that the obstacle may act as an intermediary target, in order to develop our obstacle avoidance scheme. The objects, which are tagged as obstacles after propagating the forward model, are treated as intermediary targets for the visuomotor system. This approach leads to a simple and computationally lightweight scheme for obstacle avoidance. As an alternative to computationally costly sampling-based algorithms (Kavraki et al., 1996; Kuffner Jr and LaValle, 2000), our approach uses the ability of Dynamical Systems to instantly re-plan the motion in the presence of perturbations. In our obstacle avoidance scheme, the gaze is an important element of the coupled visuomotor mechanism that is actively controlled and tightly bound to manipulation requirements and plans. We validate the usefulness of this model for robot control, by implementing it in experiments involving the visually-guided prehensile motion with obstacle avoidance, in simulation and the real humanoid robot iCub (Metta et al., 2010).

We next provide a short review of the state of the art in robotic visually-aided manipulation.

3.1 Background research

3.1.1 Visually-aided robotic reaching and grasping

Solutions to robotic visual-based reaching follow either of two well-established approaches: techniques that learn visuomotor transformations (Hoffmann et al., 2005; Natale et al., 2005, 2007; Hulse et al., 2009; Jamone et al., 2012), which operate in an open-loop manner, or visual-servoing techniques (Espiau et al., 1992; Mansard et al., 2006; Natale et al., 2007; Chaumette and Hutchinson, 2008; Jamone et al., 2012), which are closed-loop methods. Techniques that learn the visuomotor maps are very appealing because of their simplicity and practical applications. However, these methods suffer from several drawbacks. Models of the visuomotor transformations are learned by using exploratory schemes employed by a robot that are similar to babbling employed during infant development (Vernon et al., 2010). The number of exploratory movements that the robot needs to visit during the exploration is usually of the order of several thousand, or even higher. Such extensive exploration, needed to learn a model, limits the applicability of these methods because it is highly inefficient in time and energy spent. The accuracy of the reaching movement is limited by the accuracy of the eye-arm mapping estimate. Moreover, during the online control, there is no coordinated control of the effectors, in terms of the active, online modulation between the gaze and the arm. These methods often employ the first-fixate-then-start-reach strategy, which is not biologically plausible, considering that the humans simultaneously issue eye and arm commands in head-free visually-guided reaching and grasping tasks (Johansson et al., 2001; Pelz et al., 2001; Hayhoe et al., 2003). Finally, the reaching path is often generated by relying on interpolation between the starting arm state and the computed goal arm state.

On the other hand, visual servoing approaches control the speed of the arm, based on measurements of the visual error between the hand and the target. This approach ensures zero-error reaching, but it requires having the target object and the hand simultaneously in the field of view. Visual servoing does not allow us to produce a family of human-like motion profiles in reaching tasks. The previous work done on the visuomotor coordination did not explicitly address the synchronization pattern of the arm transport and grip component.

A control policy of a robotic hand (or a gripper) is usually a pre-programmed routine that is invoked after the arm reaches the target object, thus its control mechanism is not embodied in the coupled eye-arm control, as in humans.

3.1.2 Robotic obstacle avoidance

Robots operating in cluttered environments have to be able to plan their motion, avoiding collisions with objects in the workspace. There is a large number of obstacle avoidance methods and providing a broad review is not our intended goal. We now provide a brief synopsis of the main trend across these approaches. Recently the most popular methods are sampling-based algorithms (Kavraki et al., 1996; Kuffner Jr and LaValle, 2000). Sampling-based algorithms are very powerful, but cannot meet the demands of rapid motion planning that humans perform almost effortlessly in a fraction of a second. Additionally, robotic obstacle avoidance methods do not consider how the gaze control is involved in the process of obtaining information about the state of obstacles and targets, they usually assume that the environment is somehow known beforehand. Seara et al. (2003) developed an algorithm to actively control the gaze of a humanoid robot in order to support visually guided walking with obstacle avoidance. However, in robotic obstacle avoidance applications involving manipulation information about the environment is obtained either by using passive stereo systems (Khansari-Zadeh and Billard, 2012), or by relying on some special sensors such as Microsoft Kinect^{TM}, laser rangers, etc.[1] (Srinivasa et al., 2012). Having a gaze control strategy for obstacle avoidance is crucial in order to fixate obstacles. Fixations at the obstacles

[1] These sensors are not controlled in terms of the active vision paradigm.

provide accurate visual information about their state, and this information is used to proactively guide the arm-hand system. Failure to provide visual information about obstacles can result in fatal collisions.

3.2 Computational approach and system architecture

In the first part of this section, we introduce the principle of robot control by using time-invariant Dynamical Systems (DS) and the probabilistic approach for estimating the parameters of the system. Furthermore, we extend this formulation for modeling and control of coupled dynamics. Finally, we show how the basic model of eye-arm-hand coordination in the obstacle-free grasping can be extended to handle the obstacle in the workspace.

3.2.1 A single DS and GMM/GMR

The motion of our system is represented through the state variable $\xi \in \mathbb{R}^d$, symbolizing retinal coordinates representing the gaze state, Cartesian coordinates for the arm state, and finger joint angles for the hand state. N recorded demonstrations of the task yield the data set $\left\{\xi_t^n, \dot{\xi}_t^n\right\}$, $\forall t \in [0, T_n]$; $n \in [1, N]$, of the robot's states and state derivatives at particular time steps t, where T_n is the number of samples in the n-th demonstration. We posit that the recorded data samples are instances of the motion governed by a first-order autonomous differential equation:

$$\dot{\xi} = f(\xi) + \epsilon \tag{3.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous and continuously differentiable function, with a single equilibrium point $\dot{\xi^*} = f(\xi^*) = 0$, and ϵ is a zero-mean Gaussian noise term. The noise term encapsulates both sensor inaccuracies and errors inherited from human demonstrations. Time-invariance provides inherent robustness to temporal perturbations. In order to achieve robustness to displacement in the position of the target, the robot's state variable ξ is represented in the target's reference frame.

We use the Gaussian Mixture Model (GMM) to encode the motion in a probabilistic framework. The GMM defines a joint probability distribution function $\mathcal{P}(\xi_t^n, \dot{\xi}_t^n)$ over the set of data from demonstrated trajectories as a mixture of K Gaussian distributions (with π^k, $\mu^k$and Σ^k being the prior probability, the mean value and the covariance matrix of the k-th Gaussian, respectively):

$$\mathcal{P}\left(\xi_t^n, \dot{\xi}_t^n\right) = \sum_{k=1}^{K} \pi^k \mathcal{N}(\xi_t^n, \dot{\xi}_t^n; \mu^k, \Sigma^k), \tag{3.2}$$

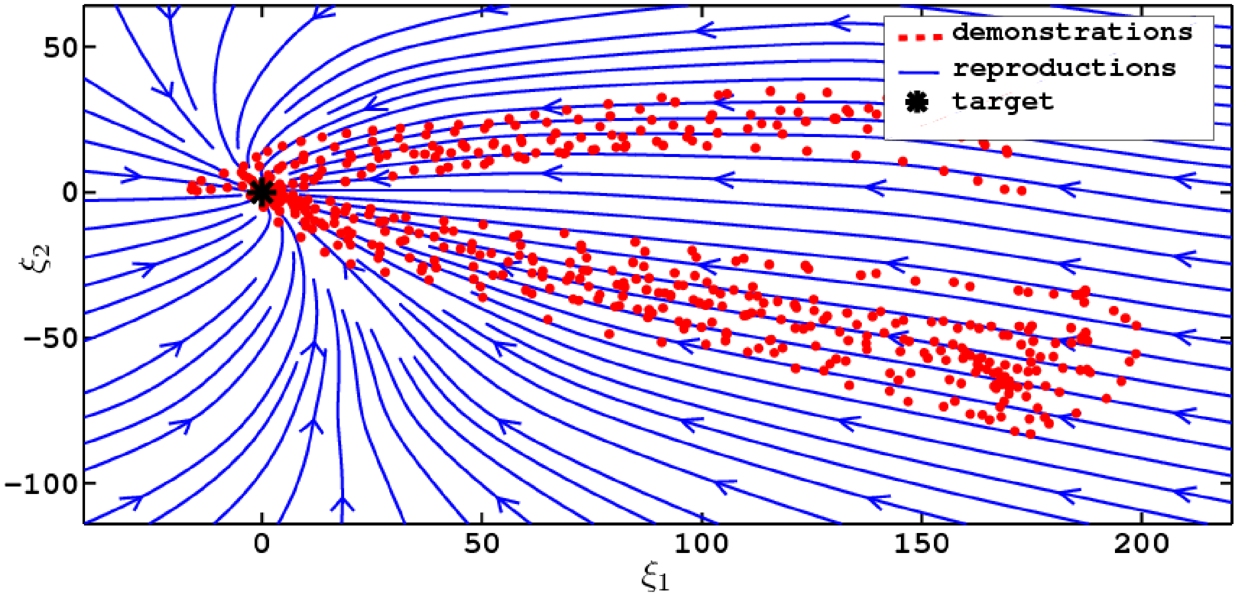


Figure 3.1: Learning and reproducing a motion with a single time-invariant DS. Given a set of demonstrations (red points), we build an estimate of the underlying dynamics. The asymptotic stability of the DS guarantees that the target (black star) will be reached. The DS, for a given robot state, computes a velocity vector that moves the robot state toward the target, hence it can be illustrated with streamlines (blue lines) in the state space that steer the robot state toward the target.

where each Gaussian probability distribution is defined as:

$$\mathcal{N}(\xi_t^n, \dot{\xi}_t^n; \mu^k, \Sigma^k) = \frac{1}{\sqrt{(2\pi)^{2d} \mid \Sigma^k \mid}} e^{-\frac{1}{2}(([\xi_t^n, \dot{\xi}_t^n]-\mu^k)^T(\Sigma^k)^{-1}([\xi_t^n, \dot{\xi}_t^n]-\mu^k)}, \tag{3.3}$$

where the mean and the covariance matrix are defined as:

$$\mu^k = \begin{pmatrix} \mu_\xi^k \\ \mu_{\dot{\xi}}^k \end{pmatrix} \text{ and } \Sigma^k = \begin{pmatrix} \Sigma_{\xi\xi}^k & \Sigma_{\xi\dot{\xi}}^k \\ \Sigma_{\dot{\xi}\xi}^k & \Sigma_{\dot{\xi}\dot{\xi}}^k \end{pmatrix}. \tag{3.4}$$

We use the Stable Estimator of Dynamical Systems (SEDS) (Khansari-Zadeh and Billard, 2011) to compute the GMM parameters. The SEDS ensures global stability of the noise-free estimate of the underlying dynamics, denoted as $\hat{f}$.

Taking the posterior mean estimate of $\mathcal{P}(\dot{\xi}_t^n \mid \xi_t^n)$ yields an estimate of $\hat{\dot{\xi}} = \hat{f}(\xi)$, a function that approximates the model dynamics through a mixture of K Gaussian functions:

$$\hat{\dot{\xi}} = \sum_{k=1}^{K} h^k(\xi)\left(A^k \xi + b^k\right), \tag{3.5}$$

where $h^k(\xi)$, A^k and b^k are defined as:

$$\begin{cases} h^k(\xi) = \frac{\pi^k \mathcal{N}(\xi;\mu^k,\Sigma^k)}{\sum_{i=1}^{K} \pi^i \mathcal{N}(\xi;\mu^i,\Sigma^i)} \\ A^k = \Sigma^k_{\dot{\xi}\xi}(\Sigma^k_{\xi\xi})^{-1} \\ b^k = \mu^k_{\dot{\xi}} - A^k \mu^k_{\xi}. \end{cases} \tag{3.6}$$

A toy example with a 2-dimensional DS, which illustrates the principles of encoding the demonstrated motion and robot control by using a time-invariant DS, is presented in Figure 3.1.

3.2.2 Coupled Dynamical Systems

Recent work (Shukla and Billard, 2011) has shown the benefits of explicitly learning a coupling between the arm DS and the finger DS over modeling motions of the physical systems with a single extended DS. The problem associated with learning one high-dimensional dynamical model that guides the motion of two physical systems is that an explicit following of correlations shown in demonstrations between the two coupled dynamics is not guaranteed. This could be a problem if the robot is perturbed far from the region of the demonstrated motion, as the behavior of the dynamical systems may not be correctly synchronized. The loss of coordination between the reach and grasp components might lead to failure of the overall prehensile task even when the individual dynamical systems converge to their attractors. An approach adopted in Shukla and Billard (2011) is to separately learn two dynamics and then learn a coupling between them. This approach ensures that the two DS will converge to their attractors, following a learned pattern of coordination between them. The approach, where the arm and hand DS are learned separately and then coupled explicitly, ensures that the behavior of the two systems is correctly synchronized, even when the motion is abruptly perturbed far from the motion recorded in human demonstrations. For more details about general properties of the CDS, see Shukla and Billard (2011).

Extended CDS architecture and learning

We extend the original CDS architecture with in total five building "blocks": three dynamical systems and two coupling blocks between them. They are organized in the following order: eye dynamics → eye-arm coupling → arm dynamics → arm-hand coupling → hand dynamics, where the arrow direction indicates the direction of control signals. The gaze DS is the master to the arm DS, and the arm DS is the master to the hand DS. There is a coupling block between each master and its slave. The major assumption is that the modulation signals between them flow only in the direction from the master to the corresponding slave, i.e. the dynamics of the slave is modulated with control signals coming from its master, not vice versa. The master system evolves independently of its slave. Figure 3.2 illustrates the architecture of the CDS, and the principles of

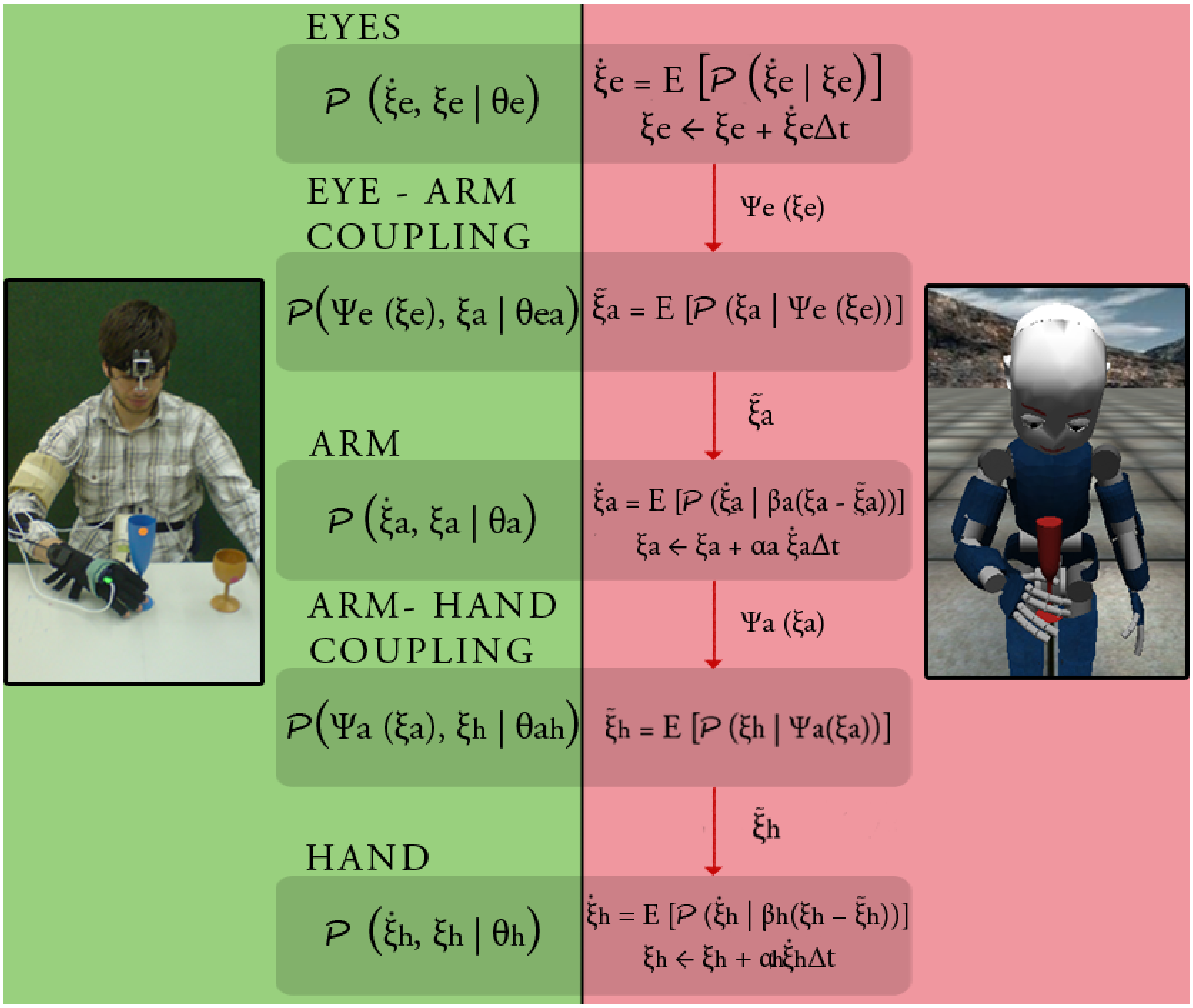

EYES
$P(\dot{\xi}_e, \xi_e \mid \theta_e)$
$\dot{\xi}_e = E[P(\dot{\xi}_e \mid \xi_e)]$
$\xi_e \leftarrow \xi_e + \dot{\xi}_e \Delta t$
EYE - ARM COUPLING
$\Psi_e(\xi_e)$
$P(\Psi_e(\xi_e), \xi_a \mid \theta_{ea})$
$\tilde{\xi}_a = E[P(\xi_a \mid \Psi_e(\xi_e))]$
$\tilde{\xi}_a$
ARM
$P(\dot{\xi}_a, \xi_a \mid \theta_a)$
$\dot{\xi}_a = E[P(\dot{\xi}_a \mid \beta_a(\xi_a - \tilde{\xi}_a))]$
$\xi_a \leftarrow \xi_a + \alpha_a \dot{\xi}_a \Delta t$
ARM- HAND COUPLING
$\Psi_a(\xi_a)$
$P(\Psi_a(\xi_a), \xi_h \mid \theta_{ah})$
$\tilde{\xi}_h = E[P(\xi_h \mid \Psi_a(\xi_a))]$
$\tilde{\xi}_h$
HAND
$P(\dot{\xi}_h, \xi_h \mid \theta_h)$
$\dot{\xi}_h = E[P(\dot{\xi}_h \mid \beta_h(\xi_h - \tilde{\xi}_h))]$
$\xi_h \leftarrow \xi_h + \alpha_h \dot{\xi}_h \Delta t$

Figure 3.2: CDS-based robotic eye-arm-hand coordination. Left (green) part of the figure shows how the CDS model is learned. Reproduction of the motion on the robot is shown on the right side of the figure (red part). CDS consists of five building "blocks": three dynamical systems (the eyes, the arm and the hand) and two coupling models: eye-arm coupling and arm-hand coupling.

learning and the reproduction of the coordinated motion.

The state of the eyes is denoted by $\xi_e \in \mathbb{R}^2$, the state of the arm is $\xi_a \in \mathbb{R}^3$, and the state of the hand is $\xi_h \in \mathbb{R}^9$. The eye state ξ_e is represented as the distance between the position of the gaze and the position of a visual target in retinal coordinates (i.e. retinal error). The arm state ξ_a is represented as the distance in Cartesian coordinates between the palm center and the final palm position with respect to the target object. The hand state ξ_h is expressed as the difference between the current hand configuration and the goal hand configuration, i.e. hand configuration adopted when the target object is grasped. In other words, the attractors of the eye, arm and hand DS are placed at the target projection in the retinal plane, its Cartesian position in the workspace and at the corresponding hand configuration when the target is grasped, which is formally expressed as:

$\xi_e^* = 0$, $\xi_a^* = 0$ and $\xi_h^* = 0$, respectively.

Our CDS model of eye-arm-hand coordination is built in the following manner. We first learn separately joint probability distributions that encode the eye dynamics $\mathcal{P}(\dot{\xi}_e, \xi_e \mid \theta_e)$, arm dynamics $\mathcal{P}(\dot{\xi}_a, \xi_a \mid \theta_a)$ and the hand dynamics $\mathcal{P}(\dot{\xi}_h, \xi_h \mid \theta_h)$. Then we learn the joint distribution for eye-arm coupling $\mathcal{P}(\Psi_e(\xi_e), \xi_a \mid \theta_{ea})$ and arm-hand coupling $\mathcal{P}(\Psi_a(\xi_a), \xi_h \mid \theta_{ah})$, where θ_e, θ_a, θ_h, θ_{ea} and θ_{ah} denote the GMM parameters, and $\Psi_e(\xi_e)$ and $\Psi_h(\xi_h)$ denote the coupling functions. The GMMs that encode the dynamics of the eyes, arm dynamics and the hand dynamics are learned using the SEDS algorithm, for more details see Khansari-Zadeh and Billard (2011). The GMMs that model eye-arm and arm-hand coupling are learned with the Expectation-Maximization (EM) algorithm (Bishop, 2007).

Two open parameters, α and β, allow for an additional fine-tuning of the characteristics of the slave response (a and h subscripts denote whether they modulate the arm motion or the hand motion, respectively). The speed is modulated with the scalar α, and the amplitude of the motion is tuned by changing the value of the scalar β. Some robots can move faster than humans, hence by using larger values for α_a and α_h, one can exploit the robot's fast reaction times. One can tailor the amplitudes of reactions to perturbations, suitable for a robot platform and a given task, by modulating the values of β_a and β_h.

Figure 3.3 illustrates the CDS model learned from demonstrations.

CDS reproduction

Algorithm 1 shows how the robotic eye-arm-hand coordination is performed with the CDS. The eye DS evolves independently in time and leads the whole system. The eye state velocity $\dot{\xi}_e$ is generated by conditioning the eye dynamics model on the current eye state. The learned GMMs are conditioned by computing the Gaussian Mixture Regression (GMR) function (Eq. 3.5), for more about the GMR see Sung (2004). The eye state variable is incremented by adding the computed velocity multiplied by the time step Δt to its current value ξ_e. The desired arm state value $\tilde{\xi_a}$ is inferred from the eye-arm coupling model by conditioning on the eye-arm coupling function $\Psi_e(\xi_e)$. The arm velocity $\dot{\xi}_a$ is computed by conditioning the arm dynamics model on the difference between the current and desired value $\xi_a - \tilde{\xi_a}$. The arm state variable is incremented by adding the computed velocity multiplied by Δt to its current value ξ_a. The desired hand state value $\tilde{\xi_h}$ is obtained by conditioning the arm-hand coupling model on the arm-hand coupling $\Psi_a(\xi_a)$. The hand velocity $\dot{\xi}_h$ is inferred by conditioning the hand dynamics model on $\xi_h - \tilde{\xi_h}$. Finally, the hand state variable is incremented by adding the computed velocity multiplied by Δt to its current value ξ_h. The eyes, arm and hand reach commanded states and the loop is reiterated until the target object is grasped.

3.2.3 Eye-arm-hand coordination for obstacle avoidance

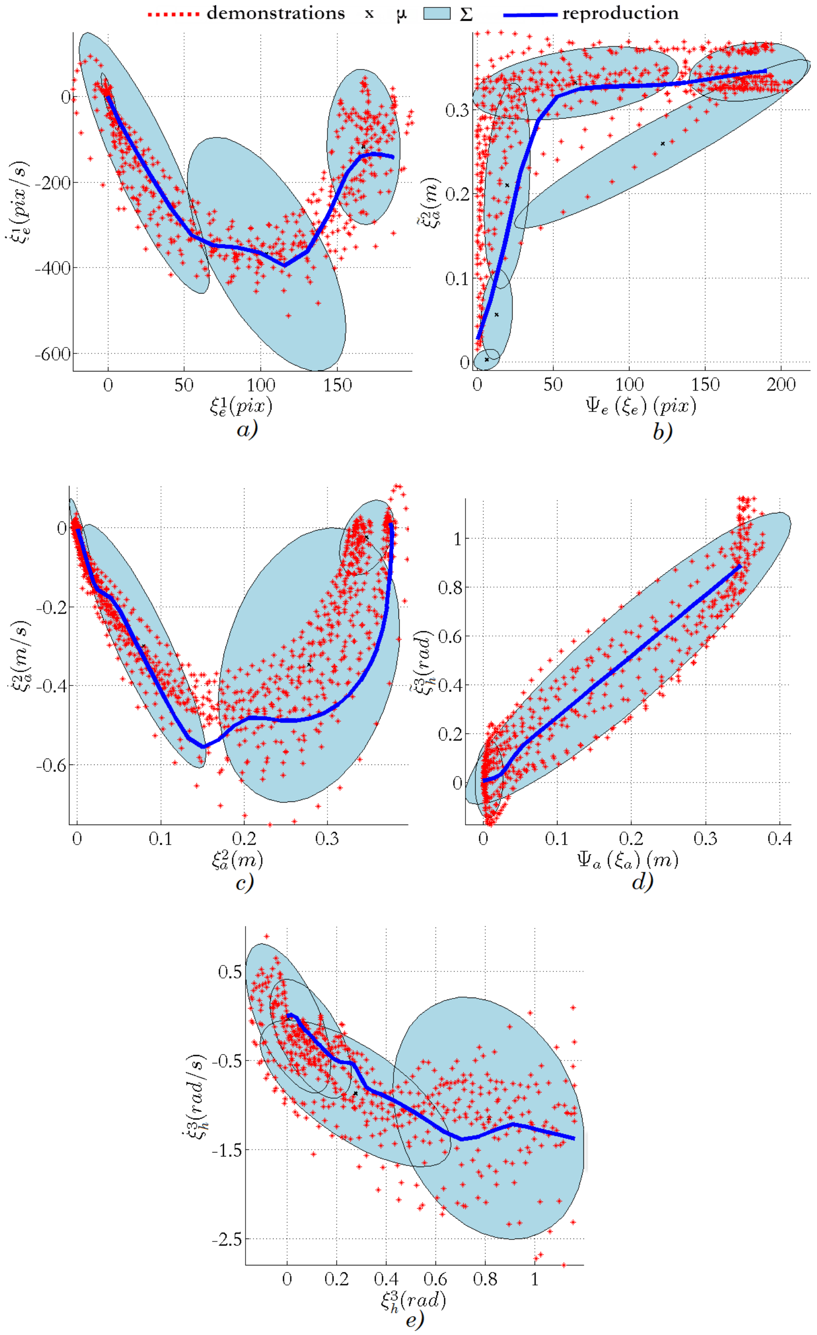


Figure 3.3: Learned CDS eye-arm-hand coordination model: a) eye dynamics, b) eye-arm coupling, c) arm dynamics, d) arm-hand coupling and e) hand dynamics. For simplicity of graphical representation, we plotted the CDS model for one gaze position, one arm position and one hand position. The eye state is presented with horizontal gaze coordinate, denoted as ξ_e^1 .The arm state is presented with Cartesian coordinate that corresponds to the direction of the major hand displacement in the task, denoted as ξ_a^2. The hand state is represented by the thumb proximal joint, denoted as ξ_h^3. Superposed to the datapoints, we see the regression signal (plain line) and the different Gaussian distributions (elliptic envelopes) of the corresponding Gaussian Mixture Models.

The extension of the CDS eye-arm-hand controller for obstacle avoidance is grounded on our hypothesis that the obstacle acts as the intermediary target for the visuomotor system in reaching and grasping tasks, see Chapter 2.

In order to define which objects in the workspace are obstacles for the realization of the intended reach-and-grasp tasks, we use a planning scheme to estimate the consequences of future actions. More specifically, the motion of the arm toward the target is estimated by integrating the dynamics of the extended CDS until each DS reaches its attractor. We integrated only the eye-arm part

```
do
  $General:$
  – query frames from cameras
  – read the current hand position from forward kinematics
  – read the hand joints from encoders
  – recognize and segment the target object
  – estimate the position of the target in both retinal
    and Cartesian coordinates
  – compute $\xi_{\mathrm{e}}$, $\xi_{\mathrm{a}}$ and $\xi_{\mathrm{h}}$
  $Gaze:$
  if gaze is not at target then
    $\dot{\xi}_e \leftarrow \mathbb{E}\left[\mathcal{P}\left(\dot{\xi}_e \mid \xi_e\right)\right]$
    $\xi_e \leftarrow \xi_e + \dot{\xi}_e \Delta t$
    – solve gaze IK
    – move the eyes and head to new joint conf.
  end if
  $Eye - arm\ coupling:$
  $\tilde{\xi}_a \leftarrow \mathbb{E}\left[\mathcal{P}\left(\xi_a \mid \Psi_e\left(\xi_e\right)\right)\right]$
  $Arm:$
  if the arm is not at target then
    $\Delta\xi_a \leftarrow \xi_a - \tilde{\xi}_a$
    $\dot{\xi}_a \leftarrow \mathbb{E}\left[\mathcal{P}\left(\dot{\xi}_a \mid \beta_a \Delta\xi_a\right)\right]$
    $\xi_a \leftarrow \xi_a + \alpha_a \dot{\xi}_a \Delta t$
    – solve arm IK
    – move the arm and the torso to new joint conf.
  end if
  $Arm - hand\ coupling:$
  $\tilde{\xi}_h \leftarrow \mathbb{E}\left[\mathcal{P}\left(\xi_h \mid \Psi_a\left(\xi_a\right)\right)\right]$
  $Hand:$
  if the hand is not at target then
    $\Delta\xi_h \leftarrow \xi_h - \tilde{\xi}_h$
    $\dot{\xi}_h \leftarrow \mathbb{E}\left[\mathcal{P}\left(\dot{\xi}_h \mid \beta_h \Delta\xi_h\right)\right]$
    $\xi_h \leftarrow \xi_h + \alpha_h \dot{\xi}_h \Delta t$
    – move the hand to new joint conf.
  end if
until object grasped
```

Algorithm 1: CDS eye-arm-hand coordination

of the whole CDS, ignoring the hand's DS, as our collision checking scheme is relatively simple. The arm end-effector is modeled as a point that moves along the estimated trajectory. Obstacle objects in the workspace are modeled as cylinders. The dimensions of a modeling cylinder should enclose the actual dimensions of the object, but should also account and compensate for the fact that the hand was modeled as a point. This is achieved by expanding the modeling cylinder for some predetermined, fixed distance (we used 5 cm for both radius and height) from the dimensions where it fits exactly around the object. By taking this approach, we are able to reliably detect collisions with the fingers in our forward planning scheme, even though the hand is modeled as a point. The argument for using this simplistic collision checking scheme is our attempt to minimize additional computational load in the control loop.

An object is tagged as an obstacle when the trajectory of the end-effector intersects with a cylinder modeling the object (certain collision), or when the cylinder lies within the area where it is very likely that it will collide with the forearm (very likely collision). For the motions we consider here and by observing the iCub's body, we define this area as the slice of the workspace enclosed by the estimated trajectory of the end-effector and the coronal plane of the body.

As suggested earlier on, we consider the eye-arm-hand coordination as a composition of two segments: a motion from the starting position toward the obstacle and from the obstacle toward the target object. Individual segments of the coordinated motion (from the starting point to the obstacle, and from the obstacle to the target) are performed in a manner presented in Algorithm 1. In the first part of the task, the arm DS moves under the influence of the attractor placed at the via-point. The hand DS is driven by the attractor placed at the hand configuration when the palm reaches the closest point (along the trajectory computed ahead of time) to the obstacle. Coupling the hand motion with respect to the obstacle is advantageous because it provides a preshape of the hand such that collisions between the fingers and the obstacle are eluded during obstacle avoidance manipulation, even in scenarios where the obstacle is suddenly perturbed during the ongoing task (see Figure 3.5). Our approach for adapting the reaching hand motion to avoid obstacles is motivated by several studies that have reported significant effects of the obstacle on all aspects of grasp kinematics (e.g. grip duration, grip aperture, time to peak aperture, distance to peak aperture, etc) (Saling et al., 1998; Tresilian, 1998; Mon-Williams et al., 2001). Tresilian (1998) interprets these effects as subtle adjustments of the transport and grip components that support obstacle avoidance. In their obstacle avoidance experiment, Saling et al. (1998) observed a systematic high correlation of arm transport parameters (transport time, time to peak velocity, time to peak acceleration, etc.) with almost all grip kinematic parameters (grip closure time, time to peak aperture, time to peak opening velocity, grip opening velocity, etc.). This result is a very strong indication that the arm and the hand remain coupled even when obstacles cause considerable alternations of the prehensile motion, compared to the no-obstacle condition.

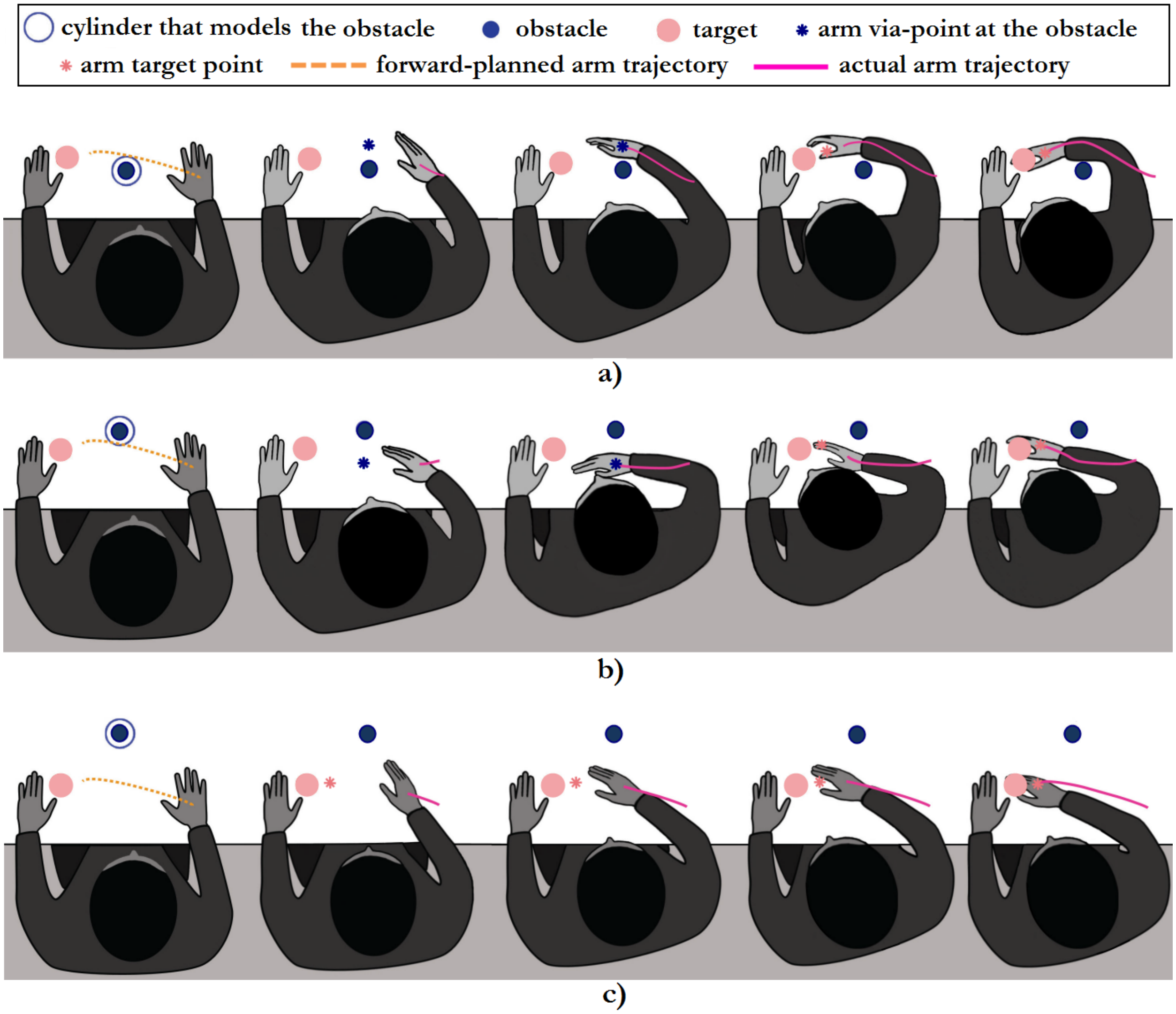
The goal hand configuration for passing the obstacle at the closest distance is obtained by observing the average hand configurations of our subjects in obstacle avoidance trials. We adapted, with slight modifications, the computed average hand configuration to match the kinematics of the iCub's hand. We did a similar procedure to obtain the goal hand configurations with respect to the target object.

The position of the via-point is determined with respect to the obstacle, such that its displacement vector from the obstacle position is oriented in either an anterior or posterior direction, for the length that corresponds to some safety distance d_{safety} between the centroid of the palm and the obstacle. We choose the direction of a displacement of the via-point (anterior or posterior) to correspond to the side of the obstacle where a collision is estimated to occur. In the second part of the task, after the obstacle is passed, the CDS is driven toward the object to be grasped. As mentioned before, hand adaptation, with respect to the obstacle, serves to support collision avoidance; whereas hand adaptation, with respect to the target, ensures coordinated and stable grasping of the target as the arm reaches it. Predefining the safety distance at which the hand passes the obstacle is based on the study of Dean and Brüwer (1994), who found that participants kept a minimum distance between the pointer and obstacles when performing planar pointing arm movements. In our human study, the measured mean value of this safety distance is 0.142 m with a small value of standard deviation 0.01 m, which can be considered as a consistent observation of the mechanism employed by the motor control system to keep the limb at the safety distance from the obstacle, as presented in Dean and Brüwer (1994).

The arm end-effector passing through the via-point at d_{safety} from the obstacle and hand adaptation, with respect to the obstacle, ensures that the hand will not collide with the obstacle. However, the end-effector obstacle avoidance mechanism, we just described, considers solely collisions with the end-effector and hence ignores a collision with the rest of the arm. We benefit from controlling the arm in Cartesian coordinates and from having an efficient inverse kinematics (IK) solver (Pattacini et al., 2010) that is able to handle two tasks: to find suitable joint configuration (primary task) and to keep solutions as close as possible to a desired arm rest posture (secondary task). By having the IK method that can solve for the goal Cartesian position by trying to keep joints close to a given rest posture, we can modulate the robot's motion in the operational space by providing joint rest postures suitable for obstacle avoidance. Our approach to the problem of finding suitable joint postures is to learn these joint postures from human demonstrations, as human demonstrations in obstacle avoidance tasks encode inherently favorable joint configurations.

Here we learn correlations between the joints that provide major contributions to obstacle avoidance manipulation and arm position in the operational space. The joints chosen to define the rest position are torso pitch and yaw, and shoulder joints corresponding to adduction-abduction and flexion-extension. Hence, we proceed with learning the joint probability distribution $\mathcal{P}(q, x)$, where $q \in \mathbb{R}^4$ denotes the joint rest posture and $x \in \mathbb{R}^3$ denotes the Cartesian position of the palm.

An adaptation of the arm posture for obstacle avoidance is done in the following manner. When reaching for a visuomotor target (the obstacle object or the grasping object), the CDS system infers the state velocities, as explained earlier. By integrating the arm velocity, we obtain a new arm state. By taking the posterior mean estimate of $\mathcal{P}(q \mid x)$, we infer a favorable rest posture. Finally, the IK solver optimizes for joint angles that correspond to the desired Cartesian position, while


cylinder that models the obstacle
obstacle
target
arm via-point at the obstacle
arm target point
forward-planned arm trajectory
actual arm trajectory

a)

b)

c)

Figure 3.4: A scheme that illustrates forward planning and obstacle avoidance. After forward integrating the CDS model, an obstacle object (dark blue disk) is identified as an obstructing object if the estimated arm motion (dashed orange line) intersects with a cylinder (dark blue circle) that models the obstacle (certain collision), or when the cylinder lies within the area where it is very likely that it will collide with the forearm (very likely collision). If the obstacle object is identified to obstruct the intended motion, then the motion of the visuomotor system is segmented: from the start to the obstacle and from the obstacle to the target. When reaching to avoid the obstacle, the arm DS moves under the influence of the attractor placed at the via-point with respect to the obstacle (dark blue star). The direction of a displacement of the via-point (anterior or posterior) is chosen to correspond to a side of the obstacle where a collision is estimated to occur: anterior side (a) or posterior side (b). If forward planning scheme does not detect a collision with the obstacle object (c), the visuomotor system is driven to the target object, i.e. the obstacle is ignored. The light red star represents the goal arm position with respect to the target object (light red disk). Figures show execution of eye-arm-hand coordination from the start of the task (left) until successful grasp completion (right).

attempting to keep the four joints as close as possible to the suggested values from the model. Figure 3.4 illustrates the obstacle avoidance scheme. While this does not ensure that the robot's arm will never collide with the obstacle, in practice, we found that this resulted in a successful obstacle avoidance motion.

3.2.4 Robot vision system

The requirements for real-time adaptation to perturbations in dynamic environments impose the demand for real-time update of information obtained from the sensory system. In order to compute the position of objects in every cycle of the control loop, the total time devoted to visual computation in our system has to be reduced to the order of ~10 ms for both cameras in the binocular setup of the iCub robot. This is a very hard constraint to achieve in a robotic system, even by using modern computing hardware with multicore processing units. In order to achieve the aforementioned requirement, we designed the visual system to use minimal computational resources.

We use an image processing scheme similar to the one proposed in Metta et al. (2004). We convert 320×240 images streamed from the cameras to 150×150 log-polar images. By transforming the images to the log-polar domain, we reduce the amount of visual information to be processed, affecting neither the field of view nor the image resolution at the fixation point. Besides the computational benefits, log-polar mapping is biologically plausible because it approximates the cone distribution in the retina and the mapping from the cone cells to the primary visual cortex of primates (Bernardino and Santos-Victor, 1999; Javier Traver and Bernardino, 2010). The image processing is done in the RGB color space, by using a pixel-by-pixel color segmentation algorithm. The same procedure is applied for detection of the target and the obstacle, thus for simplicity of explanation we will here use the term "object". After the images are segmented, we apply binary morphological operations to remove outliers, and we group segmented regions in blobs. The centroid of the biggest blob in each image is back-projected from the log-polar domain to the original image coordinates. The distance between the principal point of one camera (we chose the right camera) and the center of the object blob in the visual field represents the eye state ξ_e, which is the input to the gaze DS. The position estimation of the objects in the workspace is done by triangulating the centroids of the blobs for the left and right camera. The other camera is controlled in a coordinated manner such that both cameras have a fixation point at the estimated head-object distance in the Cartesian coordinates. The distance between the hand and the estimated position of the object represents the arm state ξ_a that is the input to the arm controller. Algorithm 1 illustrates the flow of visuomotor information processing in our model.

The decreasing visual acuity from the fovea to the periphery implies that we get a more precise estimate of the object position at the point of fixation, and the less accurate estimation in the periphery of the visual field. Because we control the gaze and embed the gaze state to the motor control mechanism, we can inherently and efficiently deal with imprecision in the position estimation associated with non-uniform visual acuity in log-polar images. The CDS drives the gaze, arm and hand toward the object using the pose information (in retinal and Cartesian coordinates) obtained from the vision system. As the gaze moves toward the object in every cycle of the control loop, we update the system with a more precise re-estimate of the object position. Before the hand comes close to the object, the gaze fixates the object, and we get the precise information about the object position, which is crucial for successful grasping and obstacle avoidance. Our time-independent CDS automatically adapts to the re-estimate of the object positions obtained from such a non-uniform resolution processing scheme.

For experiments with the real iCub robot, we use the Viola-Jones detector (Viola and Jones, 2001) in addition to the basic color-based segmentation. We use the additional detector in order to eliminate false-positives detections that are a common consequence of color-based segmentation in an unstructured workspace. In other words, we use this detector to verify our color-based detection. The Viola-Jones detector operates on the images streamed from the camera, not in the log-polar domain. When both detectors agree, we update information about the positions of the objects in the workspace, when the detectors do not agree we rely on the previously agreed position. Because the Viola-Jones detector is more computationally demanding, we run it once in every 4 cycles of the control loop.

3.3 Results

3.3.1 Model learning

We learn the CDS model by using the data gathered during the human trials, described in Chapter 2. The parameters of the SEDS algorithm (i.e. maximum number of iterations, optimization criterion, etc.) and the number of Gaussian mixtures (Section 3.2.1) are determined by using a grid-search with 10-fold crossvalidation on the RMSE between the recorded motion and retrieved trajectories from the model. The list of parameter combinations is sorted in ascending order with respect to the value of the RMSE. For each combination of parameters, we visually assess regression plots retrieved from the model. This method is necessary because the small value of the RMSE between the trajectories retrieved from the model and the demonstrated trajectories does not necessarily imply that the inferred paths always have natural-looking and smooth profiles. In other words, the measure of the RMSE provided an initial pool of good candidates, whereas we made the final choice based on the smoothness and the “natural” profile of retrieved paths. The plots for the model we chose are represented in Figure 3.3.

We use $\Psi_e(\xi_e) = \| \, . \, \|$, $\Psi_a(\xi_a) = \| \, . \, \|$ and the values of parameters α_a, α_h, β_a and β_h are set to 1. For the choice of the eye-arm coupling function, we tested performance of four different coupling functions: (1.) $\Psi_e(\xi_e) = \xi_e^2$ (vertical gaze coordinate), (2.) $\Psi_e(\xi_e) = \xi_e^1$ (horizontal gaze coordinate), (3.) $\Psi_e(\xi_e) = \xi_e$ (both gaze coordinates) and (4.) $\Psi_e(\xi_e) = \| \, . \, \|$. We used the average

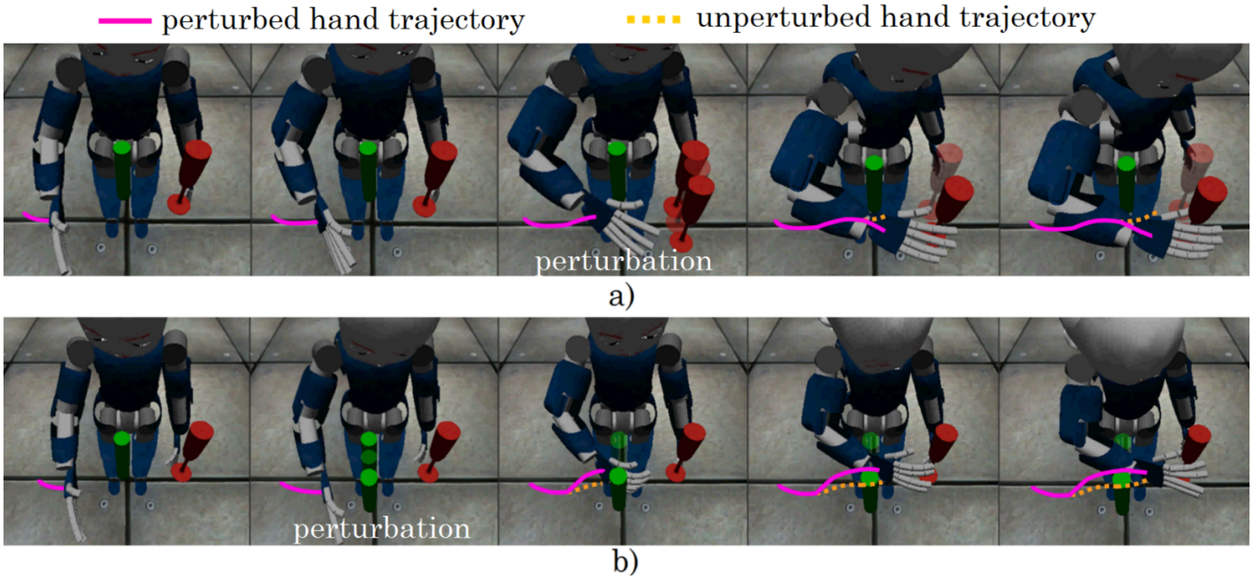
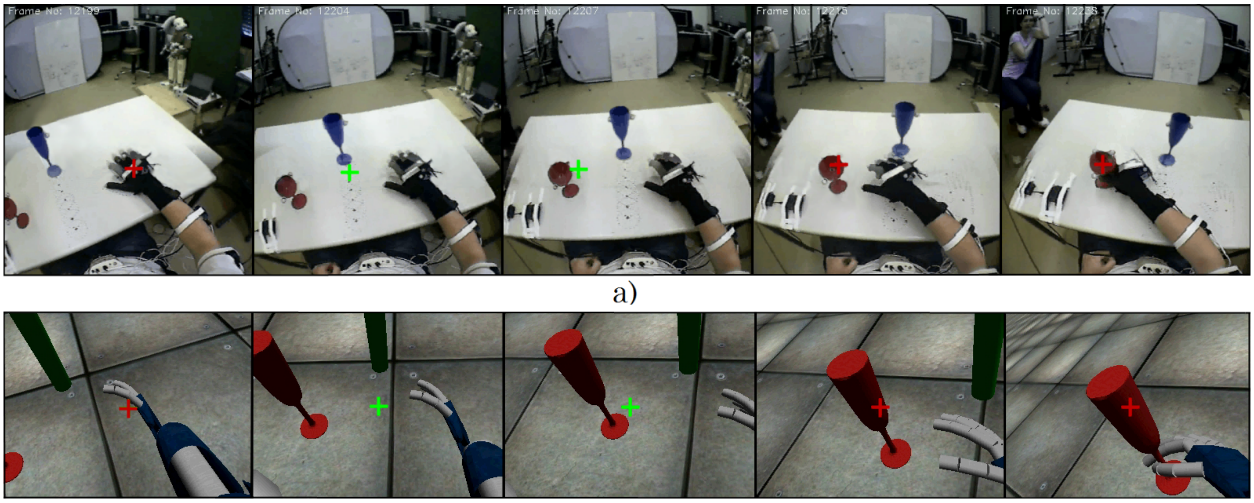


Figure 3.5: Experiments of visually-guided reaching and grasping in the iCub's simulator, with the presence of the obstacle and perturbations. The obstacle is an intermediary target for the visuomotor system, hence obstacle avoidance is divided into two sub-tasks: from the start position to the obstacle (via-point) and from the obstacle to the grasping object. Figures show execution of eye-arm-hand coordination from the start of the task (left) until successful grasp completion (right). Figures in the upper row (a) present a scenario when the target object (red champagne glass) is perturbed during the motion (perturbation occurs in the third frame from the left). Visuomotor coordination when the obstacle is perturbed during manipulation is shown in the bottom row (perturbation in the second frame). The orange line shows the trajectory of the hand if there is no perturbation. The purple line is the actual trajectory of the hand from the start of the unperturbed motion, including the path of the hand after perturbation, until successful grasping. In both scenarios (target perturbed and obstacle perturbed), the visuomotor system instantly adapts to the perturbation and drives the motion of the eyes, arm and the hand to a new position of the object.

absolute point-to-point differences from all demonstrated trajectories and retrieved trajectories from the models as a measure of how well these coupling functions perform. The best results are obtained by the norm coupling function. Our motivation for using $\| \, . \, \|$ function for arm-hand coupling is based on the previous work on hand-arm coupling, see Shukla and Billard (2011). The choice of these particular coupling functions can be considered biologically plausible. The choice of $\| \, . \, \|$ for arm-hand coupling is supported by the physiological studies (Haggard and Wing, 1991, 1995) that reported strong coupling of the hand preshape with respect to the distance from the target object in reach-for-grasping tasks. The choice of $\| \, . \, \|$ for eye-arm coupling function is supported by the fact that retinal distance in foveated vision directly affects the quality of visual information that is used by the motor system for planning and performing manipulation, as visual acuity decreases with distance from the fovea (Land et al., 1999; Land, 1999; Liversedge and Findlay, 2000; Hayhoe and Ballard, 2005). All α and β parameters are set to 1 in order to ensure an unaltered reproduction profile of visuomotor coordination learned from recorded human demonstrations.



a)

b)

Figure 3.6: The visuomotor system ignores an obstacle object when it is not relevant to manipulation, i.e. the obstacle object that does not affect the intended motion is not visually salient for the gaze. Analysis of the WearCam recordings from the human trials (a) reveals that subjects do not fixate the obstacle object (blue champagne glass) in the workspace when it does not obstruct intended reaching and grasping movements. Our CDS eye-arm-hand model shows the same behavior (b), ignoring the obstacle object (green cylinder), when the forward planning scheme estimates that the object does not obstruct the prehensile movement. The snapshots show task from the start (left) until completion of the successful grasp (right).

3.3.2 Model validation for robot control

We conduct a set of experiments with the iCub robot to evaluate the performance of our approach for the visuomotor coordination. Due to hardware constraints of the real robot, we perform perturbation experiments and experiment with obstacle avoidance in the iCub simulator. Unperturbed obstacle-free reaching and grasping experiments are conducted with the real robot.

In our experiments, we validate the ability of the CDS controller on the iCub robot to reproduce the same task of visually guided obstacle-free reaching and grasping similar to the one that humans performed in our trials, together with the advocated robustness of the model to perturbations and the ability to handle the obstacles in the workspace.

We present here the most demanding experiment we perform to validate our approach. In each run, the object to be grasped is placed at a randomly computed position within a 15 cm cube in the workspace. Figure 3.5 shows an obstacle scenario where we test coordinated manipulation with sudden perturbations of the target object and the obstacle, respectively. To introduce perturbations on-the-fly during reaching for the target, we implement a pre-programmed routine in the simulator to abruptly change the position of the object (target or obstacle) when the hand approaches it at some predefined distance, which varies from trial to trial from 0.09 m to 0.15 m. The robot's end-effector avoids the obstacle when reaching for grasping in two task segments: (1.) start position →

via-point at d_{safety} from the obstacle and (2.) via-point at d_{safety} from the obstacle → grasping object. This safety distance in the human trials is $d_{safety} = 0.142 \pm 0.01\,\text{m}$. We rescale the safety distance from human trials by 2, because the dimensions of the iCub are similar to those of a 3.5-year-old child, hence it has a smaller workspace than our adult subjects. Once the obstacle is reached, the target for the visuomotor system is changed, and the eye-arm-hand motion is directed to the object to be grasped. The IK solver adapts the arm rest posture to be as close as possible to the output inferred from the model learned from human demonstrations. Figure 3.6 shows how human subjects ignore the obstacle when it does not obstruct the intended motion, and the same pattern produced by our visuomotor robotic controller.

Because the eye state is the distance between the position of gaze and the position of a visual target in retinal coordinates, and the arm state is represented with respect to the position of the object in the Cartesian space, both variables are instantly updated when the perturbation occurs, see Figure 3.7. The DS of the eyes adapts independently to the perturbation. The behavior of the DS of the arm is modulated via the eye-arm coupling function, and the hand DS is modulated via the arm-hand coupling. Such modulation ensures that the learned profile of eye-arm-hand coordination will be preserved, and that the hand will re-open as the object is perturbed away from it, see Figure 3.5. Besides the anthropomorphic profile of visuomotor coordination (Figure 3.8), the gaze-arm lag allows for enough time to foveate at the object, to re-estimate object's pose and to compute suitable grasp configuration for the hand before it approaches too close to the object.

In setups where the arrangement of the obstacle and target differs to a moderate extent compared the setup used in the human demonstrations, the robot successfully grasps the target object, in both obstacle avoidance and no-obstacle tasks, as shown in the experiments presented in the paper and in the accompanying online video. Scene setups that are significantly different, often imply a substantially different approach of the hand to the target object than the one seen in the demonstrations. In our case, this occasionally results either in collision of the fingers with the object prior to grasping or incomplete closure of the fingers on the target object. This is not due to our gaze-arm-hand controller, but rather is due to the fact that we rely on a predefined set of the final hand configurations obtained from human trials. With moderate changes to how the hand approaches an object with complex geometry, like the champagne glass in our experiment, the set of stable hand configurations sometimes can change significantly. In order to increase the rate of grasping in scenarios that substantially differ from the setup in the demonstrations, we would need to use one of the robotic grasp synthesis algorithms to generate the final hand configuration (Sahbani et al., 2012).

The experiments presented here, with several additional experiments, are available online at `http://lasa.epfl.ch/videos/downloads/LukicBiologicalCybernetics2012.mp4`.

3.4 Summary and discussion

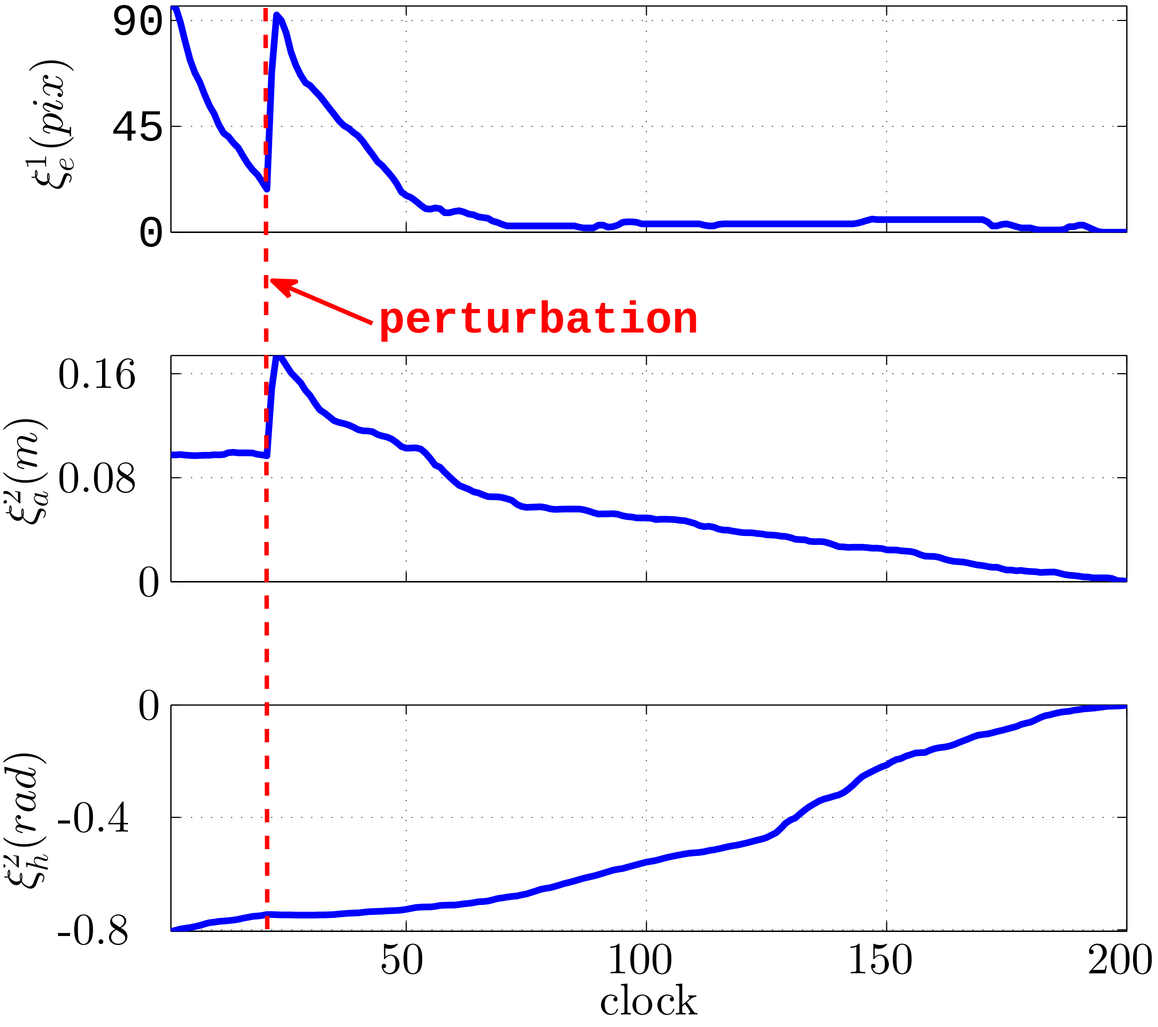


Figure 3.7: Visuomotor adaptation to perturbation during the task, generated by a sudden displacement of the target object. The upper part of the graph shows how the eye state variable, represented by ξ_e^1, adapts to perturbation. The middle graph part of the graph shows the arm state variable denoted by ξ_a^2, and the lower part shows the hand state variable ξ_h^2. Gaze DS adapts independently to spatio-temporal perturbations, whereas DS guiding the arm motion is modulated via the coupling function $\Psi_e(\xi_e)$, and the arm motion modulates hand DS via $\Psi_a(\xi_a)$. The figure shows that all three systems successfully reach the target when perturbed.

In order to design a robotic model for coupled control of the gaze-arm-hand systems, we used the findings and the data from the human study that was presented in Chapter 2. A stable model of the high-dimensional visuomotor coordination was learned by using only several human demonstrations, making it a very efficient, fast and intuitive way to estimate parameters of a robot visuomotor controller. The generalization abilities of the CDS framework (Shukla and Billard, 2011) ensure

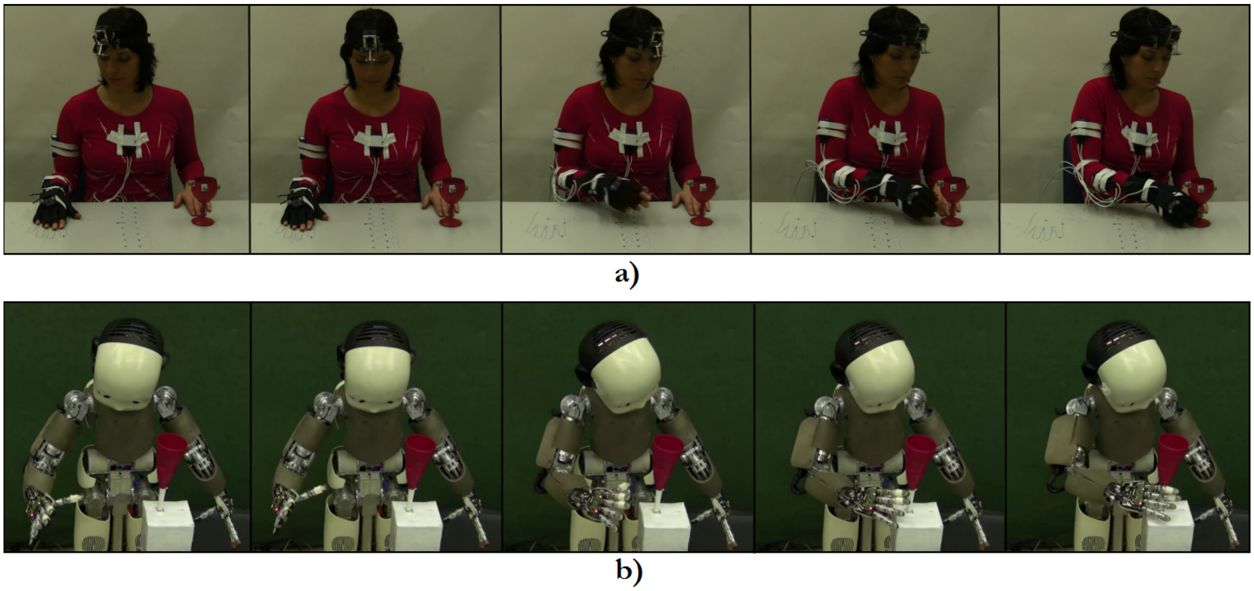


Figure 3.8: A comparison of human visuomotor coordination and visuomotor behavior of the real robot. The visuomotor coordination profile the robot produces (b) is highly similar to the pattern of coordination that was observed in the human trials (a). The figures from left to right show snapshots of the execution of eye-arm-hand coordination from the start of the task (left) until successful grasp completion (right).

the coordinated behavior of the visuomotor controller, even when the motion is abruptly perturbed outside the region of the provided human demonstrations. Similarly to visual servoing (Espiau et al., 1992; Mansard et al., 2006; Natale et al., 2007; Chaumette and Hutchinson, 2008), it performs a closed-loop control, hence it ensures that the target can be reached under perturbations. Coupling profiles for eye-arm and arm-hand systems can be modulated, thus allowing us to adjust the behavior of each slave system with respect to control signals flowing from the corresponding master system. Our eye-arm-hand controller drives the arm-hand motion in synchronization with the gaze and the arm motion. This provides a means to build a compact model of the visuomotor coordination, in a biologically inspired manner, without pre-programming the hand control policy. The major building blocks that constitute the architecture of our controller are the gaze DS, the arm DS and the hand DS. These blocks are coordinated by using the gaze-arm and the arm-hand coupling functions. Each coupling function transfers the information about the state of a master controller to signals that modulate the behavior of a slave controller. The gaze controller is the master controller of the arm, and the arm controller is the master of the hand. This control architecture is supported by the existing evidence of gaze leading the arm motion (Abrams et al., 1990; Johansson et al., 2001; Hayhoe et al., 2003) and the existing reports on coupling between the transport and the grip component in the studies of prehensile movements (Haggard and Wing, 1991, 1995).

Based on the findings from our human study, we then extended the CDS framework for visuomotor coordination on obstacle avoidance such that the task is executed in two segments: from the start to the obstacle and from the obstacle to the target. In our obstacle avoidance mechanism, the gaze is as a constituting element of the overall visuomotor mechanism, and it is actively controlled and intermingled with manipulation requirements and plans, as corroborated in the human study. During obstacle avoidance, the primary modulation of the arm is controlled in the operational space, which, together with controlled hand preshape, ensures that the end-effector avoids the obstacle. The rest postures suitable for obstacle avoidance are provided to the IK solver. We learned these rest postures from the data gathered when the subjects avoided the obstacle in reach-for-grasping.

It is important to mention that our obstacle avoidance scheme does not have the full strength of methods such as Rapidly-Exploring Random Trees (RRTs) (Kuffner Jr and LaValle, 2000) for reaching in very complex workspaces, but it endows the visuomotor system with instant reactions to perturbations, thus providing a means for the rapid handling of a relatively simple obstacle in the workspace.

Limitations

In spite of the human-like behavior the model can produce, which is also useful for robotic visuomotor control, the controller faces a number of limitations. In the controller, we programmed gaze movements in retinal coordinates, whereas solving for the eye-neck joints was outsourced to an external gaze inverse kinematics (IK) optimization solver (Pattacini, 2011). This approach required visual feedback during saccades, which is not biologically plausible and sometimes not convenient to have in a robotic system, due to occasional failures in camera drivers that can cause the loss of visual input. The IK solver demands the exact mathematical model of the gaze kinematic chain, which is sometimes difficult to obtain in a real robot, due to kinematic imperfections. Additionally, the demand for the exact mathematical model of the kinematic chain is somewhat counterintuitive when we consider the well-known kinematic plasticity of the gaze motor system (Desmurget et al., 1998b; Robinson and Fuchs, 2001; Xu-Wilson et al., 2009).

Next, in this model we used Cartesian representation for programming arm motor commands. This representation is consistent with reference frames reported to be used when humans and primates perform arm movements during highly constrained tasks such as obstacle avoidance (Desmurget et al., 1998a). Cartesian motor programming requires the simultaneous computation (in the loop) of a desired Cartesian position and solving an optimization problem to compute inverse kinematics for the purpose of transforming the desired Cartesian state to a set of joint angles of the redundant arm. The evidence from the human and monkey studies suggests that this transformation can be adapted with respect to the changed sensorimotor mapping (e.g., this change could be induced by using prism goggles) (Clower et al., 1996; Andersen and Buneo, 2003; Kurata and Hoshi, 1999; Meeker et al., 2002). Such an adaptation is not possible to accomplish with the IK-solver that requires the exact predetermined kinematics, as the one we used in our controller (Pattacini, 2011).

Additionally, evidence both from behavioral experiments and single neuron recordings suggests that proximal limb movements in unobstructed prehensile movements are programmed in joint coordinates (Shadmehr and Mussa-Ivaldi, 1994; Desmurget et al., 1998a; Kakei et al., 1999). Programming arm movements in joint coordinates could offer some practical computational benefits over programming in Cartesian coordinates. The inverse kinematic map can be computed only once, at the beginning of the movement, to obtain the goal configuration, and later only if we detect that the target object is spatially perturbed.

Finally, in our model the online coupling between the gaze-arm movements is based on retinal coordinates, as the state of the gaze controller, and Cartesian coordinates, in which the state of the arm system is represented. While this scheme is able to provide the human-like coordination pattern on the robot, its biological plausibility is questionable. The studies of Vercher et al. (Gauthier et al., 1988; Vercher and Gauthier, 1988; Lazzari et al., 1997) suggest that this coupling is most likely implemented on the interchange of the proprioceptive information between the gaze and arm.

In the next chapter, we will address the aforementioned limitations by studying the neuroscientific principles in visuomotor coordination in humans and monkeys. The neuroscientific principles serve as the basis for a number of improvements of the controller.

4 Improvements of the Robotic Visuomotor Controller Based on the Lessons from Neuroscience

In this chapter, we redesign the controller presented in Chapter 3 to address a number of its limitations. The controller is solely developed by considering the results of human behavioral studies, including our study with humans presented in Chapter 2. In order to further introduce some improvements to the controller, we here focus on the evidence from neuroscience that is obtained from neurophysiological, lesion and imaging studies in humans and non-human primates.

Namely, we first review the principles behind the interaction between the cortex and the cerebellum, the role of the cerebellum in computing multi-joint limb movements and coupling movements between the effectors. We stress the important aspect of the exchange of motor states between the cortex and the cerebellum, and how the cerebellum uses this information for synchronous motor control of the eyes, head, arm and hand.

Furthermore, we complement our theoretical work with a functional, computational model implemented in a humanoid robot. From our investigation of the neuroscientific literature, we extract a number of computational properties and the organizational structure of the primate visuomotor system on which we ground several improvements that we bring to our model presented in Chapter 3.

More specifically, we revise the gaze control block such that the target remains selected in the retinal coordinates. A learned inverse model based on the algorithm presented in Damas and Santos-Victor (2013) is used to provide the goal eye-neck joint configuration. Once the desired eye-neck joint set is computed by querying the learned model, the gaze system is driven by using the internal feedback loop consisting of a dynamical system (DS) that iteratively evolves the gaze toward the desired joint configuration and takes into account the efference copy of joint motor commands (Quaia et al., 1999; Optican, 2005). This allows us to generate eye-head saccades without visual feedback during saccades. Generating saccades without visual feedback is both biologically plausible and useful for robotic active vision, because visual feedback introduces time delays due to visual processing, and it is sometimes unavailable due to occasional issues with camera drivers. However, if visual feedback and the eye-neck proprioceptive readings are available, the gaze control could be easily switched to the mode of operation with visual and proprioceptive feedback signals.

For the arm control, we take into account the principle of programming arm movements in joint coordinates, while the benefits of instant motor re-programming and coupled motor control are retained from the previous version of the controller. We model this gaze-arm transformation of target encoding by taking inspiration from the transformation that goes from the gaze centered representation of arm reaching targets in the posterior parietal cortex (PPC) to the representation in the arm joint angles in the premotor cortex (PM) and primary motor cortex (M1). Similar to the changes we bring to the gaze controller, computing the arm joints by using the inference from the learned model (Damas and Santos-Victor, 2013) is more efficient than by using an iterative inverse kinematics optimization solver.

Finally, the modified gaze-arm coupling, now based on the transformation of the proprioceptive information from the eye-neck joints of the gaze system to the arm state in joint coordinates, brings better biological plausibility to our controller (Gauthier et al., 1988; Vercher and Gauthier, 1988; Lazzari et al., 1997).

In this work, we aim to contribute to both robotics and systems neuroscience by proposing a functional framework that integrates the cortical reference frames and the cerebellar coupled control, which have been considered so far as mostly independent research problems in both areas. This framework appears to unify a number of independent experimental observations from the primate visual neuroscience. The properties of the proposed computational model offer, within a compact framework, several attractive benefits for visually-driven manipulation in humanoid robots. We show that this controller is capable of reproducing several experimental results from monkey and human studies, namely, the saccade adaptation in target-jump tasks and the profile of decoupled arm-hand movements similar to cerebellar patients. Additionally, we propose a novel behavioral experiment that can either confirm or refute our model.

In the next section, we provide a short review of the state of the art in neuroscience regarding the investigation of the visuomotor principles, tackle a number of well-known models and outline the missing pieces we aim to fill with this work.

4.1 Background research

4.1.1 Neuroscientific models of gaze control and visuomotor coordination

In this section, we first summarize the main focuses of research in neuroscience of human and monkey visuomotor control. We then focus on several well-known neuroscientific models concerning the gaze control, hand control and gaze-arm coupling.

Considering the work on neural structures involved in visuomotor control, two complementary streams of research appeared. The first stream of research has been focused on investigating cortical structures such as the posterior parietal cortex, the premotor and the motor cortices, including the superior colliculus, a subcortical structure, and studying their role in reference frame transformations (Rizzolatti et al., 1997; Goodale and Haffenden, 1998; Batista et al., 1999; Andersen and Buneo, 2003; Crawford et al., 2004; Andersen and Cui, 2009; Beurze et al., 2010; Crawford et al., 2011; Goodale, 2011). The second major stream of research interest has been focused on the cerebellar plasticity, modeling its role in compensating delays in the motor loop, movement generation and synchronization of limb movements (Thach et al., 1992; Wolpert et al., 1998; Thach, 1998b; Kawato, 1999; Wolpert and Ghahramani, 2000; Wolpert et al., 2001; Miall and Reckess, 2002; Ohyama et al., 2003). In systems neuroscience, not many attempts have been made to propose computational functional models of how the cortex and the cerebellum interact in the context of visually driven prehension (Castiello, 2005; Castiello and Begliomini, 2008; Middleton and Strick, 2000).

For gaze control, Optican and coauthors proposed a set of models of the cerebellar interaction with the superior colliculus and frontal eye fields in saccadic eye movements (Lefèvre et al., 1998; Quaia et al., 1999; Optican, 2005). Their model, well-grounded in the neurophysiological evidence, represents a very detailed schematic of the interaction between the cerebellum, superior colliculus and brainstem nuclei for driving and stabilizing the eye movements. The most prominent feature of their modeling is the role of the cerebellum (namely, the oculomotor vermis and the caudal fastigial nucleus) as the key element in the local feedback loop that monitors the efference copy of the gaze commands and, based on it, adaptively steers the saccade to the target end-position. On the other hand, in their model, the superior colliculus and the cortical areas (frontal eye field (FEF), lateral intraparietal area (LIP)) are responsible to determine the desired target in the retinal encoding. Although this model is probably the most detailed and most prominent model of saccade generation, it has a number of shortcomings when transferred to our problem. The model is solely concerned with head-fixed, 2D saccades. The architecture of the model does not include the interaction with the reach and grasp components, similar to the majority of the other saccade models. Additionally, their mathematical model is defined by a number of hand-preset parameters, it is, therefore, difficult to learn and apply the model to different setups.

Furthermore, for modeling visually-driven grasping, the majority of the work has been focused on modeling the interaction between the anterior intraparietal area (AIP) and the premotor cortex (PM) (Rizzolatti and Luppino, 2001; Fagg and Arbib, 1998). In this modeling, the three classes of neurons (visual, mixed visual and motor and motor neurons) in the AIP transform visual representation of the object to be grasped, over an intermediate visuomotor representation, to a motor representation suitable to control the hand (Sakata et al., 1995; Fagg and Arbib, 1998; Murata et al., 2000). The visual features of graspable objects that are initially encoded in the AIP are the size, shape and orientation (Sakata et al., 1995; Murata et al., 2000). The hand motor configuration computed in the AIP is projected to the PMd and PMv (Rizzolatti et al., 1997; Luppino et al., 1999; Castiello and Begliomini, 2008), where the finer elaboration of motor actions is devised (Castiello, 2005; Culham et al., 2006; Olivier et al., 2007; Castiello and Begliomini, 2008). The PMv provides finer selection and segmentation of grip actions based on affordances provided by the AIP and this information is further transferred to the PMd (Rizzolatti et al., 1988), which has the role of keeping, monitoring and visually updating memory representation of hand motor configurations for grasping (Raos et al., 2004; Castiello, 2005; Castiello and Begliomini, 2008). The PMd motor-related temporally segmented information is transferred to the M1, which is involved in issuing low-level control commands for performing precise, independent finger movements (Lang and Schieber, 2003; Castiello and Begliomini, 2008). Yet, this line of modeling misses to include the cerebellum for hand control, which is the important element involved in the computation of synergistic finger movements (Jueptner et al., 1997a,b) and coupling the grasp with the reach component (Rand et al., 2000; Zackowski et al., 2002).

Finally, for the interaction between the visual control system and the arm, Vercher and coauthors proposed a series of models based on their monkey and human studies regarding the interaction between the smooth pursuit and arm motor system in tracking tasks (Gauthier et al., 1988; Vercher and Gauthier, 1988; Lazzari et al., 1997). In their high-level conceptual model (Gauthier et al., 1988), they stressed the important aspect of the interchange of the proprioceptive information between the smooth pursuit and the arm system, and proceeded with building the computational model (Lazzari et al., 1997), which can faithfully replicate a number of interesting observations from behavioral experiments. The visuomotor coupling block of their model corresponds to the cerebellum, namely, it models the high level interaction between the flocculus, responsible for smooth pursuit eye movements, and the dentate nucleus, responsible for eye-arm coupling and arm motor control. However, this model is limited to producing 2D eye movements and planar arm movements, which obviously represents a hard constraint for representing the complex coordination between head-free eye-head saccadic movements and unrestricted, three-dimensional arm movements. Furthermore, the hand control and arm-hand coupling are not included in their model.

Interestingly, in the context of the full eye-head-arm-hand coordination, to the best of our knowledge, there is no such model yet, even at the functional level of abstraction. In this chapter, we aim to fill this gap by proposing both theoretical, schematic model, and its computational implementation in the robot. The computational implementation of this model is expected to bring a number of practical improvements over our robotic controller presented in the previous chapter.

4.2 Schematic model of the central nervous system for visuomotor control

In this section, we present a schematic model of the elements of the central nervous system (CNS) that are involved in visuomotor target encoding and coordinated visuomotor control. Before proceeding with further reading, the reader should note that, in our modeling, we jointly take into account the results obtained from monkey and human studies. Most of the neurophysiological data reported in the literature were obtained from monkeys. We include results obtained from humans, whenever applicable. The intermixing of the presented results is not problematic because the visuomotor coordination principles and their anatomical substrates in humans and monkeys are regarded as highly similar. For example, the eye-head saccade system of monkeys is very similar to the one in humans (Desmurget et al., 2000; Saeb et al., 2011). The anatomical and functional homologues of the superior colliculus and lateral intraparietal area, parietal reach region and anterior intraparietal area, the areas involved in eye movements, reaching and grasping, respectively, are well established in both lesion studies and brain imaging, as reviewed in (Andersen and Buneo, 2002; Castiello, 2005; Culham et al., 2006; Castiello and Begliomini, 2008; Vesia and Crawford, 2012). Similarly, the organization of the visuomotor coordination, and the nature of the cerebellar contribution to it, appear to be very similar between the species (Gauthier et al., 1988; Vercher and Gauthier, 1988). Some subtle differences that arise, for example, from different values of mechanical parameters of the gaze system (Saeb et al., 2011), or from the differences in the time course of adaptation of reactive saccades (Desmurget et al., 2000), are not an issue at the level of abstraction we take in our modeling. Figure 4.1 presents the most relevant brain areas involved in gaze-arm-hand target encoding and coordinated motor control.

4.2.1 Cortical reference frames for target encoding

Our modeling starts with the well-known hypothesis that motor commands are programmed in egocentric coordinates (i.e. in coordinates relative to some parts of the body. This seems to be the default and fundamental characteristics of the vision-for-action system (Goodale and Haffenden, 1998; Crawford et al., 2004; Goodale, 2011; Crawford et al., 2011). There is ample evidence that initial targets for gaze movements are encoded in relative retinal coordinates, and unconstrained arm and hand movements are encoded in relative joint coordinates. We next discuss in more detail how we use this information in our modeling.

Reference frames for target encoding in gaze control

A number of cortical areas are involved in selecting targets for gaze control (Figure 4.1(a)): the lateral intraparietal area (LIP), frontal eye fields (FEF), supplementary eye fields (SEF) and the superior colliculus (SC) (Andersen and Buneo, 2003; Krauzlis, 2005; Culham et al., 2006; Constantin et al., 2007). These areas are strongly interconnected and constitute a distributed network devoted to generating saccadic eye movements (Blatt et al., 1990; Andersen et al., 1990; Matelli and Luppino, 2001; Sparks et al., 2001; Paré et al., 2001; Ferraina et al., 2002; Andersen and Buneo, 2003). In our model, the targets that trigger eye movements are encoded in relative retinotopic coordinates, i.e. the distance vector between the retinal target projection and the location of the fovea, as reported in LIP (Colby and Duhamel, 1996; Andersen and Buneo, 2002; Constantin et al., 2007), SC (Freedman and Sparks, 1997; Krauzlis et al., 2000; Klier et al., 2001, 2003b; Bergeron et al., 2003; Constantin et al., 2004; Krauzlis, 2005; DeSouza et al., 2011), FEF (Dassonville et al., 1992; Russo and Bruce, 1993; Tu and Keating, 2000; Constantin et al., 2007; Monteon et al., 2013) and SEF (Russo and

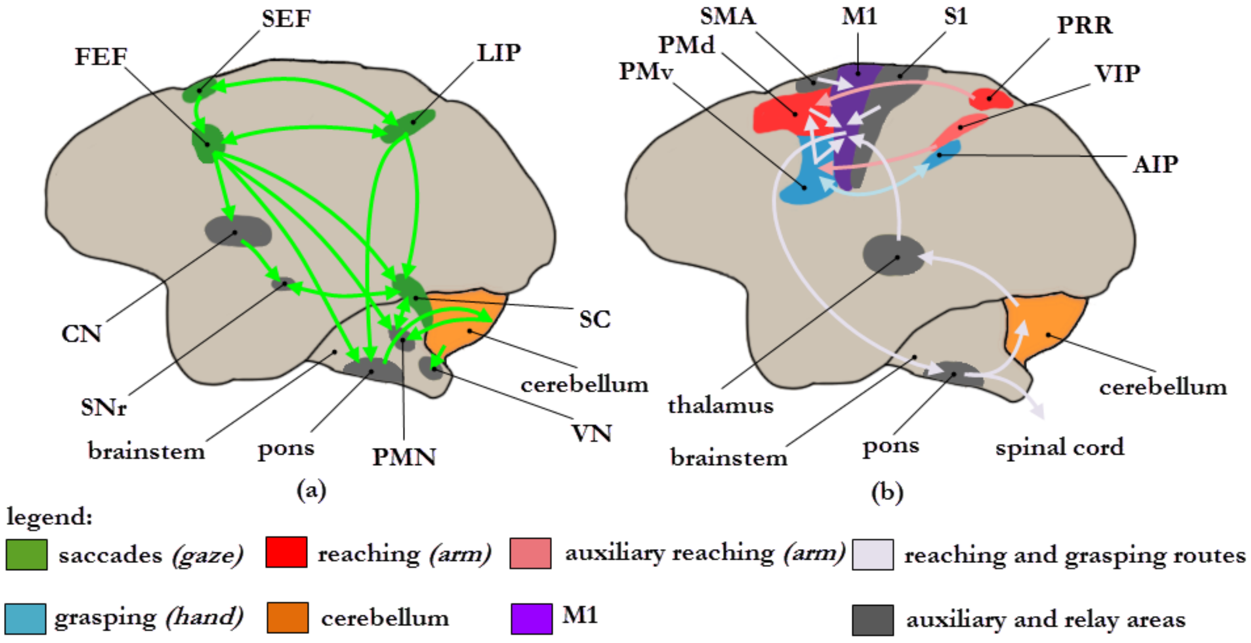


Figure 4.1: Outline of the primary anatomical regions and pathways for visuomotor control in the macaque CNS. For clarity of graphical representation of the corresponding areas and signal routes between them, we separately present: (a) neural circuitry for saccades (*eyes*) and (b) neural circuitry for reaching and grasping (*arm and hand*). For the same reason, the visual cortex, the extrastriate visual cortical areas and some additional areas that are involved in the higher aspects of visuomotor control (e.g. the inferior temporal cortex and the prefrontal cortex) are not presented here as well. List of abbreviations: AIP: anterior intraparietal area; VIP: ventral intraparietal area; LIP: lateral intraparietal area; PRR: parietal reach region; S1: primary somatosensory cortex, M1: primary motor cortex; SMA: supplementary motor area; PMd: dorsal premotor cortex; PMv: ventral premotor cortex; SEF: supplementary eye fields; FEF: frontal eye fields; CN: caudate nucleus of the basal ganglia; SNr: substantia nigra pars reticulate; SC: superior colliculus; PMN: brainstem premotor nuclei; VN: vestibular nuclei. The figures are adapted from (Kandel et al., 2000; Rizzolatti and Luppino, 2001; Krauzlis et al., 2004; Krauzlis, 2005; Vesia and Crawford, 2012)

Bruce, 1996; Russo et al., 2000; Martinez-Trujillo et al., 2004; Constantin et al., 2007)[1].

The LIP, SEF, FEF and SC, encode the saccadic targets in relative retinal coordinates (Krauzlis, 2005), whereas the conversion from the retinal commands to eye and head joint movements is implemented in the downstream structures, where the cerebellum takes the predominant role (Klier et al., 2003b,a; Crawford et al., 2011). For example, in patients with specific cerebellar lesions, the Listing's law for eye movements does not hold, which suggests that the retinal coordinates to joint angle conversion occurs in the cerebellum. The initial target encoding in retinal coordinates (performed in the LIP-SEF-FEF-SC network), and transformation of these coordinates to a set of goal eye-neck joint angles to drive gaze movements by the internal feedback loop (i.e. internal model; done by the cerebellum and the other regions of the brainstem) is the feature we implement

[1] Interestingly, neural recordings and electrical stimulation of the SEF have revealed that this area uses multiple reference frames, including retinal, head-centered and space-centered coordinates for encoding saccadic targets.

in our model, see Section 4.2.2 for more on the cerebellar contribution to gaze control.

Reference frames for target encoding in arm control

In our model, in the starting stage of visuomotor transformations, arm reaching targets are encoded in gaze-centered coordinates, according to the evidence of such encoding in the parietal reach region (PRR) of the superior parietal lobule (SPL) (Andersen et al., 1985; Batista et al., 1999; Buneo et al., 2002; Medendorp et al., 2003; Buneo and Andersen, 2006; Bhattacharyya et al., 2009; Crawford et al., 2011). The PRR is involved in encoding spatial targets for reaching, not for issuing motor commands per se, which supports the view that the primary role of the PRR in target selection and in sensorimotor transformations for target representation (Fernandez-Ruiz et al., 2007; Crawford et al., 2011).

Gaze centered encoding in the SPL projects to the dorsal premotor cortex (PMd) (Kurata, 1991; Johnson et al., 1996; Galletti et al., 2003), and via the PMd to the primary motor cortex (M1) (Johnson et al., 1996; Lacquaniti and Caminiti, 1998). The PMd, PMv and M1 are found to be strongly active during visually guided reaching and pointing movements (Sasaki and Gemba, 1986; Georgopoulos et al., 1988; Kettner et al., 1988; Schwartz et al., 1988; Caminiti et al., 1991; Fogassi et al., 1992; Kurata and Hoffman, 1994). Evidence that the flow of information from the posterior parietal to the frontal areas is mostly involved in sensorimotor transformations, but not in directly issuing motor commands, comes from the studies that have revealed that the PMd is not essential for the direct generation of reaching movements but for encoding of the sensory representation about the target location (Johnson et al., 1996). Along this sequence of sensorimotor transformations, Kakei and coauthors in a series of their single-cell recording experiments found a spatial transition between neurons in the PM to the M1 shows a gradual shift in coding from predominately spatial encoding to a primary pattern of movement encoding in joint/muscle activations, respectively (Kakei et al., 1999, 2001, 2003). Furthermore, sensorimotor-related neural activations on average occur earlier in the PM than in the M1, which comes in support of the hypothesis of the sequential reference frame transformation model directed from the parietal to the frontal areas (Kakei et al., 2001). In their study, Beurze et al. (2010) found a similar, gradual transition from gaze-centered encoding in the PPC to body-centered, joint-based coordinates in the M1[2]. Motivated by the existing evidence, we represent the final target representation of the arm target in arm joint coordinates.

Additional evidence about the reference frames used for programming arm movements comes from the analysis of the kinematic measures of arm movements in behavioral studies. Similar to eye movements, arm reaching movements are believed to be programmed in relative joint coordinates, as observed in behavioral studies (Desmurget et al., 1998a; Crawford et al., 2004; Buneo and Andersen, 2006; Blohm et al., 2008). The behavioral study of Soechting and Lacquaniti (1981) has shown the invariant pattern of covariation between the shoulder and the elbow joints during movements

[2] In the final stage of sensorimotor transformations, the regions of the M1 and PMd specialized in reaching are found to project to the spinal cord (He et al., 1993; Johnson et al., 1996; Scott, 2003).

of the arm. On the other hand, the pattern of spatial arm trajectories has shown substantial variability when compared with the almost linear relations between joint activations. Soechting and Lacquaniti (1981) interpreted the invariance between arm joints as the evidence that the arm movements are programmed in joint coordinates. These results have been corroborated by a number of subsequent studies (Soechting and Lacquaniti, 1983; Lacquaniti et al., 1986; Rosenbaum et al., 1995; Desmurget and Prablanc, 1997; Osu et al., 1997). Desmurget et al. (1997) found in their study a difference between two major strategies in motor programming of reaching movements. They found that unconstrained reaching movements are planned in joint coordinates. However, their results suggested that highly contained reaching movements, such as obstacle avoidance, are programmed in Cartesian coordinates, as in our robotic model for obstacle avoidance presented in Chapter 3.

Prism adaptation studies show that learning of sensorimotor reference frame transformation for visually guided reaching occurs across the PPC (Clower et al., 1996; Andersen and Buneo, 2003), the PMv (Kurata and Hoshi, 1999) and the PRR (Meeker et al., 2002). Motivated by the results of the prism adaptation studies, the reference frame transformation from gaze centered to arm centered encoding in our model is not rigid, it can be adapted.

Reference frames for hand control

The anterior intraparietal area (AIP) of the IPL is involved in transforming visual, shape based representation of the object to be grasped, over an intermediate visuomotor representation, to a motor representation suitable to control the hand (Sakata et al., 1995; Fagg and Arbib, 1998; Murata et al., 2000). The hand motor configuration computed in the AIP is projected to the PMd and PMv (Rizzolatti et al., 1997; Luppino et al., 1999; Castiello and Begliomini, 2008), where the finer elaboration of motor actions is devised (Castiello, 2005; Culham et al., 2006; Olivier et al., 2007; Castiello and Begliomini, 2008). Both the PMv and PMd are known to be active in visual control of hand movements while grasping (Rizzolatti et al., 1988; Raos et al., 2004). The PMd motor-related temporally segmented information is transferred to the M1, which is concerned with lower-level motor control of grasping (Brochier et al., 2004). The M1 is involved mostly in issuing control commands for performing precise, independent finger movements (Lang and Schieber, 2003; Castiello and Begliomini, 2008), whereas synergistic finger movements are probably computed elsewhere. “Vectorial programming” of the whole-hand movements is most likely computed in the cerebellum and these commands are sent to the M1 for segmentation and low-level control, via the loop between the M1 and the cerebellum (Section 4.2.2). The grasping areas of the CNS are presented in Figure 4.1(b).

4.2.2 Cerebellar dynamical control and motor coupling

Our modeling of the cerebellar contribution to visuomotor control is built around the hypothesis that the cerebellum plays a crucial role in the coupled control of movements of different motor effectors, including guiding and synchronizing visuomotor actions (Miall et al., 2000, 2001). Its extensive network of anatomical connections with a great number of cortical visuomotor structures, including the PPC, M1, LIP-SEF-FEF-SC network and the low-level downstream structures such as the brainstem, aid the role of the cerebellum as a motor coordinator (Stein, 1986; Thach, 1998b,a; Middleton and Strick, 2000). Lesions of the cerebellum induce substantially more dramatic impairments of multi-joint movements compared to motor abilities to perform simple, single-joint movements (Thach et al., 1992; Thach, 1998a; Miall et al., 2001). This suggests that one of the primary roles of the cerebellum is in multi-joint movement coordination, indeed[3]. Furthermore, single cell recordings from the dentate and interpositus nuclei of the cerebellum suggest that the cerebellum is recognized to command motor correction signals on a real time basis in order to adapt to perturbations induced during ongoing movements (Thach et al., 1992; Miall et al., 2001). Among its multiple roles, the cerebellum is also known to have a role as a state predictor, for the purpose of compensating delays in the sensorimotor loop and for estimating consequences of intended tasks (Paulin, 1993; Wolpert et al., 1998; Wolpert and Ghahramani, 2000; Wolpert et al., 2001)[4].

Traditionally, computational models of the cerebellum have been primarily concerned with modeling the role of the cerebellum in compensation of delays in the sensorimotor loops (Miall et al., 1993), predicting contexts based on internal forward models (Wolpert and Ghahramani, 2000; Wolpert et al., 2001), providing computational models for cerebellar motor learning (Kawato and Gomi, 1992a,b) and head-fixed saccade control (Lefèvre et al., 1998). However, few attempts in computational modeling have tackled the involvement of the cerebellum coordinating natural, unrestricted eye-head-arm-hand actions. This is a niche where our neuroscientifically-inspired modeling effort is concentrated.

Cerebellar contribution to gaze control

In our model, saccadic targets are encoded in relative retinal coordinates as presented in the LIP-SEF-FEF-SC network (Section 4.2.1). On the other side of this transformation, low-level structures such as the reticular formation saccade generator of the brainstem already have access to information about the joint angles of the eyes and the neck (Crawford et al., 2011). This indicates that the computation of the saccade command velocity and conversion of the retinal error to eye-head joint rotations must be utilized somewhere between these structures, most likely by the cerebellum and the brainstem. Figure 4.1(a) presents the aforementioned gaze control routes involving the cerebellum. In our model, we take into account the transformation of the retinal error to the desired gaze eye-neck joints that define the end-point fixation. It is well-known that the

[3] The cerebellum is a multivariate motor controller, as expressed by the methodology of control theory.

[4] The cerebellar estimation of consequences of motor actions has inspired us to develop the forward planning mechanism for detection of obstacles presented in Chapter 3.

cerebellum takes a role in long-term motor adaptation of forward and inverse models for saccadic and smooth pursuit eye movements (Desmurget et al., 1998b; Robinson and Fuchs, 2001; Xu-Wilson et al., 2009). The adaptive nature of this mapping has inspired us to introduce the machine learning approach to implementing the retino-motor mapping for gaze control.

In addition to the role of the cerebellum in the conversion from the retinotopic encoding to gaze motor commands, the cerebellum is a major feedback structure for online steering of gaze movements, where a number of cerebellar regions are involved: the ventral paraflocculus (VPF) (for smooth pursuit) and the oculomotor vermis and the underlying caudal fastigial nucleus (for saccades) (Quaia et al., 1999; Lefèvre et al., 1998; Robinson and Fuchs, 2001; Krauzlis, 2005; Optican, 2005). Thus, we model the internal feedback loop that steers the eye-neck system to the end-point fixation as defined by the set of eye-neck angles. The output velocity commands derived from the internal forward model are integrated to command the eye-head posture. The discrepancy between the observed remarkable final accuracy of gaze end-points and the considerable inherent variability in the gaze motor commands could not be due to an open loop controller, which suggests the existence of an internal feedback loop that correct the gaze in flight (Scudder et al., 2002; Chen-Harris et al., 2008; Xu-Wilson et al., 2009). Based on this evidence, a number of subsequent works have suggested that the cerebellum is this internal feedback element that monitors and corrects gaze joint motor commands in an online fashion (Robinson and Fuchs, 2001). Because the proprioceptive and visual feedback is too slow to be used in online control, the cerebellum relies on the efference copy of the motor commands to perform online correction of movements based on the residual motor error (Lefèvre et al., 1998; Quaia et al., 1999; Xu-Wilson et al., 2009).

Cerebellar contribution to arm control and gaze-arm coupling

Following the evidence that the cerebellum has a prominent role in controlling goal-directed arm movements, we include the cerebellar contribution to arm control in our model. From the fast routing inputs from the M1 via the pons and the spinal cord (Thach et al., 1992), the cerebellum can access to the cortical target representation for arm movements (Section 4.2.1). (The connections between the cerebellum, the arm-hand cortical motor areas and the brainstem are outlined in Figure 4.1(b)). Cerebellar patients show kinematic deficits while performing arm movements such as reaching (Becker et al., 1990, 1991; Bastian et al., 1996; Zackowski et al., 2002), pointing (Bonnefoi-Kyriacou et al., 1998) and throwing (Timmann et al., 1999). Cerebellar patients exhibit greater end-point errors in arm reaching, and movements are performed slower compared to healthy subjects (Zackowski et al., 2002), with improper inter-joint coordination (Becker et al., 1991). The magnitudes of angular joint velocities were impaired, and the loss of proper temporal synchronization between shoulder and elbow joints was observed (Becker et al., 1991). Based on the evidence that the cerebellum computes arm joints in a simultaneous manner, in our model arm joint movement commands are computed jointly, in a vectorial fashion. The main feature of arm movements in cerebellar patients is the loss of coordination across many joints involved in the task (Bastian et al., 1996). Cerebellar patients have shown the "one-joint-at-the-time" strategy while reaching, while healthy controls simultaneously controlled arm joints.

Next, the gaze-arm coupling block of our model is based on the hypothesis that the interpositus and dentate nucleus are responsible for such coupling, and that the nature of this coupling is based on the model that stores correlations between the gaze motor error and arm motor error, hence gaze-arm coupling is state-based not time-based. Dysfunctions of the cerebellum produce substantial drops in the performance of coordinated eye and arm movements (Bekkering et al., 1995; Miall et al., 2000, 2001; Miall and Reckess, 2002). In patients with cerebellar ataxia, motor latencies for movement initiation were increased when the patients were performing a step-tracking task with simultaneous engagement of eye and arm movements compared to a task that required individual eye or limb movements (Brown et al., 1993; van Donkelaar and Lee, 1994). Miall et al. (2001) found that the cerebellar activation parametrically increases with the level of required visuomotor coordination in a tracking task. Vercher and Gauthier (1988) have shown that lesions of the dentate nucleus produce uncoupling of eye and arm movements. The input signal about the gaze state used for the gaze-arm coordination is most likely utilized in the form of the efference copy of gaze motor commands (Cotti et al., 2011). The visuomotor coordination between the gaze and the arm is believed to be based on the cerebellar mapping between non-retinal gaze motor errors and motor errors of the arm (Miall et al., 2000). Additional support that the input for gaze-arm coordination is the gaze motor commands and not the retinal error are the results of visuomotor experiments that suggest that the pattern of visuomotor coordination is preserved even in the total darkness (Lazzari et al., 1997).

Cerebellar contribution to hand control and arm-hand coupling

The cerebellum is involved in coupling of arm reaching and hand grasping movements, as well. In our model, based on the evidence from the literature, the interpositus and dentate implement arm-hand coupling. In their brain imaging study, Jueptner and coauthors found significant activations of the cerebellar nuclei during learning and reproducing a set of finger movements (Jueptner et al., 1997a,b).

Regarding the arm-hand coupling, Rand et al. (2000), in their study with cerebellar patients, found that the kinematic measures of the arm and hand systems in cerebellar patients varied significantly compared to healthy control subjects, who exhibited very tight coupling between the arm and the hand kinematic parameters (namely, times of the maximum velocity of the wrist and the maximum grip aperture). Similarly, in the study of Zackowski et al. (2002) with cerebellar patients, individual arm and hand components were affected. Interestingly, in their study, the deficits in coordination of these components were even more striking. The coupling between the components was severely deteriorated compared to healthy controls. The patients frequently dropped the object due to improper synchronization between the reaching and grasping components. The similar loss of synchronization of the hand preshape with respect to goal-directed arm movements together with

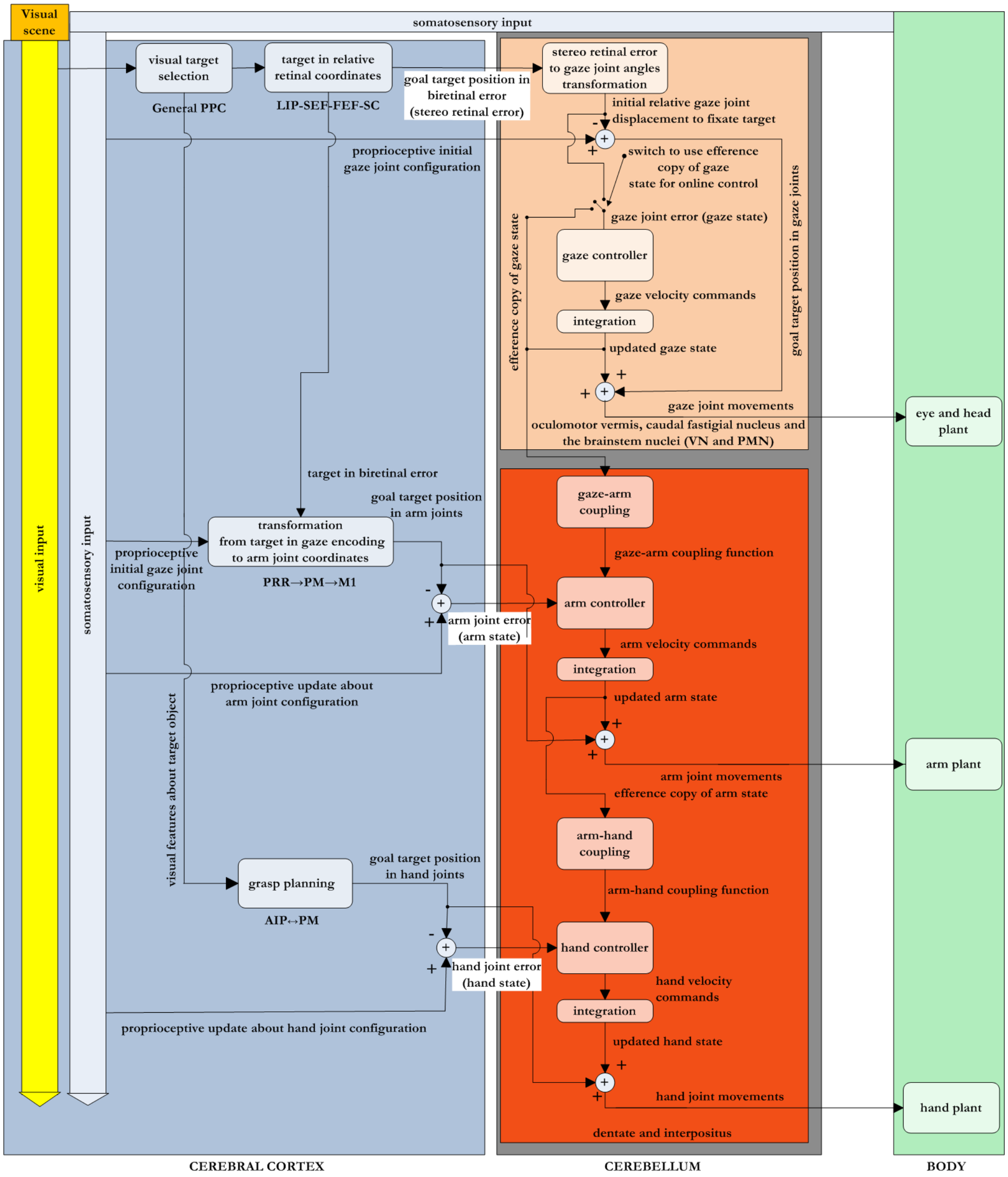

Visual scene
somatosensory input
visual target selection
General PPC
target in relative retinal coordinates
LIP-SEF-FEF-SC
goal target position in biretinal error (stereo retinal error)
stereo retinal error to gaze joint angles transformation
initial relative gaze joint displacement to fixate target
proprioceptive initial gaze joint configuration
switch to use efference copy of gaze state for online control
gaze joint error (gaze state)
gaze controller
gaze velocity commands
integration
updated gaze state
efference copy of gaze state
goal target position in gaze joints
gaze joint movements
oculomotor vermis, caudal fastigial nucleus and the brainstem nuclei (VN and PMN)
eye and head plant
visual input
somatosensory input
target in biretinal error
goal target position in arm joints
transformation from target in gaze encoding to arm joint coordinates
proprioceptive initial gaze joint configuration
PRR→PM→M1
arm joint error (arm state)
gaze-arm coupling
gaze-arm coupling function
arm controller
arm velocity commands
integration
updated arm state
proprioceptive update about arm joint configuration
arm joint movements
efference copy of arm state
arm plant
visual features about target object
arm-hand coupling
arm-hand coupling function
grasp planning
AIP↔PM
goal target position in hand joints
hand controller
hand joint error (hand state)
hand velocity commands
integration
updated hand state
proprioceptive update about hand joint configuration
hand joint movements
dentate and interpositus
hand plant
CEREBRAL CORTEX
CEREBELLUM
BODY

Figure 4.2: Based on our literature investigation, we propose a block scheme of visuomotor coordination. The left part corresponds to the cerebral cortex, with the corresponding blocks relevant for visual target selection, representation of saccade targets, reference frame transformation for arm reaching and grasp planning. The right part of the figure corresponds to the cerebellum and blocks responsible for online control of the gaze, arm and hand and their synchronization. Parts that are not directly relevant to our modeling, such as mid-stop relay stations such as the pons, SNr, etc. are not represented in the diagram in order to simplify the graphical representation. Similarly, ascending output signals from the cerebellum to the M1 are directly represented as arrows carrying motor commands to their corresponding plants.

the increased variability in arm trajectories was observed a throwing task, as well (Timmann et al., 1999). The study of Mason et al. (1998) where they selectively inactivated the dentate and interpositus nuclei by muscimol injections provided further insights on how arm and hand movements are coupled in the cerebellum. They found that the inactivation of the anterior interpositus and adjacent dentate impaired control of grasping, leaving arm reaching mostly intact. On the other hand, inactivation of the posterior interpositus and adjacent dentate affected reaching without affecting grasping. Moreover, the study of Mason et al. (1998) suggested that the connections of the anterior hand regions and the posterior reaching regions contribute to coupling of arm and hand movements.

A series of perturbation studies of Haggard and Wing (Haggard and Wing, 1991, 1995, 1998) has provided solid evidence that the arm and the hand are coupled by the state-based, time-invariant principle of coordination. Considering that the dentate and the interpositus implement this coupling, as revealed by neuroscientific studies, we propose that state-based, time-invariant, coordination is a principle of visuomotor coupling implemented in the dentate and the interpositus. This principle is implemented in our computational model.

Based on our literature investigation, we created a functional schematic model of cortico-cerebellar involvement in reference frame transformations and motor control for visuomotor coordination. This model is presented in Figure 4.2.

4.3 Computational model for visuomotor control

4.3.1 Model premises

We next present a model for visuomotor control that is built upon the investigation presented in the previous section. Our model represents a computational implementation of the schematic illustrated in Figure 4.2. It shares, on a functional level of abstraction, a number of resembling features with the corresponding parts of the primate cortico-cerebellar circuitry involved in visuomotor control. The model incorporates: (a) target encoding transformations and representations for eye, arm and hand control that are reported to be used in the cortex and (b) the coupled control principles that are found in the cerebellum. It should be emphasized that this architecture is neither a detailed model nor an exhaustive model of the CNS. Rather, it is a functional mathematical abstraction that shares a number of similarities with the biological, functional organization and biological computational principles involved in primate visuomotor control. Our primary reason why we focus on a functional model, instead of a very detailed model, is to keep the model detailed enough to be a useful abstraction for systems neuroscience, but still limited in scope to make the model computationally tractable for robotic implementation.

The biological features of our model are reference frames (and their transformations) used for motor control and the organization of the coupled motor control between different effectors. We model the reference frame transformations by using the Infinite Mixture of Linear Experts (IMLE) algorithm (Damas and Santos-Victor, 2013). We devise the cerebellar coupling by using the Coupled Dynamical Systems framework, the same computational approach that has been presented in the previous chapter. It is important to note that both algorithms are not implemented by using the connectionist, more biologically plausible, approach. However, the functionality they provide is particularly useful to mimic some fundamental aspects of the motor control circuitry reported from neuroscience studies in primates. The aimed contribution of this work to systems neuroscience is to propose a functional and a mathematical model of visuomotor coordination that shares a resemblance to the functional high-level organization and the interaction between the cerebellum and the cortex. The features of our model that share a resemblance to the corresponding features of the primate visuomotor circuitry are, namely:

1. Visual targets for gaze commands are selected in retinal coordinates (summarized in Section 4.2.1) whereas gaze commands are programmed multi-joint in eye-neck coordinates (summarized in Sections 4.2.1 and 4.2.2)

2. Arm movements are programmed in joint coordinates (summarized in Section 4.2.1)

3. Hand movements are programmed in joint coordinates (summarized in Section 4.2.1)

4. Eye, arm and hand motor commands are represented in terms of the motor error, i.e. relative coordinates (summarized in Section 4.2.1)

5. Retinal errors are converted to eye and neck joint commands in the cerebellum and the other brainstem nuclei (summarized in Sections 4.2.1 and 4.2.2)

6. The cerebellum monitors the gaze movements by observing the efference copy of gaze motor commands (summarized in Section 4.2.2)

7. Multi-joint motor commands are programmed synchronously (summarized in Section 4.2.2)

8. Gaze and arm motor commands are coupled based on the efference copy derived from the gaze motor commands (summarized in Section 4.2.2)

9. Arm and hand motor commands are coupled based on the efference copy derived from the arm motor commands (summarized in Section 4.2.2)

10. Motor commands and the target representation are not memorized and stored in a long-term fashion; they are updated and computed in real-time (more on this can be found in (Goodale and Haffenden, 1998; Goodale, 2011))

11. Motor coupling is based on the transformation of the motor error between the two effectors (summarized in Section 4.2.2)

12. Motor coupling is time invariant, i.e. it only dependent on the state, not on time (summarized in 4.2.2)

All enlisted properties can be summed up and jointly tackled under the two "umbrella" problems: the problem of reference frame transformations and the problem of coupled motor control. We next proceed with our computational modeling.

4.3.2 Modeling reference frame transformations

The first fundamental problem of visuomotor coordination is a computation of a sequence of transformations from the gaze-centered encoding to the representations suitable to generate arm and hand movements. The direct kinematic problem is defined as computing the output variables (e.g. Cartesian position of the end effector) based on the inputs (e.g. arm joint angles). This problem is a well-defined mapping, suitable for both learning approaches and analytical solutions. On, the other hand, the inverse problem of computing the inputs (arm angles) based on a set of desired outputs (e.g. desired Cartesian position of the end effector) is a far more complicated mapping. For such inverse problems in highly redundant systems, such as the gaze and arm systems in primates and humanoid robots with a high number of degrees of freedom, infinitely many inverse solutions may exist. Until recently, this problem could not be successfully tackled by using learning approaches. The two main benefits of the learning algorithms, as compared to alternative optimization based approaches (Pattacini, 2011) are: (a) no need to have the prior information on the precise kinematic model and (b) although learning can be a time consuming iterative process, inference can usually be solved rapidly, as compared to iterative optimization computations, which makes learning very attractive for the problems we tackle here[5]. The IMLE learning algorithm for multi-valued regression (Damas and Santos-Victor, 2013), recently developed in our laboratory (VISLAB IST), has demonstrated to be successful in simultaneously providing forward and inverse kinematic predictions in highly-redundant systems. We next provide a brief description of this algorithm.

IMLE algorithm

The IMLE algorithm is a probabilistic learning algorithm. It is built on the main assumption that the input-output data mapping can be approximated by a mixture of local linear experts (i.e. local models). The algorithm has the ability to learn multi-valued functions, by associating different linear models to share the same region of the input space. If an input training point is presented as $z_i \in \mathbb{R}^d$ and a corresponding output is $x_i \in \mathbb{R}^D$, then the generative model of the IMLE algorithm is described as follows:

[5] The discussed properties of learning algorithms are the universal properties applicable to the majority of machine learning algorithms, however, some particular machine algorithms can differ in various aspects, including the time needed for learning and inference.

$$\mathcal{P}\left(x_i \mid z_i, w_{ij}; \Theta\right) \sim N(\mu_j + \Lambda_j(z_j - \nu_j), \Psi_j), \tag{4.1}$$

$$\mathcal{P}\left(z_i \mid w_{ij}; \Theta\right) \sim N(\nu_j, \Sigma_j), \tag{4.2}$$

where the mean ν_j and the covariance matrix $\sum_j$ define a Gaussian input region for each expert j. Parameters μ_j, the mean, and the matrix of regression coefficients Λ_j define the linear relation from inputs to outputs of each expert. Ψ_j matrix models the uncorrelated noise at the output dimensions. The latent variable w_{ij} assigns training data points to particular experts. The parameters of the IMLE, jointly represented as Θ, are estimated by using the extended expectation-maximization (EM) procedure.

Once the model parameters are learned, when performing multi-valued inverse predictions, for each query point the IMLE finds a minimal set of predictions by performing post-hoc clustering procedure and statistical hypothesis testing in order to assess the validity of the predictions. The IMLE algorithm has very low computational complexity, which makes it very suitable for both online learning and forward and inverse predictions. For more on this algorithm, please consult the original article (Damas and Santos-Victor, 2013).

Gaze reference frame transformation

In order to provide head-free (the eyes and head) saccadic eye movements, the cerebellum implements a mapping from the retinal target representation to the eye-neck joint angles (Sections 4.2.1 and 4.2.2). To provide saccadic eye movements with vergence, i.e. being able to fixate in stereoscopic depth, this transformation must take into account the biretinal target representation (the retinal error from the left and right image planes) (Tweed, 1997).

We represent the gaze joint angle displacement that provides object end-point fixation as $\triangle q_g \in \mathbb{R}^6$ (defined in the following order: neck pitch, neck roll, neck yaw, eyes tilt, eyes version and eyes vergence angle, respectively) and the biretinal error as $\xi_g \in \mathbb{R}^4$,

$$\xi_g = \begin{bmatrix} p_c - p_{t,right} \\ p_c - p_{t,left} \end{bmatrix}, \tag{4.3}$$

where the position of the fovea is p_c and the position of a visual target in retinal coordinates is $p_{t,i}$, $i = \{left,\, right\}$. Then a direct mapping can be formulated as:

$$\xi_g = f_g(\triangle q_g). \tag{4.4}$$

By randomly moving the fixation target and by using the fixation behavior provided by another gaze module (Pattacini, 2011), we obtain training data points $(\triangle q_{g,i}, \xi_{g,i})$ to train the IMLE algorithm to estimate this mapping. Once the IMLE model is trained, in the run-time after the visual target is segmented and the biretinal error is obtained ξ_g^*, we query the IMLE module for the inverse solutions, i.e. to provide the desired joint displacement that will provide successful target fixation:

$$\widetilde{\triangle q_g} = \widetilde{f_g^{-1}}(\xi_g^*). \tag{4.5}$$

Once the desired displacement is inferred, the goal target position in gaze joints q_t can be obtained as:

$$q_t = q_c - \widetilde{\triangle q_g}, \tag{4.6}$$

where q_c is the vector of eye-neck proprioceptive joint readings from encoders.

Gaze-arm reference frame transformation

The flow of reference frame transformations from the gaze-centered target encoding in the PPC to the encoding of the target position in arm joint coordinates in the PM/M1 provides primates with very successful visually guided reaching abilities (Section 4.2.1). This functionality drives the modeling we present in this section. We formulate this reference frame transformation as follows:

$$x_t = f_{ga}(q_t), \tag{4.7}$$

where q_t represents the fixated target position in the gaze joint reference frame, and x_t is the goal arm joint configuration at the target object (arm joints are defined in the following order: shoulder pitch, shoulder roll, shoulder yaw, elbow, wrist pronation-supination, wrist pitch and wrist yaw, respectively). It is worth to note that this representation constrains natural-looking visually-driven reaching behavior, because it assumes that the gaze first fixates the target in order to provide the reference frame transformation for the arm. After the gaze lands on the target, the movements of the arm can be programmed. This constraint prohibits producing natural looking visuomotor coordination profiles, because it has been consistently observed in psychological studies that the gaze and arm are simultaneously controlled when performing prehensile movements driven by head-free gazing (Johansson et al., 2001; Hayhoe et al., 2003; Lukic et al., 2014b). By using the functionality of the gaze reference frame transformations we presented in the previous section, we can compute the desired reference frame transformations for the arm that allows reaching and grasping a visual target that lies outside the fovea. For the extrafoveal target, we first obtain the biretinal error ξ_g and use Eqs. 5.2 and 5.4 to compute a final set of gaze joints when the target is fixated q_t. Once we have this, we can obtain x_t by using Eq. 5.10.

Similar to the gaze system, the gaze arm-reference frame transformation is learned by using the IMLE. The input-output training point pairs $(q_{t,i}, x_{t,i,})$ are obtained by moving the arm to randomly selected points in the workspace and subsequently fixating the center of the palm.

4.3.3 Modeling coupled motor control

The second fundamental problem of visuomotor control is how to generate movements of the eyes, arm and hand and how to appropriately coordinate the movements of these effectors. This has been addressed in Chapter 3. In this chapter, we use the same approach with some subtle changes, which we will briefly describe here. In our model, the CDS corresponds to the part of the model that amounts for the coupling effects in the cerebellar nuclei.

In our case, the gaze state $\xi_e \in \mathbb{R}^6$ is represented as the distance between the current gaze joint configuration q_c and the goal position of the target in gaze joint coordinates q_t (i.e. gaze joint error), $\xi_e = q_c - q_t$, where q_t is obtained as explained in Section 4.3.2. This is the difference with respect to the previous version of the controller, where the dynamics was encoded in the form of retinal error. The gaze control scheme now implemented in the gaze joint coordinates is a biologically plausible way and it is most likely implemented in the oculomotor vermis and the caudal fastigial nucleus (Sections 4.2.1 and 4.2.2). Similarly, the arm state $\xi_a \in \mathbb{R}^7$ is represented as the distance in joint coordinates between the assumed arm configuration x_c and the goal arm configuration when the target object is reached x_t: $\xi_a = x_c - x_t$. In the previous version of the controller, we used Cartesian encoding for the arm movements. The current, relative encoding of arm movements in arm joint coordinates is a more biologically plausible strategy than in the previous version of the controller, where we used Cartesian encoding for the arm movements. Arm movement generation in joint coordinates is a pattern observed both in neural and behavioral studies (Section 4.2.1). The hand state $\xi_h \in \mathbb{R}^9$ is expressed as the difference between the current hand configuration h_c and the goal hand configuration when the object is grasped h_t: $\xi_h = h_c - h_t$, the same representation we used in Chapter 3.

We first select a target for saccadic eye movements in the retinal coordinates and encode it in the form of the retinal error. The initial retinal error encoding is a biologically plausible strategy orchestrated by the LIP-SEF-FEF-SC network (See Sections 4.2.1 and 4.2.2). Then, we transform the biretinal error in the gaze joint coordinate encoding (Section 4.3.2). The gaze movement velocity vector is computed based on the gaze joint error.

Two stages of gaze control: the first, the conversion of the retinal coordinates to eye and neck joint angle and, the second, gaze programming in joint coordinates is an organization observed in the primate visuomotor control (Section 4.2.2).

In order to control the arm in joint coordinates, we need to have an available representation of the target in this reference frame. For this, we use the steps presented in Section 4.3.2.

Algorithm 2 presents our implementation of the block scheme for reference frame transformations and motor coupling proposed in Figure 4.2 based on our literature investigation.

4.4 Results

```
do
  General:
  – query frames from cameras
  – recognize and segment the target object
  – compute the position of the object in
  – retinal coordinates : $\mathrm{p_{t,i}},\ \mathrm{i} = \left\{\mathrm{left, right}\right\}$
  – biretinal error : $\xi_{\mathrm{q}} = \left[\mathrm{p_{fovea}} - \mathrm{p_{t,right}}; \mathrm{p_{fovea}} - \mathrm{p_{t,left}}\mathrm{right}\right]$
  – read the head joints from hand encoders : $\mathrm{q_c}$
  – read the arm joints from arm encoders : $\mathrm{x_c}$
  – read the hand joints from hand encoders : $\mathrm{h_c}$

 Gaze:
  if gaze is not at target then
    $q_t \leftarrow$ QueryIMLEforEyesAndNeckJoints($\mathrm{xi_q}$)
    $\xi_e \leftarrow q_c - q_t$
    $\dot{\xi}_e \leftarrow \mathbb{E}\left[\mathcal{P}\left(\dot{\xi}_e \mid \xi_e\right)\right]$
    $\xi_e \leftarrow \xi_e + \dot{\xi}_e \Delta t$
    $q_c \leftarrow \xi_e + q_t$
    MoveEyesAndNecktoUpdatedJoints($\mathrm{q_c}$)

  end if

  Eye − arm coupling:
  $\tilde{\xi}_a \leftarrow \mathbb{E}\left[\mathcal{P}\left(\xi_a \mid \Psi_e\left(\xi_e\right)\right)\right]$

  Arm:
  if the arm is not at target then
    if the first pass or the target is perturbed then
        $x_t \leftarrow$ QueryIMLEforGazeArmJointTransform($\mathrm{q_t}$)
    end if
    $\xi_a \leftarrow x_c - x_t$
    $\Delta\xi_a \leftarrow \xi_a - \tilde{\xi}_a$
    $\dot{\xi}_a \leftarrow \mathbb{E}\left[\mathcal{P}\left(\dot{\xi}_a \mid \beta_a \Delta\xi_a\right)\right]$
    $\xi_a \leftarrow \xi_a + \alpha_a \dot{\xi}_a \Delta t$
    $x_c \leftarrow \xi_a + x_t$
    MoveArmToUpdatedJoints($\mathrm{x_c}$)
  end if

  Arm − hand coupling:
  $\tilde{\xi}_h \leftarrow \mathbb{E}\left[\mathcal{P}\left(\xi_h \mid \Psi_a\left(\xi_a\right)\right)\right]$

  Hand:
  $\xi_h \leftarrow h_c - h_t$
  if the hand is not at target then
    $\Delta\xi_h \leftarrow \xi_h - \tilde{\xi}_h$
    $\dot{\xi}_h \leftarrow \mathbb{E}\left[\mathcal{P}\left(\dot{\xi}_h \mid \beta_h \Delta\xi_h\right)\right]$
    $\xi_h \leftarrow \xi_h + \alpha_h \dot{\xi}_h \Delta t$
    $h_c \leftarrow \xi_h + h_t$
    MoveHandToUpdatedJoints($\mathrm{h_c}$)
  end if
until object grasped
```

Algorithm 2: Improved algorithm for eye-arm-hand coordination

4.4.1 Improvements with respect to the previous controller

While our new controller is able to reproduce the same task of visually guided obstacle-free reaching and grasping similar to humans in our motion capture trials and as the controller presented in Chapter 3, there are several important differences between this new controller and the previous controller. Figure 4.3 shows several snapshots of learning and testing of the gaze controller and the arm controller, respectively.

We re-designed the arm controller to encode the motion in joint coordinates instead of using the Cartesian coordinates, and the gaze controller to encode gaze commands in gaze joints instead of retinal error. The second difference is that we use the full state coupling instead of the norm-coupling functions for the gaze-arm and arm-hand coupling. Both of these changes increased the dimensions of the joint probability density functions that model the corresponding dynamical systems of the gaze and arm and coupling blocks. Hence, in theory, these changes should increase the computational complexity of the problem. However, the time of computation of all 5 blocks of the CDS, as for the previous version, remains under 1 ms[6], while we gain several advantages.

For gaze control, interestingly, both gaze IMLE inference and gaze IK controller have comparable computational complexity, with the computation time under 1 ms. Hence, for the gaze controller, introducing the IMLE does not significantly change the computational time, as in the arm's case. Nevertheless, the benefit of being able to adapt this mapping is the improvement over the previous version of the controller with the IK solver that required the mathematical model of the gaze system kinematics. The second advantage for the gaze control, once we changed encoding of the gaze controller, is that we are able to command gaze movements in a visual-open loop manner, which is attractive in terms of the computational efficiency (no need to rescan the stereo images after each integration step if the target is static) and more convenient, as well (if the camera drivers occasionally fail, this situation is not problematic, because there is no requirement to segment the images after each integration pass, as in the previous controller).

The advantage for arm control is that, by directly programming arm movements in joint coordinates, we avoid the computation of IK in each pass of the control loop, the task that requires on average 25 ms (this time can be up to 40 ms). For computation of the target encoding in arm joints, the gaze-arm joint mapping is only computed at the beginning of reach-to-grasp movements and when the visual system detects that the object to be grasped is perturbed. The time to compute the gaze-arm joint mapping by using the IMLE is under 1 ms. For most of the real-world tasks, even in dynamic, unpredicted scenarios when perturbations normally happen, the proportion of the time when the object is steady is usually significantly greater than the time during which it is being perturbed. In other words, most of the objects to be grasped are more in a steady state than they are perturbed in the workspace. We exploit this premise to gain the computational efficiency.

[6] The CDS code is implemented in C++ and the presented tests are run on a computer with an Intel i7 2.7 GHz dual-core processor and 4 GB of RAM. All reported times are averages calculated for 200 passes through the control loop.

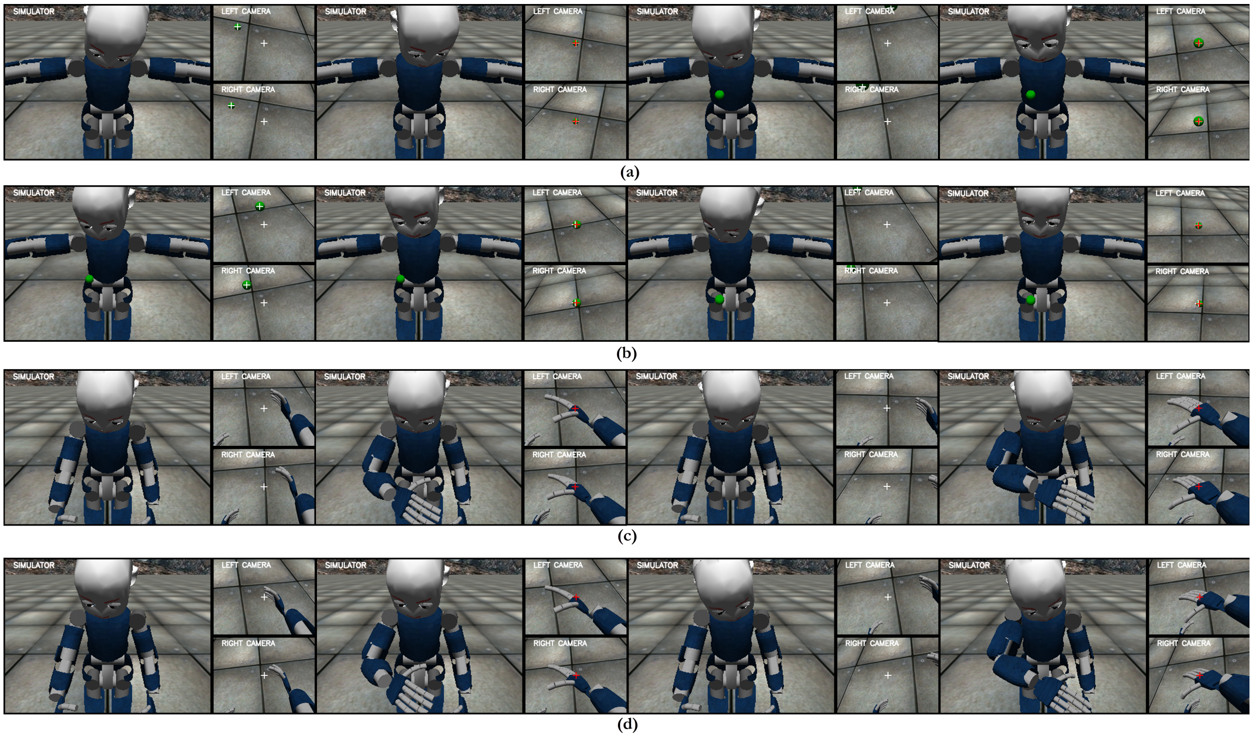


Figure 4.3: Biretinal-gaze joint mapping learning (a) and performing (b), and gaze-arm mapping learning (c) and performing (d), respectively. The figures in each row from left to right show snapshots of the execution two fixation saccades (biretinal-gaze mapping) and the execution of two visually-driven reaches (gaze-arm mapping), respectively. The figures are ordered to correspond to before and after fixations snapshots from the simulator. The first row (a) corresponds to saccades used to train the gaze mapping with the IMLE. The second row (b) shows the performance of the controller once the IMLE and gaze DS are learned. The IMLE is used to obtain the inverse mapping and the gaze DS, an internal feedback element, steers the saccade to the end point. Similar to the gaze rows, row (c) corresponds to visually guided reaches by using babbling to train the gaze-arm mapping with the IMLE and row (d) shows the performance of the learned model.

The second benefit is the use of the full coupling between the dynamical systems, which yields better visuomotor coordination in practice. Consider, for example, the arm DS, which is now a 7 dimensional mapping instead of the previously used 3-dimensional Cartesian representation. If we conditioned on the norm function of the gaze state, it would be difficult to reliably infer a 7 dimensional vector of desired arm joints based a scalar value. With the use of the full state coupling, this mapping becomes robust because it is conditioned on more complete information. Similar rationale applies to the choice of the arm-hand coupling as well.

In the next sections, we further validate our model by replicating two mechanisms widely observed in human and monkey experiments: the decoupling of the reach and grasp components after cerebellar lesions, and saccade adaptation task. In addition, we propose a novel behavioral study.

4.4.2 Replicating the observations from Rand et al. study

In order to further validate our model, we attempt to replicate some effects observed in the human studies. Rand et al. (2000), in their study with a group of cerebellar patients, found that the most striking difference between the cerebellar patients and healthy controls was observed in the coupling between the reach and grasp components. While healthy control subjects exhibited very tight coupling between the arm and hand kinematic parameters, the arm-hand coupling in cerebellar patients was significantly affected. The uncoupling between the grip and the transport component was most obvious in the profile of the grip aperture. In the cerebellar patients, the grip aperture had a steeper rising profile, and the time of the maximum grip aperture was achieved, on average, ∼20% earlier compared to the healthy controls. Their study suggests that the cerebellum is a center responsible for arm-hand coupling, and for that reason in the cerebellar patients the arm and hand movements were decoupled (i.e. movements of the grip were not properly adjusted with respect to the reaching component, they behaved as if they were independent). The subsequent study of Zackowski et al. (2002) corroborated the arm-hand decoupling effects reported by Rand et al. (2000).

If the hypothesis that the dentate and interpositus nuclei of the cerebellum implement arm-hand coupling is indeed valid, artificially "lesioning" our model by removing the arm-hand coupling block, which in our model correspond to the dentate-interpositus of the cerebellum, should produce the arm-hand decoupling effects as observed in the cerebellar patients (Rand et al., 2000). This validation would favorably support our model. The plots from Figure 4.4 show that removing the arm-hand coupling block produces the profiles of the grip movements, which compared to normal (coupled) arm-hand movements, are very similar to the results observed in Rand et al. (2000). More specifically, decoupling the arm and hand causes that the hand system evolves independently of the arm, which is observed in the steeply ascending profile of the grip opening and significantly earlier achieving the maximum grip aperture. In other words, virtual lesioning of our model by removing the corresponding coupling block produces similar effects to those observed in patients with cerebellar pathology. After removing the arm-hand coupling, the reaching and grasping were

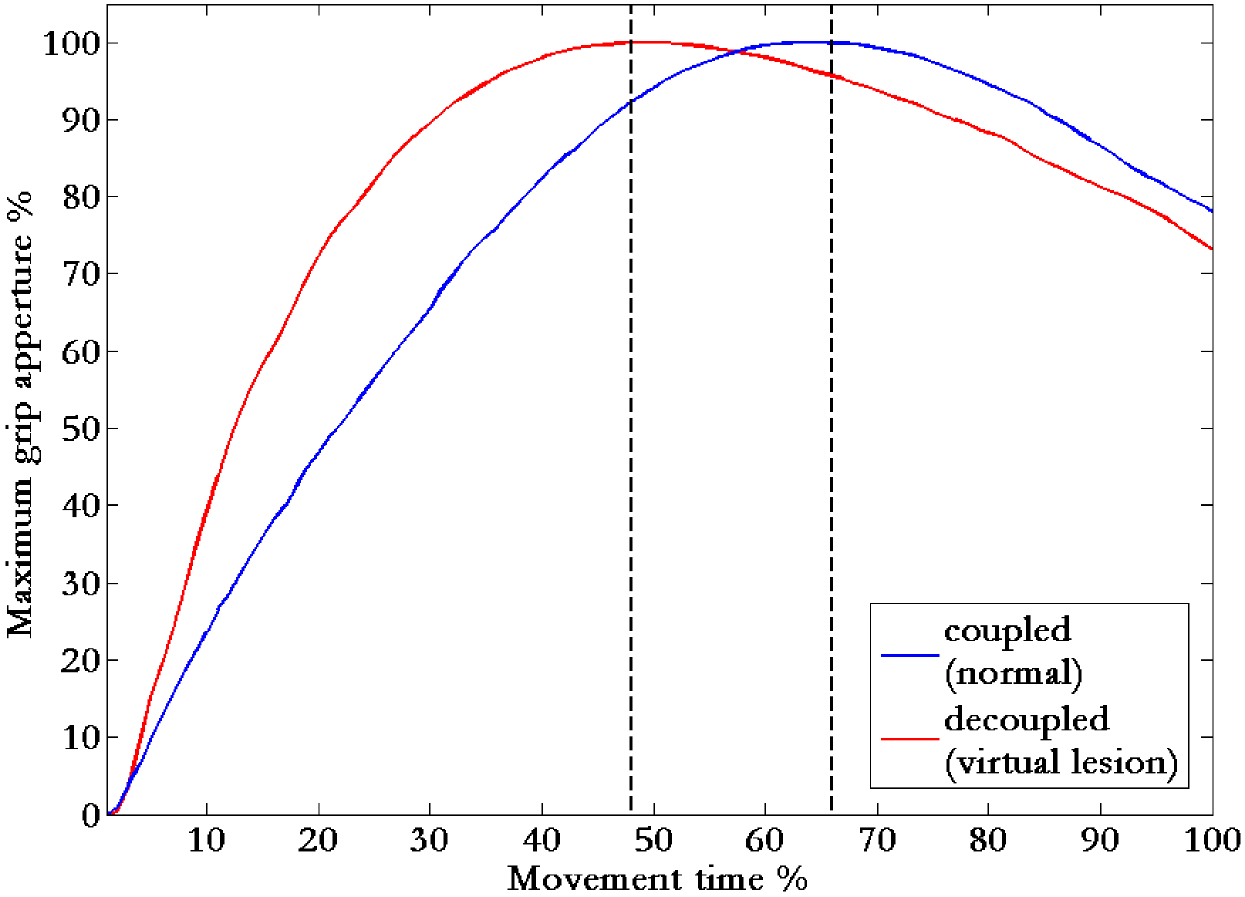
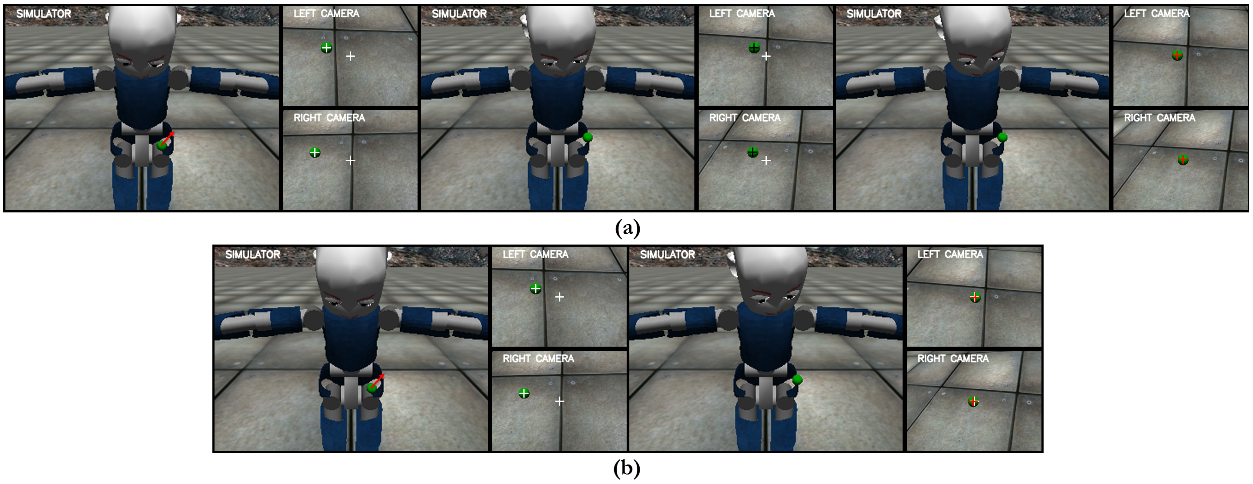


Figure 4.4: Comparison of the grip profile when the arm and hand are coupled (blue) versus the case where the (virtual) lesion removes arm-hand coupling and the arm and the hand systems evolve independently (red). The grip aperture is normalized between 0 (starting hand posture) and 100% (maximum grip aperture). The time is normalized, as in the graphs of Rand et al. (2000). The time difference of ~20% between the maximum grip aperture of normal (coupled arm-hand) and cerebellar patients in the experiment of Rand et al. (2000) is observed in our study, as well.

not actively synchronized during the task, hence our robot occasionally did not manage to grasp the object in a successful manner, similar to the cerebellar patients (Zackowski et al., 2002).

4.4.3 Double-step saccade adaptation experiment

Here, we demonstrate the saccade adaptation ability of our model. The double-step saccade paradigm is a widely used experimental protocol for studying the adaptation of the gaze control system in humans and monkeys (Robinson and Fuchs, 2001; Optican, 2005; Tian et al., 2009). In the most common variant of this experiment, a saccadic target is first presented at one position in the visual periphery (i.e., extrafoveal location). Immediately after a saccade to the target is elicited, the target makes a jump (onward, backward, top, down or a combination of them) to another position in the neighborhood of the first location. This artificially induces an error signal in the saccade programming that should be compensated by adapting the gaze system. Because saccades are fast ballistic eye movements, due to visual blur, they do not rely on visual feedback in flight. This means that after the initial saccade is elicited, the gaze first lands at the first target position (before the target jump). Once visual feedback is available during the fixation, the corrective saccade can be programmed to land at the second, final target position. If the target is systematically displaced in these double-step trials, after several hundred (in humans) up to ∼1500 saccades (in monkeys), the CNS learns to adapt the saccade mapping in order to directly, in a single saccade, hit the expected location of the target (second target position, after the jump). This adaptation of the mapping is a gradual process.



(a)

(b)

Figure 4.5: Double-step saccade adaptation experiments. The first row (a) shows the behavior of the model at the beginning of the saccade adaptation trials. The first snapshot corresponds to the initial configuration of the robot and the scene before the saccade is initiated. Based on the biretinal error, the saccade is programmed and initiated. Immediately after this the target is perturbed to a new position (indicated by the red arrow). The gaze lands at the position where the target would be if the perturbation did not occur. The corrective saccade (third snapshot, first row), based on the updated biretinal error of the target after the fixation, is issued, and the gaze successfully lands at the target. After 1000 iterations, the IMLE learns the adapted mapping, and it is able to compute the gaze joints that provide a direct target fixation after the saccade jump (b). The saccade adaptation in double-step target jump experiments is the behavior widely observed in monkey and human studies.

Once we have the initial biretinal-gaze joint mapping (Figure 4.3), we proceed with the double-step adaptation of our model. Similar to the babbling-like exploratory procedure outlined in Figure 4.3(a), we randomly place the initial target position in the workspace. The target is perceived and encoded in the form of the biretinal error, based on which the saccade end position is computed by querying the corresponding IMLE model, as explained in Section 4.3.2. Immediately after the saccadic gaze movement is initiated by the gaze DS (Section 4.3.3), the target jumps to the second position, by the fixed displacement vector. To make the learning more challenging, we displace the target both along the y-axis of the workspace (left-right direction from the robot's view) and z-axis (vertical displacement), by 2.5 cm in both directions. This spatial displacement is perceived as vertical and horizontal offset in the stereo image planes. At the beginning of the double-step saccade adaptation experiment, the model behaves exactly as humans or monkeys. It initiates the gaze commands to the first target position, then the gaze lands at that position (without using visual feedback during the flight). When the first saccade is completed, a corrective saccade is issued. The corrective saccade steers the gaze to the final target position (position after the jump, Figure 4.5 (a)). Once the object is fixated, an updated relative gaze joint displacement is computed $\tilde{\triangle} q_g = q_{before\_1st\_saccade} - q_{after\_2nd\_saccade}$, and the IMLE mapping (Eq. 5.10) is incrementally adapted.

After 1000 double-step saccade performing-and-learning trials, the number comparable to human and monkey saccade adaptation studies (Robinson and Fuchs, 2001; Optican, 2005; Tian et al., 2009), the model learns to successfully accommodate for the target jump by directly issuing the saccade to the expected position after the target jump (Figure 4.5 (b)), as humans or monkeys do. In our model, the gaze IMLE model corresponds to the oculomotor vermis-caudal fastigial nucleus complex. The oculomotor vermis and caudal fastigial nucleus are reported to take part in gaze learning in double-step saccade adaptation experiments, both in humans and monkeys (Desmurget et al., 2000; Robinson and Fuchs, 2001; Optican, 2005; Tian et al., 2009).

4.4.4 Predictions of the model: new experiment with primates

Because our model shares a high level of parallelism with the visuomotor control principles implemented in the cerebellum and the cerebral cortex, some fundamental testable predictions could be established. In our model, gaze-arm-hand motor coupling is implemented in a manner that the efference copy of the gaze motor commands from the oculomotor vermis and the caudal fastigial nucleus (gaze control) is transferred to the dentate-interpositus (arm-hand coupled control). Therefore, it is important to propose a real-world experiment that could validate this assumption of our model. To achieve this, we again take some inspiration from the neuroscientific literature.

In their fMRI study, Miall et al. (2000) found that the activation of the oculomotor vermis, the cerebellar area traditionally related to movement of the eyes (Section 4.2.2), was increasingly active in combined manual and ocular tracking compared to ocular tracking alone. On the other hand, in the interpositus, the area related to the arm-hand movements and coupling (Section 4.2.2), Robinson (2000) has observed a significant number of neurons that respond during saccade related activity.

Expanding this line of research, from our model, we can propose an experiment with humans or monkeys that could shed more light on the coordination principles implemented in the cerebellar nuclei. Namely, our model suggests that the perturbation (for example, induced by the transcranial magnetic stimulation (TMS) pulse) of the oculomotor vermis-caudal fastigial nucleus system during the simultaneous control of the gaze and the arm in a visually guided reaching task, would not only affect the gaze movements by the expected change of the efference copy of the gaze motor commands, but that the perturbation would be observed in the arm's kinematics, as well. The rationale behind this is that the disrupted efference copy of the gaze motor commands from the vermis-CFN would be transferred to the dentate-interpositus, responsible for arm-hand coupled control, as in our model.

4.5 Summary and discussion

In the preceding sections, we have tackled the target encoding used for visuomotor control and the coupled motor control of the eyes, arm and hand. Many neural centers are involved in visuomotor actions. Although the eyes, arm and hand represent different systems, their motor actions share common principles and during execution of a visuomotor task, these systems are carefully coordinated. The neural structures for visuomotor control need to solve two main tasks: (a) an appropriate representation of reference frames used for the respective effectors and (b) coordinated control of these effectors.

The reference frames used for motor visuomotor control have three common principles: (a) they are egocentrically represented, (b) they are represented in terms of the difference between the current and the desired state, (c) they are updated on a real-time basis. The eye movements are programmed on the basis of retinal error, and the retinal commands are in the later stages converted to eye and neck joint angles. Unconstrained arm and hand movements are encoded in relative joint coordinates. These reference frames are not stored in an offline manner; they are updated in an online fashion as the task progresses.

Similarly, motor commands for these effectors have a number of common principles: they are feedback controlled based on the aforementioned motor error representation, the movements of many joints of an effector are synchronously programmed and inter-effector commands are synchronized in the loop. The reference frames for visuomotor control are represented across a network of cortical areas that are connected to the cerebellum via the recurrent signal routing loops. The cerebellum is the primary neural center for computing synchronous multivariate motor commands.

In this chapter, we have presented the first model, to our knowledge, that is able to unify, on a functional level of abstraction, a number of principles observed from neuroimaging data, studies of brain lesions and neurophysiological results. This model is not a very detailed model of neural circuitry, its contribution is rather to serve as a sketch of the main computational principles involved in visuomotor control and the functional interaction between the cortex and the cerebellum.

This model is also useful for robotics, because it combines desired properties of the model learning methods and visual servoing, which are often considered as separate approaches in visuomotor control. The architecture of the model is modular, which makes it suitable for further biological modeling and extending. For example, we did not include modules that share a resemblance with the inferior temporal cortex and the prefrontal cortex, the areas involved in higher level object recognition, task planning and motor sequencing. The modular architecture of the assumed approach could easily provide further integration of such models. The functional abstraction of our approach in modeling cortical-cerebellar visuomotor control could make possible more detailed biological modeling of the gaze, arm and hand subsystems within the IMLE and CDS frameworks as a “computational umbrella”, as long as the main CDS requirements for the stability of coupled dynamical systems are respected. The modular architecture of our approach could make it possible to add new modules when new evidence is gathered.

Finally, in this chapter, we have proposed a novel experimental paradigm that can provide additional insight into the nature of the cerebellar motor coupling, and consequently, confirm or reject our model.

5 Models of Motor-primed Visual Attention for Humanoid Robots

If we imagine a robot bartender in a real-world context, equipped with an active stereo camera system that has the task to grasp a glass, fill it with a beverage of choice, and serve it to a guest. In a visually-aided manipulation, based on the standard computer vision processing approach, during reaching and grasping for the target object, in every cycle of the control loop, vision scans every part of both stereo images searching for the target object and potential obstacles, in order to update the robot's knowledge about their state (position, orientation and other properties of interest that might change during a task). Assume that the motion of the arm has been initiated and is directed toward a specific object, say a wine glass (the obstacles will by definition be all objects that obstruct an intended movement). Here, a question arises: why would one want to scan the peripheral parts of the stereo images for obstacles, since they correspond to regions in the workspace ten meters or so from the wine glass that is at around 30 cm from the hand? Clearly, the space scanned should be restricted to a region of space that is motor-relevant.

Contrary to robots, humans and non-human primates have the ability to rapidly and graciously perform complicated tasks with a limited amount of computational resources. The attentional system efficiently selects only a subset of information relevant for reaching and grasping among the plethora of visual information. The attentional system operates efficiently and routinely manages the challenging task of selective information processing, in a seemingly effortless manner, by means of highly customized attentional mechanisms. When dynamically changing environmental conditions demand rapid motor reactions, there is no time to compute the full visual model of the world (Ballard, 1991; Wilson, 2002). The humans and non-human primates use attention to select important visual information, and compute only a relevant subset of them on the fly.

In visual attention, two mechanisms are recognized: *covert attention* and *overt attention* (Werner and Chalupa, 2004). Covert visual attention corresponds to an allocation of mental resources for processing extrafoveal visual stimuli. Overt visual attention consists in active visual exploration involving saccadic eye movements (Figure 5.1). These two mechanisms are instantiations of the same underlying mechanism of visual attention, hence intermingled both functionally and structurally, working in synchronization and complementing each other. Covert attention selects interesting regions in the visual field, which are subsequently attended with overt gaze movements for high-acuity foveated extraction of information (Hoffman and Subramaniam, 1995; Findlay and Gilchrist, 1998; Liversedge and Findlay, 2000). Visual attention (covert and overt) is tightly related to the motor control system. Numerous evidence from visual neuroscience and psychology suggests that visual attention is bound and actively modulated with respect to spatio-temporal requirements of reaching and grasping (Hayhoe et al., 2003; Baldauf et al., 2006; Baldauf and Deubel, 2008; Geisler, 2008; Baldauf and Deubel, 2009). While saliency-based attentional mechanisms have been very influential in robotics, on the other hand, motor-primed attentional effects have received little attention to date (Begum and Karray, 2011). Figure 5.1 illustrates how attention is drawn toward manipulation-relevant regions of the visual field, even in a common, well-rehearsed natural task such as tea serving.

In this chapter, we hypothesize that such a biologically-inspired, explicit, active adaptation of attention with respect to motor plans can endow robot vision with a mechanism for the efficient allocation of limited visual resources. This approach contributes to the state of the art in visual-based reaching and grasping, tackling visual attention from a new, alternative perspective where visual attentional relevance is not defined in terms of low-level visual features such as color, texture or intensity of the visual stimuli, but rather in terms of manipulation-relevant parts of the visual field as visually relevant regions. In our model, the attentional mechanism becomes a fundamental building element of the motor planning system and vice versa. At each cycle of the control loop, the visual and motor systems modulate each other by exchanging control signals. In this work, we show that modulation of visual processing, which emerges from the motor system, can drastically improve visual performance, in particular, the speed of visual computation, one of the most critical aspects of the system. The proposed approach is evaluated in robotic experiments using the iCub humanoid robot.

We next briefly review related work on computational modeling of visual attention, its use in robotics, and the biological evidence onto which we ground our approach to tackle the existing problems.

5.1 Background research

5.1.1 Computational modeling of attention and robotic attention

Most of the modern work on computational modeling of attention draws inspiration from the feature integration theory of attention from psychology (Treisman and Gelade, 1980). The feature integration theory argues that low-level, pre-attentive features attract visual attention in a bottom-up, task-independent manner. The intuition behind this approach is that a non-uniform spatial distribution of features is somehow correlated with their informative significance. The influence of the low-level features on capturing attention is motivated by the functions of the neural circuitry in the early primate vision and experimental findings in scene observation tasks (Wolfe, 1998; Reinagel and Zador, 1999; Geisler, 2008).

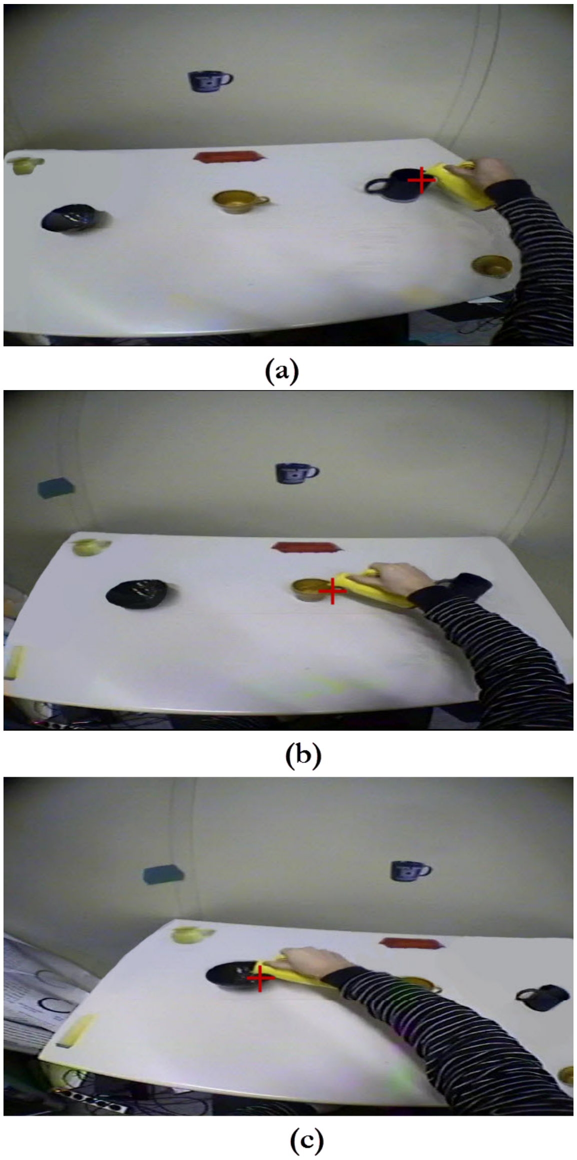


Figure 5.1: Experimental setup with a natural task. The subject is required to pour the tea into two cups and one bowl that are placed close to the horizontal midline of the table. 4 pictures of various objects are placed close to the border of the table, and 2 pictures are placed on the wall facing the subject. These pictures play the role of visually salient distractors because they share the same visual features with the objects, but remain completely irrelevant for manipulation through the entire task. The *overt attention*, i.e. gaze movements, together with the scene as viewed from the subject's standpoint are recorded by using the WearCam system (Noris et al., 2010). The order of the figures from left to right corresponds to the progress of the task. The cross superposed on the video corresponds to an estimated gaze position. It can be seen that the gaze is tightly bound to an object that is relevant to spatio-temporal requirements of the task. In spite of the presence of salient distractors, the gaze remains tightly locked on the current object of interest. This behavior cannot be predicted by the feature-based saliency maps, even with top-down extensions because in manipulation tasks perceptual processing is biased toward manipulation-relevant regions of the visual field, not toward the most textured or distinctively colored stimulus.

By far, the most influential computational implementation grounded in this theory is the concept of the saliency map (Itti et al., 1998). In the aforementioned model low-level features such as color, orientation, brightness and motion are extracted in parallel from the visual input. The visual input is represented as a digitized 2D image. Low-level features from the visual stimuli compete across local neighborhoods, and multiple spatial scales building spatial banks of features that correspond to center-surround contrast computed across different scales. The feature banks are normalized and aggregated by a weighted sum to create a master saliency map. The focus of attention is driven by the interplay between a winner-take-all mechanism (WTA) and an inhibition of return mechanism (IOR) that operates on the final saliency map. This pure bottom-up approach, driven by the early perceptual pop-out features, has been subsequently extended to guided visual search by an additional weighting of the feature channels with a top-down bias that comes from the prior knowledge about objects (Navalpakkam and Itti, 2005; Frintrop, 2006).

Related work in robotics is heavily influenced by the aforementioned Itti-Koch computational model of attention. Whereas most of the computational models implicitly assume covert attention shifts, i.e. no movements of the head and the eyes are involved, most robots are equipped with an active camera system, which makes them suitable for active, overt visual exploration. These robotic applications inherently rely on a saliency map-based scheme to evaluate visual stimuli, and then, instead of shifting covert focus of attention, they actively initiate saccadic movements of the cameras to bring the fixation to the most salient point in the visual field (Begum and Karray, 2011). A number of robotic applications are primarily concerned with implementing saliency maps in order to achieve biologically-inspired saccadic and smooth-pursuit eye movements either with a single pan-tilt camera or a complete robot head (Manfredi et al., 2006). These schemes have been extended to biologically inspired log-polar vision (Metta, 2001; Orabona et al., 2005). Saliency-based attention has been studied in conjunction with exploration, development and learning for humanoid robots (Orabona et al., 2005). Attentional-based vision has been addressed as an aid to sociable robots to improve human-robot interaction (Breazeal et al., 2001; Aryananda, 2006) and in imitation learning (Doniec et al., 2006; Ogino et al., 2006).

5.1.2 Current shortcomings of attention-based models for robot vision and their biological solutions

Although the efforts made in the robotic community have been very fruitful, expanding theoretical foundations and providing practical applications of attentional mechanisms, the most prominent use of attentional schemes still remains applied to object tracking, scene exploration, mimicking the human visual system for robotic studies of development and for providing human-like visual behavior for sociable robots (Begum and Karray, 2011). A very significant drawback of attentional models based on early perceptual saliency, for the purposes of visually driven motor control, is that an attentional relevance is computed solely on the structure determined from low-level visual stimuli projected on the retina, whereas neither the 3D structure of the environment, physical constraints such as body kinematics nor motor action plans are taken into account. The use of attention for active, real-time vision-based manipulation that relies on reliable visual information at each cycle of the control loop continues to be very limited. This is an issue we aim to address in this work. In particular, we identify the following three issues as critical: i) speed of computation, ii) distribution of focus of attention and iii) salient features.

Speed of computation and distribution of focus of attention

Attention in primates evolved as a cheap, efficient and inherently embedded mechanism to select a small subset of abundant visual information for further, high-level processing. The primary reason for this is to efficiently optimize the use of scarce computational resources. However, as previously mentioned, most work in robotics related to attention is motivated by the saliency model of Itti and Koch (Itti et al., 1998). Regardless of the massively parallel architecture, constructing a saliency map is an extremely intense computational task. The best reported times on CPU-based implementations, highly specialized for efficiency, are of an order of 50 ms for a single map (Kestur et al., 2012), the time which doubles for a stereo system, after which, in addition, some high-level visual processing is done in the later stages in the visual processing pipeline. This prohibits applications of the classical saliency map approaches for fast real-world robotic problems such as real-time adaptation to perturbations in grasping tasks with obstacle avoidance.

The majority of models of attention assume that a focus of attention, the so-called attentional spotlight, is a circular shaped region of a fixed radius (Posner et al., 1980), which is centered at a point with the highest saliency in the visual field. Zoom-lens models extend the attentional spotlight concept by allowing the radius of an attentional "window" to change with respect to task demands (Eriksen and James, 1986). Both the spotlight and zoom-lens models restrict applicability of attentional mechanisms for real-world robotic scenarios in complex tasks because only one location in the visual field is (covertly) selected as the focus of attention, toward which the further attentional interest is oriented (covertly or overtly). A number of recent studies from visual neuroscience and psychology suggest that covert attention can take on a complex spatial arrangement (Baldauf and Deubel, 2010). Baldauf et al. have found that covert attention supports pre-planning of a rapid sequence of movements toward multiple reaching goals, by distributing peaks of attention along an intended reaching path (Baldauf et al., 2006; Baldauf and Deubel, 2008). These findings show that covert attention can be distributed not only at one location, as overt attention, but rather simultaneously forms a complex "attentional landscape" in the visual field. Schiegg et al. found that covert attention can be split into multiple foci that are deployed in a way to pinpoint individual locations of intended contact points of the fingers during precision grasping (Schiegg et al., 2003). The experiments with non-human primates have shown that visual receptive fields can even adapt after several minutes of the tool use by elongating their shape to covertly overlay the tool held in the hand (Làdavas, 2002; Maravita and Iriki, 2004).

Computational models of attention have shown good performance and significant statistical similarity to human strategies in simple scene viewing and guided search tasks (Itti et al., 1998; Reinagel and Zador, 1999), but describing human gaze behavior in more complex tasks is far beyond their capabilities. We hypothesize that this is attributable to the fact that only low-level image features are taken into account by the models that compute attentional relevance, whereas the strong top-down bias from motor information is completely ignored. This is rather surprising, considering that there are numerous evidences that report on the very significant coupling between the motor system and attention allocation. Even in pure perceptual tasks, where vision does not support ongoing arm movements, the peripersonal space[1] receives a prioritized covert visual processing compared to the extrapersonal space (Maringelli et al., 2001; Làdavas, 2002; Losier and Klein, 2004), with the peaks of the attentional relevance of visual stimuli close to the hands (Reed et al., 2006; Abrams et al., 2008; Cosman and Vecera, 2010; Davoli et al., 2012). The importance of visual specialization of the peripersonal space is even observed at the level of the parts of the central nervous system. Neurophysiological studies in humans and non-human primates have revealed specialized circuits in the putamen, parietal cortex and ventral premotor cortex that are devoted to processing of visual stimuli within the peripersonal space (Fadiga et al., 2000; Weiss et al., 2000; Rushworth et al., 2001; Làdavas, 2002; Reed et al., 2006). Previc, in his well-known theory of visual field specialization, hypothesized that the visual prioritization of the peripersonal space emerges from functional relationships between the vision and motor systems (Previc, 1998). In this view, the peripersonal space is inherently more visually salient than the extrapersonal space because it supports motor activities with the hands.

Behavioral studies that analyzed the distribution of covert attention in visuomotor tasks have shown interesting results. Covert attention is brought to objects relevant to manipulation, even when reaching for multiple targets in a sequence (Baldauf et al., 2006), or in parallel by engaging bimanual manipulation (Baldauf and Deubel, 2008). The starting position of the hand (Eimer et al., 2006) and its goal position (Baldauf and Deubel, 2009) receive prioritized visual processing when preparing arm movements. Deubel and Schneider found that deployment of covert visual attention at an obstacle occurs when the obstacle obstructs intended arm movements, however, in cases when it does not obstruct intended manipulation it is not covertly attended (Deubel and Schneider, 2004). Deployment of covert attention could be modulated by motor plans as tightly as to support planned finger movements during grasping (Schiegg et al., 2003).

Very few, if none, of the mechanisms reviewed in this subsection, are utilized in the modern computational attentional methods embedded in robotic visually-driven reaching and grasping. Taken together, biological studies indicate an apparent dependence and an active modulation of

[1] The peripersonal space is defined as the space around the body within which an agent (a human, monkey or a robot) can manipulate objects without using locomotion to move the body, whereas the extrapersonal space is postulated as the space beyond the peripersonal space and its representation is used for navigation and orienting, see (Previc, 1998) for more.

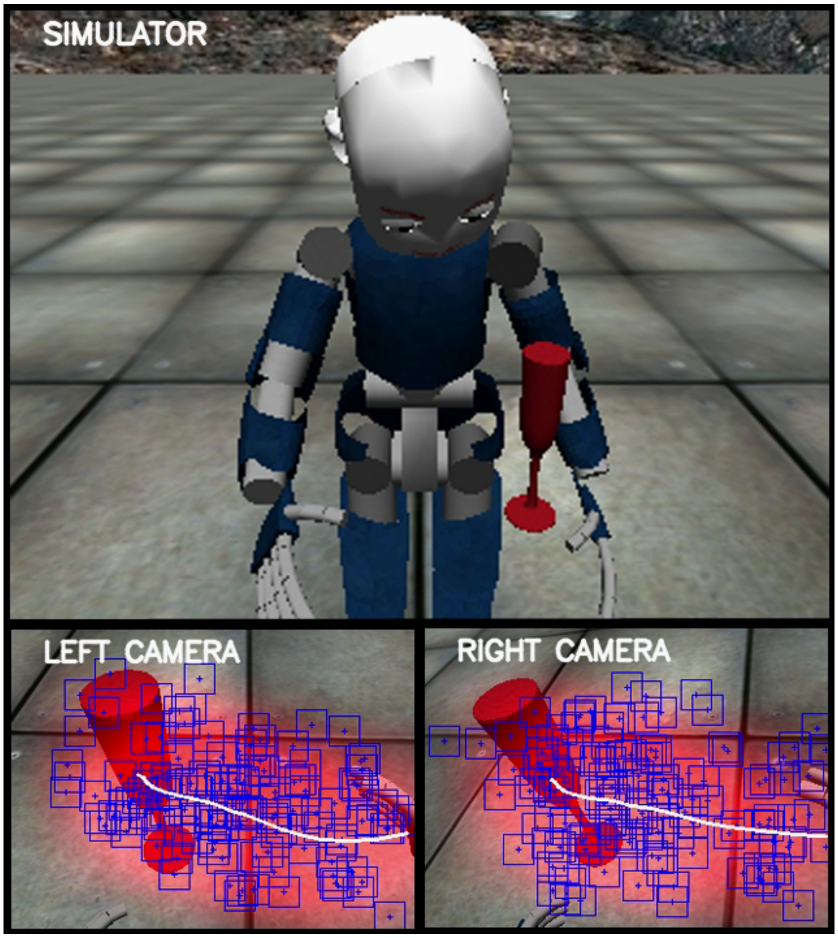


Figure 5.2: The figure displays the main idea of the proposed approach: nonuniform image processing driven by a motor-primed visual attentional landscape. Visual space is prioritized depending on its motor relevance; i.e., visual attention is biased toward motor-relevant parts of the workspace projected to the stereo images. The white line represents a forward-planned (mentally-simulated) movement toward the object to be grasped (red glass). The reddish blend superimposed on the snapshots of the left and right cameras is a visualization of the intensity of the visual attentional landscape. The attentional landscape has a higher intensity closer to motor relevant parts of the visual field. The images are processed in a manner that the spatial distribution of their attentional landscapes is taken into account (motor-relevance is prioritized). The anchors of the scanning windows (blue squares) are sampled with respect to their relevance, i.e. more dense visual scanning is done where the attentional landscape has higher values, and less dense scanning where it has low values. Ignoring irrelevant parts of the images affords significant computational savings, whereas the processing of motor-relevant parts of the visual scene supports visually-guided reaching and grasping.

visual attention on motor information. All these results suggest that low-level feature-based saliency is suppressed when an actor is engaged in visually-aided physical tasks, regardless whether the task is manipulation or navigation, whether the interaction with the object is performed in a parallel or in a sequential manner, and regardless whether gaze movements are suppressed or not. In plain words, in physical tasks, motor-relevant parts of the visual field are visually salient.

The aforementioned behaviors observed in these studies are elegantly explained and unified by the premotor theory of attention proposed by Rizzolatti and coauthors (Rizzolatti and Craighero, 2010). This theory argues that visual attention is a feature that emerges from the motor neural circuits that generate actions, i.e., cortical structures that are involved in arm movements are also responsible for constructing covert visual attention that accompanies the movements. In developing our model, we take the exact approach as argued by the premotor theory of attention: the attentional landscape is primed by the motor system. By equalizing motor-relevant as attention-salient, we aim at tackling the reviewed current weaknesses in the existing attention models. We demonstrate in this chapter that motor-primed visual attention is a very efficient mechanism. Figure 5.2 illustrates the main principles of our approach.

We proceed further with the section that describes how the peripersonal space-primed attention and motor plans-primed attention landscapes are computed.

5.2 Peripersonal space-primed and motor plans-primed attention

In this section, we proceed with modeling the influence of the motor system in the modulation of visual resources. From the evidence presented in the previous section, we took inspiration for this work. Here, we hypothesize that such a modulation of visual processing would provide more efficient visual processing compared to the standard, uniform image processing that is not modulated by the motor system. More specifically, we present two methods to bias the visual processing with respect to the state of the motor system: (a) peripersonal space-primed mechanism, and (b) motor plans-primed attentional modulation. The peripersonal space-primed mechanism is the concept of attention based on the idea that visual attention should be biased toward the reachable space of a robot. The biasing of the attention with respect to the reachable space has been observed in the monkey and human experiments, reviewed in the previous section. The motor plans-primed attention is motivated by the evidence that visual attention is dynamically bound to motor plans of a monkey/human in behavioral experiments (a robot in the case of our modeling). The peripersonal space-primed mechanism is a more general method to tailor visual attention with respect to the motor system because it biases attention to the whole reachable space. On the other hand, motor plans-primed attention is a more specific method, and computationally more efficient because it bounds the attention only to a subset of the reachable space that is defined by the current motor plan.

In order to distribute visual attention with respect to both the peripersonal space and motor plans of a robot, we first need to obtain a transformation that will map the points from the spatial coordinates to the image planes. We next describe a method to compute projections from the workspace to the image plane. Once this transformation is obtained, it is used to construct the two variants of visual attentional landscapes.

5.2.1 Mapping of the workspace to the image plane

Projections to the image plane

Let the Cartesian workspace position be represented as $x \in \mathbb{R}^3$, and the kinematic configuration at the current time of the torso-neck-arm represented with the torso, neck, and head joints as $q \in \mathbb{R}^9$, the transformation function of the form:

$$p_i = f_i(c), \tag{5.1}$$

where $c \in \mathbb{R}^{12}$, $c = \begin{bmatrix} x \\ q \end{bmatrix}$, and $p_i \in \mathbb{R}^2$ represents the projection of a 3D point, taking into account the kinematics of the torso, neck and eye, to the image plane of the i-th camera, where $i = \{left, \, right\}$.

A classical, straightforward approach would be to compute a sequence of kinematic transformations through the torso-neck-head kinematic chain in order to obtain the extrinsic camera parameters, and use them together with the intrinsic parameters of the camera to obtain the projective transformation. For a stationary camera, calibration of all the camera parameters can be easily accomplished by formulating the problem as linear regression and solving it by using the least-squares approach. However, for cameras mounted on a moving robot's head, the problem includes the torso-neck-head joints. This imposes the need for calibration of the kinematic chain, because most often a real robot differs from its nominal kinematic model. Hence, the linear problem of calibration for a static camera becomes highly nonlinear for a camera mounted on the head as we include the torso-neck-head joints as independent variables.

Clearly, an alternative solution is to rely on a non-linear approximation using any of the standard machine learning techniques for non-linear regression. Similarly to what happens with human newborns, the robot starts by exploring in a babbling-like manner a set of kinematic configurations. During this exploration it segments an object (e.g., a small colored ball) placed at a randomly chosen position from a set of known positions in the workspace. The data obtained during the exploration (encoder readings of the joints in the torso-neck-head chain, the position of the object in the workspace and its projection to the camera planes) is used to learn a mapping function. A problem associated with this approach is that the babbling-like exploration with the real robot is very costly because in order to build a reliable estimate of this nonlinear mapping, the size of a training set needs to be arbitrarily large to be representative, usually of an order a few thousand data samples.

Here, we take an intermediate step that represents a compromise between the two previously described approaches. The idea is to take advantage of the simulator of a robot in order to obtain a

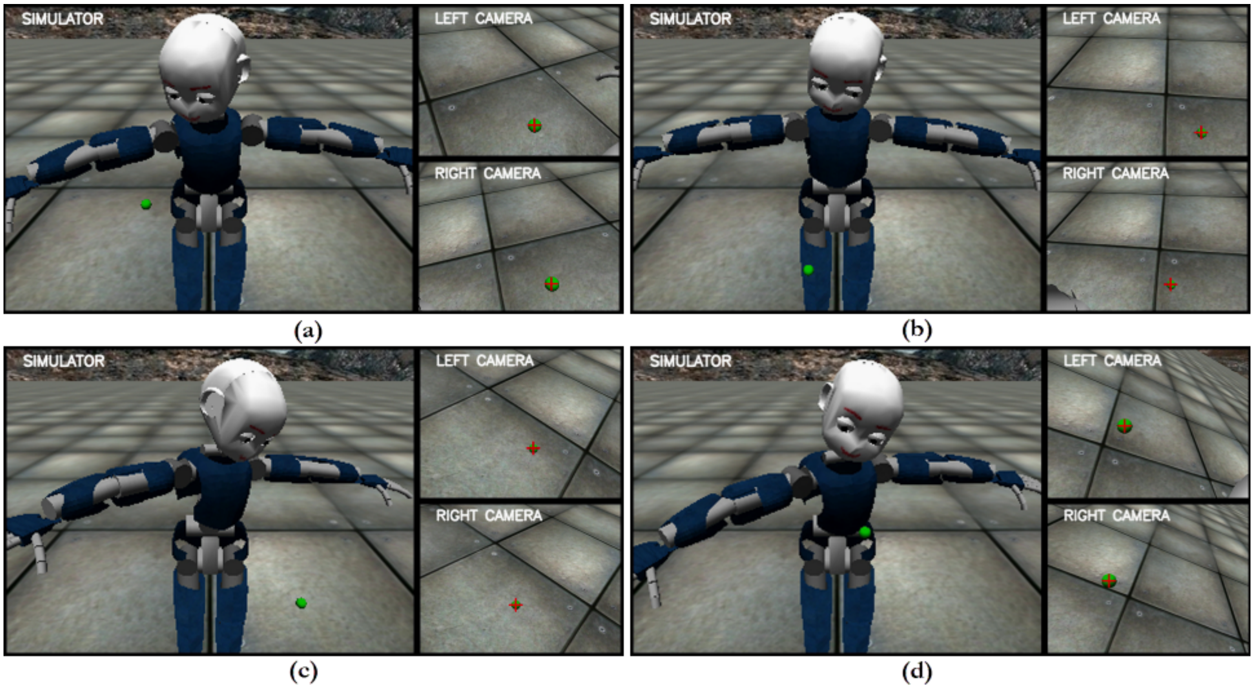


Figure 5.3: Exploratory behavior used to learn an adaptable model of the visuomotor transformation. The snapshots from the simulator (a-d) show several examples of exploratory configurations. The torso-neck-head-eye-joints (9 DoF) are sampled from the uniform distribution within their respective joint limits, and, similarly, the position of the green ball is sampled from the uniform distribution defined within the reachable space. For each sampled configuration, the encoders are read, and the locations of the segmented ball in the stereo images are obtained. After the exploration, these data points are utilized to learn a neural network model of the workspace to the stereo image projections. The advantage of having such a model is that the model can be easily adapted to data points obtained from the real robot by taking similar exploratory procedure, in order to adapt the model to the discrepancies between the mathematical model and the kinematics of the real robot.

large number of training samples by employing babbling, and use this data set to estimate an initial set of parameters for the mapping model (Figure 5.3). This model is then incrementally adapted with the data obtained from the real robot, which accounts for only a small fraction of the data obtained in the simulator.

Learning the map

A feed-forward neural network is a suitable machine learning algorithm for our application for a number of reasons (Haykin, 1998). Feed-forward neural networks can compute multi-input-multi-output functions. Their output is very fast to compute in real-time because the computation consists of a short sequence of matrix-vector multiplications, followed by (non)linear transfer functions. Feed-forward neural networks are suitable for incremental learning, either in a batch or a stochastic, online mode. This allows us to first estimate this function from the data in the simulator, and then adapt it with the data from the real robot.

The parameters of an architecture of neural networks for transformation from the workspace to the image coordinates (i.e. number of layers and the number of hidden units, etc.) are determined by using grid-search on the mean squared error (MSE) between the recorded image projections and retrieved projections from the model. We tested 10 different network architectures, and for each architecture, we performed 10 learning runs in order to ensure robustness with respect to random initialization of network parameters. We used the Levenberg–Marquardt optimization algorithm with early-stopping in order to prevent overfitting (Haykin, 1998). The recorded data set is randomly partitioned for 70 % of the data devoted to training, 15 % data for validation, and 15 % data for testing. The lowest MSE on the testing set is obtained using two hidden layers with 25 nodes in each hidden layer. Transfer functions in the hidden layer are hyperbolic tangent sigmoid, and in the output layer are linear. The data set is normalized to obtain zero mean and unity variance. In order to get the real-time performance, a network class is implemented in C++ by using linear algebra functions from OpenCV library (Bradski and Kaehler, 2008). The time needed to transform 50 points by using neural nets to the image planes of both cameras is less than 1 ms.

5.2.2 Peripersonal space-primed attention

In order to be able to distribute visual attention with respect to the peripersonal space of a robot, (a) we need to have a transformation that will map the peripersonal space to the image planes (as described in the previous section 5.2.1), and (b) we need to obtain a representation of the peripersonal space that we will map to the stereo images as the robot takes different postures. We next proceed with describing how we obtain the representation of the peripersonal space and how we learn the peripersonal space-primed attentional landscape.

Representation of the peripersonal space

For modeling the reachable space, classical methods such as polynomial discriminants and geometric approaches compute the boundaries of the robot's reachable space (reviewed by Kim et al. (2014)). The limitations of these methods are that they can only be applied to special kinematic chains and they model the boundary of the reachable space, without any notion regarding which locations of the reachable space are more likely to be attended. We take an exploratory, sampling based approach that overcomes these two difficulties.

We model the peripersonal space by commanding the robot to explore reachable positions by randomly varying the arm joint angles. More specifically, we sample the joint values from the uniform distribution defined over the feasible joint ranges, and we read the achieved 3D end-effector positions from the robot's forward kinematics. Once this exploration is carried out, we store recorded

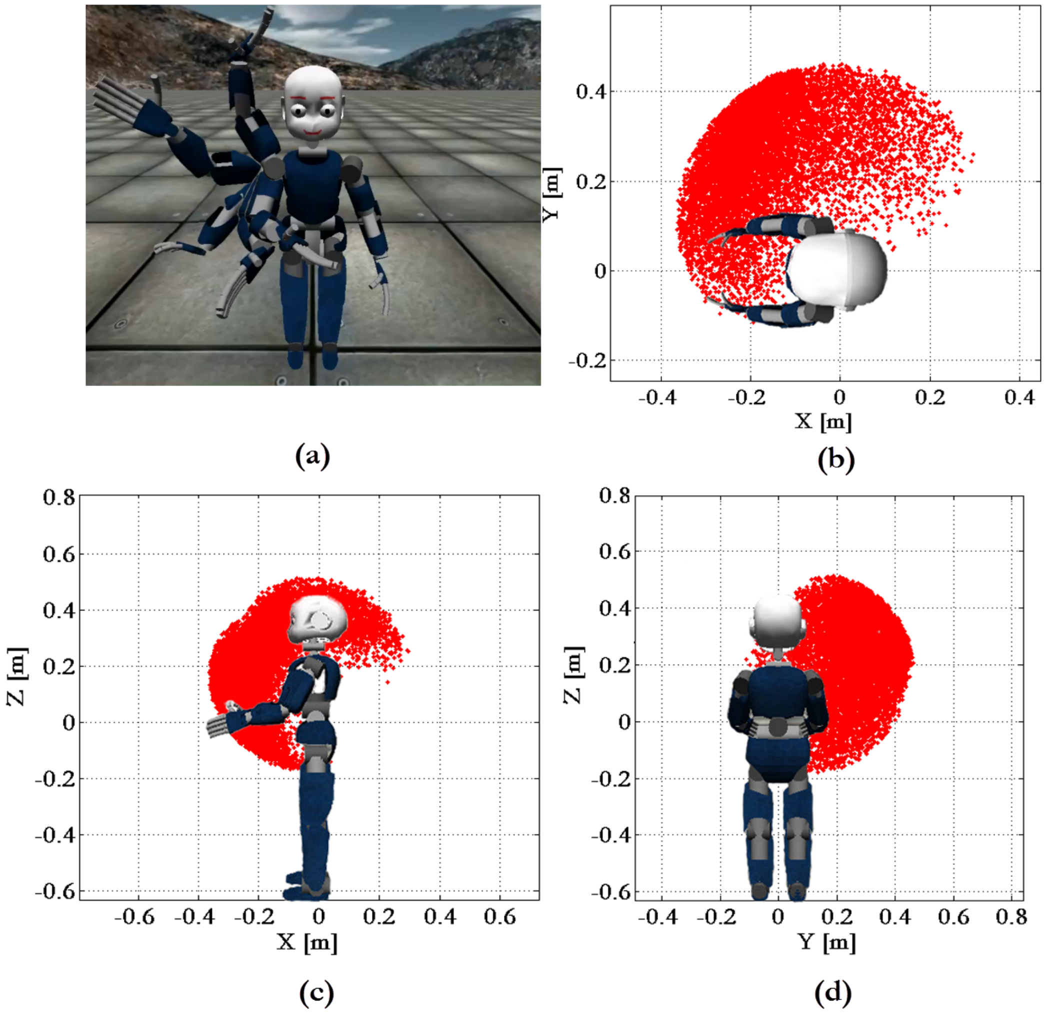


Figure 5.4: Exploratory behavior used to model the peripersonal space for the right arm of the iCub robot. Figure (a) represents several exploratory movements captured in the simulator and superimposed. Figures (b-d) show the sampled cloud of data points with respect to the robot's body in XY (b), XZ (c) and YZ plane (d).

reachable points in a database. Figure 5.4 shows the exploratory procedure that we take and the obtained, sampled representation of the peripersonal space.

Attentional landscape

After obtaining the representation of the peripersonal space, we model the distribution of attention with respect to the peripersonal space. We sample the eye-neck-head joints from the uniform distribution within the joint limits and, for each sampled configuration, we project the previously sampled cloud of the reachable space points by using the previously learned mapping (presented in Section 5.2.1) to the stereo images. This procedure is shown in Figure 5.5.

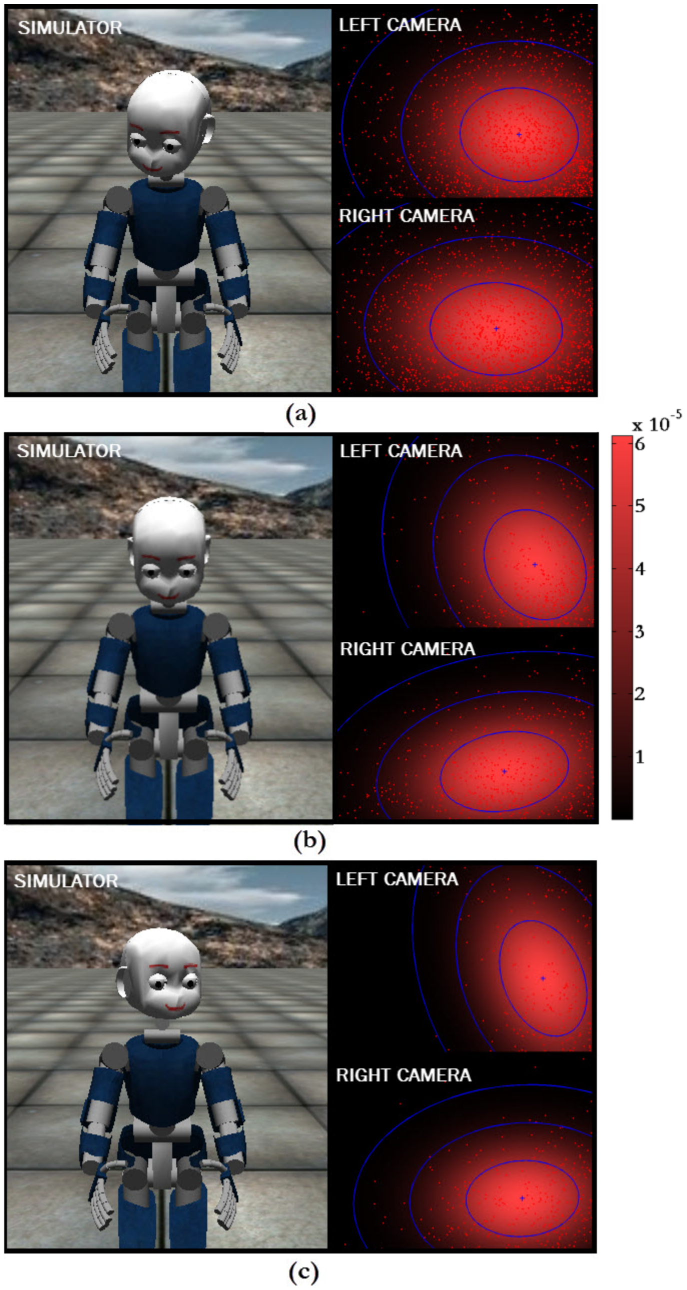


Figure 5.5: Exploratory behavior used to learn a model of peripersonal space-primed attention (a-c). The figures represent several exploratory movements. For each randomly generated neck-head-eye posture, we project the sampled set of reachable points to the stereo image planes. Using bivariate Gaussian distributions to model the elliptical envelopes of projections of the bubble-shaped cloud to the image planes is an intuitive choice. In this case, the bivariate Gaussians, one for the right and one for the left image, are parametric representations of the peripersonal space attention for the stereo setup given the current posture of the robot. The reddish heat maps correspond to the values of the density of the projections. Finally, the mapping from the neck-head-eye joint angles to the parametric representation of attention is learned by using these data. This mapping is used in the run-time to infer how peripersonal attention should be distributed given the set of the neck-head-eye joint angles.

The bubble-shaped cloud of the end-effector locations that models the peripersonal space (Figure 5.4) projects as an ellipsoid-shaped scatter to the left and right cameras (Figure 5.5). We use a bivariate Gaussian distribution to model the scatter of the projected points on the image planes, which represents a parametric representation of a peripersonal attentional landscape, as formulated:

$$\Lambda_{i,t}(p;\mu_{i,t},\Sigma_{i,t}) = \frac{1}{\sqrt{(2\pi)^4 \mid \Sigma_{i,t} \mid}} e^{-\frac{1}{2}((p-\mu_{i,t})^T(\Sigma_{i,t})^{-1}(p-\mu_{i,t}))}, \tag{5.2}$$

where t is the index of the currently sampled configurations and the corresponding projections, $i = \{left, right\}$, and $\mu_{i,t}$ and $\Sigma_{i,t}$ are the mean and the covariance matrix, respectively. In this case, the bivariate Gaussians, one for the right and one for the left image, are parametric representations of the peripersonal space attention for the stereo setup given the current neck-head-eye posture of the robot.

Before we proceed with learning of a function that maps the neck-head-eye posture to the parametric representation of the attentional landscape (the mean and the covariance), we must take into account that the covariance matrix, inferred from such a mapping and used to compute the attentional landscape, must be symmetric and positive definite to ensure the validity of the Gaussian distribution. One solution is to enforce this by projecting the inferred covariance matrix (only symmetric but no guarantees of positive-definiteness) onto the set of symmetric positive definite matrices by using the constrained convex optimization programming. However, addressing this problem involves iterative optimization procedures, which we want to avoid for maximizing computational efficiency. Here we use an alternative approach. We first decompose the covariance matrix into the product of a lower triangular matrix $L_{i,t}$ and its transpose by using the Cholesky factorization:

$$\Sigma_{i,t} = L_{i,t}L_{i,t}^T,\ L_{i,t} = \begin{bmatrix} L_{1,i,t} & 0 \\ L_{2,i,t} & L_{3,i,t} \end{bmatrix}. \tag{5.3}$$

Next we proceed with learning a mapping $\lambda_{i,t} = g_i(q_t)$, defined from the current joint angles $q \in \mathbb{R}^6$ to the tuple $\lambda_{i,t} \in \mathbb{R}^5, \lambda_{i,t} = [\mu_{1,i,t}, \mu_{2,i,t}, L_{1,i,t}, L_{2,i,t}, L_{3,i,t}]^T$, which is an ordered, column-vector arrangement of the elements of $\mu_{i,t}$ and $L_{i,t}$. This mapping is learned with a feed-forward neural network, by using a similar procedure to the one explained in Section 5.2.1. In the run-time, for a given configuration q^*, we infer $\tilde{\lambda}_{i,t}$, i.e.,$\tilde{\mu}_{i,t}, \tilde{L}_{i,t}$, from function g_i. We then compute the attentional landscape as follows:

$$\Lambda_{i,t}(p;\tilde{\mu}_{i,t},\tilde{L}_{i,t}) = \frac{1}{C} e^{-\frac{1}{2}((p-\tilde{\mu}_{i,t})^T(\tilde{L}_{i,t}\tilde{L}_{i,t}^T)^{-1}(p-\tilde{\mu}_{i,t}))}, \tag{5.4}$$

where C is a normalization constant. The reconstructed covariance matrix, computed as the product $\tilde{L}_{i,t}\tilde{L}_{i,t}^T$, is a symmetric positive definite matrix. Considering that the Cholesky lower triangular matrix represents the measure of deviation from the isotropic Gaussian, we can constrain computation of the attentional landscape within the ellipse obtained by multiplying the unit circle

$D = \{p \in \mathbb{R}^2 \mid \|p\|_2 = 1\}$ with $\tilde{L}_{i,t}$ and translating the product by $\tilde{\mu}_{i,t}$:

$$E_{i,t} = \sigma \tilde{L}_{i,t} D + \tilde{\mu}_{i,t}. \tag{5.5}$$

$\sigma \in \mathbb{R}$ is a free parameter that corresponds to the number of standard deviations at which one wants to compute the ellipse, and it is usually set at $\sigma = 3$. The value of the attentional landscape outside the 3σ-ellipsoid is insignificant to affect the distribution of attention and can be neglected. For this reason, we cut-off the attentional landscape at zero outside the 3σ-ellipsoid to avoid computing Eq. 5.4 at these pixels to gain computational efficiency.

5.2.3 Motor plans-primed attention

The peripersonal space primed attention could be seen as a general, multipurpose technique, to compute the distribution of attention to the image regions that correspond to the entire peripersonal space. It might or might not involve reaching and grasping movements. However, because peripersonal space-primed attention is bound by the whole reachable space, it does not utilize particular motor plans of a robot. Additional constraining of the attentional landscape around motor plans-relevant regions results in additional computational savings and more localized visual processing. We here present a way to further constrain the attentional landscape, with respect to motor plans of the robot. This is a more specialized technique than peripersonal space-primed attention.

We use our robotic eye-arm-hand controller, presented in Chapter 3, to generate reaching and grasping movements and to forward-plan the arm-hand reaching trajectory. Learned eye-arm-hand Coupled Dynamical Systems (CDS) are used in order to “mentally simulate” the consequences of intended actions, more specifically, to compute (i.e. plan) an intended trajectory and to identify obstacles. This mentally simulated arm reaching trajectory is transformed to the image planes of the stereo cameras. The projected mentally-simulated trajectory is used to compute an attentional landscape, i.e. a saliency map. We utilize the mentally-simulated trajectory in order to bias visual resources to motor-relevant parts of the visual field, which we describe in the next section.

Attentional landscape

The mentally-simulated trajectory of the arm, from the current position to the final position at the current time t, is represented as $x_t^n \in \mathbb{R}^3$, $\forall n \in [1, N_t]$, where N_t represents the total number of discrete samples. This mentally-simulated trajectory at every cycle of the control loop t is obtained from the CDS controller. The kinematic configuration at the current time t of the torso-neck-head is represented with the torso, neck, and head joints $q_t \in \mathbb{R}^9$, $\forall t$. We use the previously learned transformation function, presented in Section 5.2.1, to perform this mapping:

$$p_{i,t}^n = f_i(c_t^n),\ \forall n \in [1, N_t]\,, \tag{5.6}$$

where $c_t^n \in \mathbb{R}^{12}$, $c_t^n = \begin{bmatrix} x_t^n \\ q_t \end{bmatrix}$, and $p_{i,t}^n \in \mathbb{R}^2$ represents the projection of the trajectory to the image plane of the i-th camera, where $i = \{left, \, right\}$.

After we project the mentally-simulated trajectory to the image planes, we construct an attentional landscape which associates high saliency close to the mentally-simulated trajectory perceived in the image coordinates. To compute the attentional landscape, i.e. a measure of visual processing priority (saliency map), we use a bivariate kernel smoothing function, where kernels are placed at every point of the projection of the mentally-simulated trajectory to the image planes. Formally, we compute an attentional landscape for each camera i as follows:

$$\Lambda_{i,t}(p) = \frac{1}{N_t h^h h^v} \sum_{n=1}^{N_t} K(p - p_{i,t}^n), \tag{5.7}$$

where $p \in \mathbb{R}^2$ corresponds to two-dimensional pixel coordinates of the image plane,

$$K(p - p_{i,t}^n) = k\left(\frac{p^h - p_{i,t}^{n,h}}{h^h}\right) k\left(\frac{p^v - p_{i,t}^{n,v}}{h^v}\right), \tag{5.8}$$

where $k(.)$ represents a kernel, and h^h and h^v are kernel widths along the horizontal and vertical image dimensions. We tested both Gaussian kernels and triangular kernels, and we choose to use triangular kernels because they are faster to calculate. The triangular kernel is expressed as follows:

$$k(z) = \begin{cases} 1 - \left|z\right|, & \left|z\right| \leq 1 \\ 0 & otherwise \end{cases}. \tag{5.9}$$

The kernel smoothing function assigns high values of saliency close to the mentally-simulated trajectory projected to the image planes of stereo cameras, which decrease in the directions away from the trajectory (Figure 5.2). The attentional landscape is used to guide image processing in order to efficiently distribute limited visual resources. The part of the image with higher saliency draws more visual processing, and the opposite is true. In the next section, we explain how we distribute visual processing with respect to the visual attentional landscape, both peripersonal space-primed and motor plans-primed.

5.3 Attention-driven visual processing

In Section 5.2, we presented two techniques of attentional landscapes that can be utilized to distribute visual attention emerging from the motor system. In order to detect objects relevant for the task at hand, a robot must process stereo images. In this section, we propose two methods to use the attentional landscape to guide visual processing. These two techniques make our approach general enough to be used as a pre-modulating technique to almost any kind of standard image processing detectors and segmentation techniques (pixel-by-pixel color segmentation, histogram-based detectors, Viola-Jones, SIFT, SURF, etc.). The two processing schemes that will be presented apply to both peripersonal space-primed and motor plans-primed attention.

5.3.1 Thresholding and sampling

One simple approach, suitable for pixel-by-pixel color processing and interest point detectors-descriptor approaches, is to distribute visual processing to the region of the image where an attentional landscape $\Lambda_{i,t}(p)$ is higher or equal than some threshold d_i. It is easy to empirically estimate the computational time for processing the entire image and from this value estimate cost per pixel. By sorting pixels with respect to ascending values of their saliency, we can pick a number of pixels corresponding to the available computational resources. From this sorted array, we can easily compute the threshold d_i on the attentional landscape. The approximate value of the threshold can be determined in ~3 ms for 4800 subsampled pixels by using the Quick Sort algorithm.

The second attention-driven, visual processing method is concerned with modulating the image processing techniques employ image processing within a scanning window, e.g. Viola-Jones detector, histogram-based detector, Rowley-Baluja-Kanade detector, etc. Here the task is to determine the position of the scanning windows with respect to an attentional landscape $\Lambda_{i,t}(p)$, in order to have more dense scanning where the saliency is large, and less dense scanning in spatial regions with low saliency. Because we use either a kernel smoothing function or a Gaussian function to build an attentional landscape, we can treat the attentional landscape as a bivariate probability density function and use any kind of sampling techniques to sample spatial locations of scanning windows. Again, we can empirically obtain a cost associated to process the image in each window, and from the total visual resources, calculate the number of points to sample from the attentional landscape. We use the Gibbs sampling method (Murphy, 2012). We choose the Gibbs sampling instead of other sampling procedures such as the general Metropolis-Hastings algorithm[2], because the acceptance rate of sampled proposed values is 1, which makes it a very efficient procedure. The procedure operates as follows:

1. start with an initial pixel location: $p_{i,0} = [h_{i,0}, v_{i,0}]^T$
2. for $j = 1, 2, \ldots, M$
3. sample $h_{i,j}$ from the conditional distribution $\Lambda_{i,t}(h \mid v_{i,j-1})$ by using the inverse transform sampling
4. sample $v_{i,j}$ from the conditional distribution: $\Lambda_{i,t}(v \mid h_{i,j})$ by using the inverse transform sampling

[2] The Gibbs sampling algorithm can be viewed a special case of the Metropolis-Hastings algorithm.

5. store $p_{i,j} = [h_{i,j}, v_{i,j}]^T$, increment j and loop over steps 3-5 for the given number M of scanning windows

6. return the set of sampled points: $P = \{p_{i,1}, \ldots, p_{i,M}\}$ (locations of scanning windows)

The Gibbs sampler and inverse transform sampling function embedded in it are implemented with look-up tables as C-arrays for efficiency. The time for querying the Gibbs sampler is ~3 ms for an attention landscape of size 320×240 for 50 sampled scanning windows.

Adjustment when sampling from the peripersonal space-primed attentional landscape

In Section 5.2.2, we presented a method for modeling the peripersonal space attention with one bivariate Gaussian per stereo image. The bivariate Gaussian is suitable for modeling the projection of the 3D peripersonal space blob to the image plane, as we illustrate in Figure 5.5. Once this representation is obtained, it is used to perform image processing according to it. For processing by using the thresholding-based approach, this representation of the attentional landscape can be directly used, however, for the sampling-based approach, we find that it is better to slightly balance it. The steeply rising profile of the Gaussian distribution biases sampling toward its centroid. When we sample a smaller number of windows, this could lead to the case that the objects that lie closer to the boundary of the reachable space are missed. For this reason, we propose using a balanced version of the peripersonal space-primed attention (Section 5.2.2) when doing sampling-based image processing. A balanced peripersonal space-primed attentional landscape is defined in the form of a mixture between the obtained bivariate Gaussian (Eq. 5.4) and the uniform distribution $U(p)$:

$$\Lambda_{i,t}(p; \tilde{\mu}_{i,t}, \tilde{L}_{i,t}) = \pi \frac{1}{C} e^{-\frac{1}{2}((p-\tilde{\mu}_{i,t})^T (\tilde{L}_{i,t}\tilde{L}_{i,t}^T)^{-1}(p-\tilde{\mu}_{i,t}))} + (1-\pi)U(p),\; U(p) = \begin{cases} c, & c \in domain \\ 0 & otherwise \end{cases}, \quad (5.10)$$

where $\pi \in [0, 1]$ is the mixing probability, which is a parameter that can be hand-tuned according to the desired behaviors. Creating the mixture between the Gaussian and the uniform distribution flattens the original Gaussian profile, which results in more spread out sampling and, hence, better coverage of image regions that correspond to the spatial regions lying closer to the boundaries of the peripersonal space. Again, we constrain computations within the 3σ-ellipsoid $E_{i,t}$.

5.3.2 Closing the loop: from covert attentional landscape to overt eye movements and manipulation

It is noteworthy to mention that we recompute and sample the attentional landscape maps at every cycle. This implies that there is no requirement to implement the IOR mechanism and deal with the problems with the change of coordinates associated with standard saliency models (Begum and Karray, 2011), which simplifies our approach and hence reduces the overall computational time.

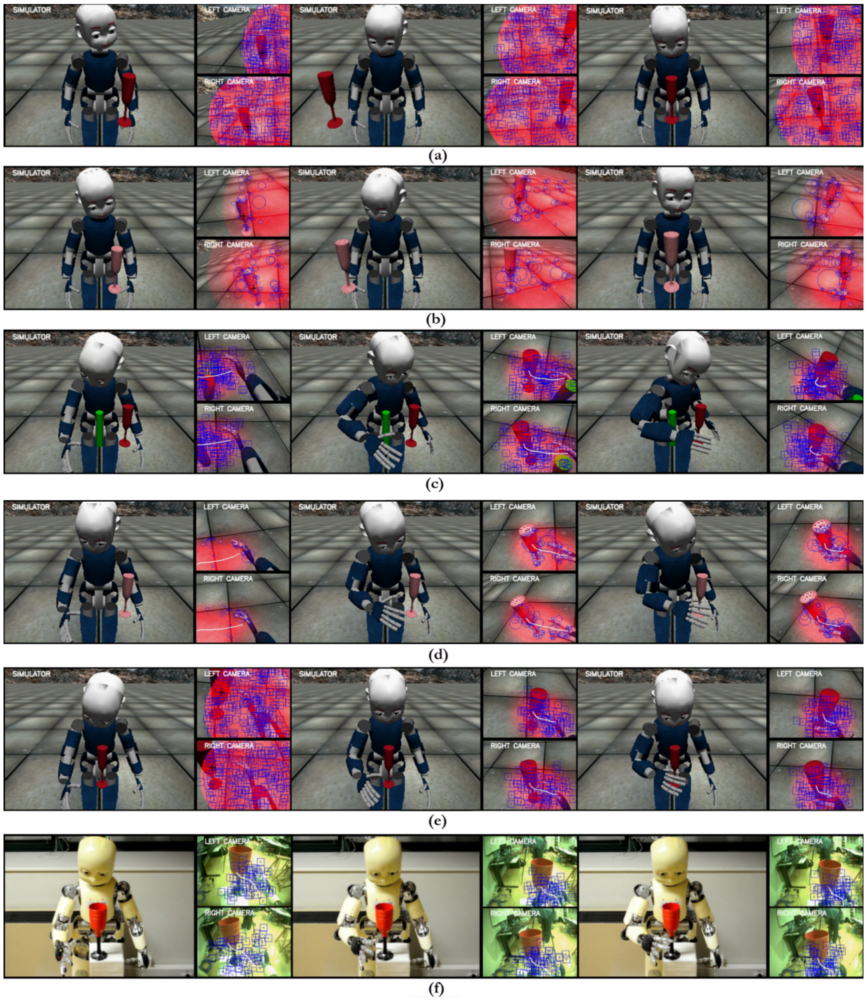
As described in the previous section, when the attentional landscape is constructed, the top-down visual scan is performed in the spatial regions that have high relevance. These two stages correspond to covert visual attention. In the case of motor-primed attention, after the targets (and/or obstacles) are detected, the overt gaze movements are initiated toward the first intermediary target in a synchronous manner together with the arm and the hand motion by using our CDS eye-arm-hand controller (Lukic et al., 2012, 2014a). In a no-obstacle task, the eye-arm-hand system is directly driven toward the target. In tasks with obstacle avoidance the eye-arm-hand system is driven toward the obstacle, which is treated as an intermediary target for the visuomotor system, as explained in Chapter 3. When the obstacle is avoided, the system is driven toward the object to be grasped.

5.4 Results

We validate our method in the iCub simulator and the real robot with a task of visual exploration for initial object detection (peripersonal space-primed attention), and reaching and grasping a kitchenware object (motor plans-primed attention). Resolution of the stereo cameras in the setup is 320×240. We verify this approach with two well-known standard image processing techniques. For the first visual detector, we select a scanning window hue-saturation histogram-based detector. We implement this detector by using functions from the OpenCV library (Bradski and Kaehler, 2008). For the second detector, we selected SURF (Bay et al., 2006)[3]. SURF is a powerful detector because it provides visual features that are robust to moderate changes of the perspective. Because it computes feature point descriptors, it provides the ability to detect partially occluded objects. However, SURF (together with a family of similar detectors like SIFT, GLOH, etc.) is very computationally demanding, with the cost being double for a binocular system, hence it has limited applicability for manipulation where the stereo vision is used in the loop. The total time to process a stereo pair of images in the standard, full-blown way, is for the histogram based detector with the window size 20×20 is 168 ms and for SURF with the Hessian threshold set to 300 is 515.5 ms.

We first test both detectors in the context of peripersonal space-primed attention. The time needed to infer the parametric representation of the attentional landscape by using feed-forward neural networks is negligibly small, close to a tenth of a millisecond. Computing the peripersonal attentional landscape image requires 35.5 ms. These are computations common for both image processing techniques. Sampling from the relevance images, for the histogram-based detector, requires 7 ms for the stereo setup for 50 image windows per image. Performing sparse image processing for

[3] We used the implementation available from the OpenCV library.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.6: Experiments of visual exploration for object detection (a-b) and visually-guided reaching and grasping in the iCub's simulator (c-e), in two different scenarios with two detectors, and the real robot (f). The reddish blend shows the superimposed attentional landscape used to drive visual processing (for the peripersonal space-primed attention with the histogram-based detector (a) we are sampling from a modified version, computed as in Eg. 5.10 with $\pi = 0.2$). The figures (a) and (b) represent snapshots from the experiments where visual processing is prioritized to the peripersonal space (peripersonal space-primed attention), for histogram-based detector and SURF, respectively. The blue squares are scanning image windows for which visual features are computed. The robot adopts a random configuration, and the object adopts a random position within the reachable space. Figures (c-f) show the context of motor plans-primed attention, namely, the execution of eye-arm-hand coordination from the start of the task (left) until successful grasp completion (right). The white line corresponds to a mentally-simulated arm trajectory that is projected to the image planes of stereo cameras. Figure (b) corresponds to the obstacle scenario with histogram-based detector. Figure (c) corresponds to the obstacle scenario with histogram-based detector. Figure (d) corresponds to the no-obstacle scenario with SURF detector. The blue circles correspond to detected strong feature points. Figure (e) shows how a combination of both approaches: the peripersonal space-primed attention is used to bootstrap initialization of the motor plans-primed attention. The bottom row (f) corresponds to the no-obstacle scenario with histogram-based detector with the real robot.

these windows takes 26.5 ms. These times sum up to 69 ms for the peripersonal-space histogram based visual detection. We can see that with our approach we can save 99 ms for each pass through the control loop (speed up factor ∼2.4 ×). For SURF, thresholding takes 6.5 ms and processing 30 % of the image pixels with the highest salience takes 280 ms, which sums up to the total time of 322 ms for our approach. We can see that this saves 193.5 ms per pass (∼1.6 × faster). Figures 5.6(a-b) show the simulated results, and Tables 5.1 and 5.2 report times for the peripersonal space-primed attention with the histogram based detector and SURF, respectively.

For motor plans-primed attention, we use a similar approach; the only difference is that this, more specialized visual attention, is used to aid the ongoing movements. For both detectors, the common computations involve a projection of the mentally-simulated trajectory to the image plane and computing a motor-primed attentional landscape. The cumulative time for calculating a projection of the forward-planned trajectory to the image planes and computing attentional image landscapes is 19 ms (1 ms for projection and 18 ms for computation of the landscapes). For the histogram-based detector, sampling time for 50 windows is the same as in the peripersonal version, 7 ms, and similarly, the image processing time is 28 ms. The overall time for motor plans-primed histogram-based image processing is only 54 ms, i.e. ∼3.1 ×(114 ms) faster than the naive image processing with a uniformly sliding window. For motor-plans primed attention with SURF, again, thresholding requires 6.5 ms and processing 30 % of the image pixels of the most relevant pixels takes 281 ms. The total time for our approach with SURF is 306.5 ms, which is ∼1.7 ×faster than the classical, full-blown image processing. Figures 5.6(c-d) show the scenarios and Tables 5.3 and 5.4 report times for the with the histogram detector and SURF, respectively. Figure 5.6(f) presents the experiments with the motor plans-primed attention and the histogram-based detector with the real iCub robot.

The presented schemes could be used independently of each other, as previously discussed, and as shown here, however, they could work even better if used together. In order to plan the movements for actions (for estimation of future movements and for updating the visual scene by using visual processing driven by motor plans-primed attention), a robot must have some initial guess where the object might be. Of course, to initialize the procedure one could scan the entire images first and then in the further iterations apply reduced processing by utilizing the motor attention and updating the knowledge about the object state from the vision system. However, for this initial exploration, we could use the peripersonal space attention to constrain the initial visual search. Once the robot starts to move, it switches to the motor plans-primed mechanism. Figure 5.6(e) shows how these two attentional mechanisms work together.

Clearly, the presented experiments show that if we choose to intelligently process the images, prioritizing valuable image resources to motor relevant plans of the images, we can speed up visual computations by up to a factor of 3 times compared to the standard uniform image processing approach, where all pixels have the same priority and hence they are processed accordingly, without any discrimination what is motor relevant from what is not.

Finally, it is important to mention that, in addition to speeding up visual processing, this

Table 5.1: CPU time for peripersonal space-primed attention with histogram sliding-window detector

inferring attentional param. with neural nets	attentional landscape	sampling	sparse image processing	**total time with our approach**	**standard (full) image processing approach**	**savings with our approach**
~ 0	35.5	7	26.5	**69**	**168**	**99 (2.4 x)**

Table 5.2: CPU time for peripersonal space-primed attention with SURF

inferring attentional param. with neural nets	attentional landscape	thresholding	sparse image processing	**total time with our approach**	**standard (full) image processing approach**	**savings with our approach**
~ 0	35.5	6.5	280	**322**	**515.5**	**193.5 (1.6 x)**

Table 5.3: CPU time for motor plans-primed attention with histogram sliding-window detector

projecting trajectory with neural nets	attentional landscape	sampling	sparse image processing	**total time with our approach**	**standard (full) image processing approach**	**savings with our approach**
1	18	7	28	**54**	**168**	**114 (3.1 x)**

Table 5.4: CPU time for motor plans-primed attention with SURF

projecting trajectory with neural nets	attentional landscape	thresholding	sparse image processing	**total time with our approach**	**standard (full) image processing approach**	**savings with our approach**
1	18	6.5	281	**306.5**	**515.5**	**209 (1.7 x)**

CPU time required to compute the peripersonal space-primed attention with histogram-based image processing (Table 5.1) and SURF computation (Table 5.2) and motor plans-primed attention with histogram-based-image processing and SURF (Table 5.3) and SURF computation (Table 5.4), respectively. For all 4 tables, we compare our approach to the standard-uniform processing with the same detectors. We show that our motor-prioritization of visual processing has the potential to afford savings up to 3.1 times faster than the standard, naive image processing. The times in the tables above are provided in milliseconds, and correspond to the computation for both cameras of the binocular setup of the iCub robot with 320×240 color input images. The times are the averages computed over 200 passes.

approach facilitates the accuracy of visual detections during an ongoing prehensile movement. The common problem with visual detections in cluttered scenes (as the one in Figure 5.6 (f)) is that there could be a significant number of false positives after image processing is done. Because we bound visual processing to motor plans of a robot, we significantly reduce false positive detections. In the context of the conducted experiments, there are no false positive detections of the objects in the parts of the visual field that are irrelevant to motor plans, because the relevant objects are not likely to be there.

The computation times presented here are the averages computed for 200 measurements. The experiments are run on a computer with an Intel i7 2.7 GHz dual-core processor and 4 GB of RAM. We have included a supplementary video file which contains the experiments presented here. The video will be available online at `http://lasa.epfl.ch/~lukic/IEEE_Tran_2014.wmv`.

5.5 Summary and discussion

In this chapter, we have presented one general approach, with two different, but complementary, computational realizations, where visual attention is computed by using modulation signals originating from the robot's motor system. In sharp contrast to the classical approach in computational models of attention and corresponding robotic implementations, where visual saliency is computed based on low-level visual features such as color, edges and intensity contrast, emphasis is put here on tuning the robot vision with respect to the notion of the peripersonal space and forward-planned reaching and grasping movements.

The approach presented here is inspired by the results from psychology and visual neuroscience suggesting that visual attention emerges from the motor system, as elegantly summarized under the premotor theory of attention (Rizzolatti and Craighero, 2010). The peripersonal space around the body (in both humans and non-human primates) inherently attracts more visual resources than the extrapersonal (beyond reach) space, with and without supporting arm movements (Graziano and Gross, 1995; Previc, 1998). A number of more recent studies with humans show that the specialization in the peripersonal space could be additionally fine-tuned in order to support reaching and grasping movements (Baldauf and Deubel, 2010).

According to the aforementioned results from the psychology and neuroscience, we have developed two attentional techniques to drive visual processing in humanoid robots: peripersonal space-primed and motor plans-primed models of visual attention. Peripersonal space-primed attention is based on the idea that visual processing supporting reaching and grasping should prioritize the reachable (peripersonal) space of the robot. On the other hand, motor plans-primed attention is constructed around the idea that during movements, the image parts corresponding to the space around motor plans should receive higher priority for visual processing. The peripersonal space-primed attention model is a more general concept and could be used for a variety of applications, including visual exploration of the reachable space, but also during the ongoing movements, as well.

Nevertheless, we advocate its use for visual exploration, but not during actual movements, because motor plans-primed attention offers a more specialized framework, which results in higher computational savings. We have taken a machine learning, data-driven exploratory approach to construct the visuomotor transformations and to obtain an implicit notion of the peripersonal space used for guiding visual processing. The benefits of such an approach are that learned models be adapted, if needed, to the visuomotor transformations involving the imperfections of the kinematics and cameras of real robot, and that it overcomes limitations of the classical methods used for representation of the peripersonal space, while still being very efficient to compute (less than a millisecond to compute the outputs of feedforward neural networks). Once the attentional landscape is computed (either peripersonal space-primed or motor plans-primed) it could be used to drive almost any standard image processing technique. We have presented experiments with two popular techniques, with the histogram-based color detector and SURF. For the histogram-based detector, we treat the attentional landscape as the bivariate probability density function and sample locations of the scanning windows by using the Gibbs sampling technique. For SURF, we apply a threshold based segmentation to constrain computation of SURF features within the parts of the image with higher motor relevance.

Furthermore, in the presented experiments, we have shown how the peripersonal space-primed and motor plans-primed attention can work together. Peripersonal space-primed attention is used to bootstrap initialization of the motor plans-primed attentional mechanism. In order to use motor plans-primed attention, the robot first needs to possess some previous belief where the object might be. This prior information about the object location is used in an iterative procedure: to compute motor plans, which are used to control the robot and for visual updating of the object location by means of motor plans-driven visual processing. The initial guess where the object might be placed could be obtained by first scanning the entire stereo images in the classical way and then proceeding with the iterative procedure until the task ends. However, peripersonal space-primed attention offers a way to constrain the initial visual search, which is a more efficient method than the naive and expensive scanning of the whole images. Once the object to be grasped (and objects to be potentially avoided) is detected, the robot then selects its motor plans, and it switches its visual attentional mechanism to the motor plans-primed, more specialized and more efficient, attentional model that supports visual processing during movements.

Taken together, in this chapter, we have shown that our approach can efficiently distribute limited visual resources in a robot system, significantly reducing resources compared to the classical uniform image processing, but still allowing for a robot to perform complicated tasks, such as manipulation with obstacle avoidance.

6 Conclusion and Future Work

In this chapter, we make a summary of the main contributions of the thesis. We bring to light some limitations of the proposed work and propose several future directions for improvements that could be natural extensions of this dissertation.

6.1 Thesis contributions

There are four main contributions of this thesis:

First, we presented a novel behavioral experiment with humans, in which we investigated the visuomotor coordination in humans in complex motor tasks, such as prehension with obstacle avoidance lead by head-free gaze movements. Our study indicates that visually-guided reaching with obstacle avoidance is organized in a sequential manner, and that the visuomotor system treats the obstacle as an intermediary target, favoring movement segmenting instead of holistic task programming. Furthermore, we found that the forward planning mechanism might be proactively involved in guiding the motor system and detecting potential obstacles guiding reaching and grasping. We have extended the well-known fact that the gaze actively leads the arm-hand system, by showing that this coupling is preserved even in the presence of an obstacle.

Second, the observations from our human study provided the basis on which we developed a robotic eye-arm-hand controller. The controller is solely estimated by using the human motion capture data. The controller is based on our extension of the Coupled Dynamical Systems framework. The properties of this framework provide the model with the ability to rapidly generate stable coordinated movements and almost instantly reprogram movements when perturbations occur, mimicking the behavior of humans. This controller shares similar properties with classical visual servoing because the movements are generated in a closed-loop fashion. However, it is also related to learning-based visuomotor robotic models because it employs machine learning techniques to learn movement generation and motor coordination.

Third, we investigated the neuroscientific literature, focusing on the main computational principles behind the target encoding, programming and coordination of visuomotor movements. In our modeling, we emphasized the hypothesis that the cerebellum uses the cortical target encoding, and, based on this representation, performs closed-loop programming of multi-joint, compound movements and movement coordination between the eye-head system, arm and hand. We unified these considerations in the block-schematic model we proposed. In addition to our theoretical modeling, we provided a computational model for robotic visuomotor control in obstacle-free prehension. In obstacle-free reaching and grasping, our computational model offers a number of improvements over the first robotic model we have developed. The improvements are, namely, in terms of the computational efficiency (faster computation), introduced plasticity of the target representation (e.g., with this plasticity it is now possible to accomplish saccade adaptation) and the improved overall biological plausibility of our model (more biologically plausible target encoding for the gaze and arm and more plausible gaze-arm coupling). To the best of our knowledge, this model represents the only functional framework to unify, on a functional level of abstraction, the aforementioned computational and organizational principles borrowed from the neural motor control in the context of the full eye-arm-hand visuomotor control, both among robotic and neurophysiological models.

Fourth, we presented a new view on the modeling of the allocation of visual resources in the form of a motor-primed visual attentional landscape. This work was motivated by recent findings in human and monkey visual neuroscience and psychology. Spatially distributing visual attention in the form of the attentional landscape is a more general and a more complex concept than the attentional spotlight and zoom-lens paradigms. Attentional motor-priming prioritizes visual processing to motor-relevant parts of the visual field. Namely, we presented two models of motor-primed visual attention allocation. The first, more general, model devotes visual attention to the reachable space of a robot. The second, more specialized, technique allocates visual attention close to motor plans of the robot. Furthermore, we presented two methods for using the attentional landscape for driving visual processing. We showed that attentional motor-priming is a very efficient mechanism in terms of saving the limited resources for the visual computation.

6.2 Limitations and Future Work

6.2.1 Gaze fixation pattern at the target

In our robotic modeling, we selected the centroid of the object (obstacle and target) as the fixation point for the gaze. However, this simplified scheme of selecting the fixation points on the object might be upgraded in order to improve both biological plausibility and the computational benefits of using active vision. From physiological studies, it is known that the gaze fixations are driven to regions of the target contact points in grasping, whereas in viewing tasks the gaze is directed to the object's centroid (Brouwer et al., 2009). The explanation for this result is that fixations during grasping are focused on the object's contact parts because the eyes provide visual feedback for motor control of the fingers in grasping scenarios. These contact parts are mostly close to the boundary of an object. The gaze is more likely to fall on the edges of obstacles, in both manipulation tasks (Johansson et al., 2001) and in navigation (Rothkopf and Ballard, 2009), which can be explained by taking visual information for path planning for obstacle avoidance. We observed the same effects in our human trials. However, at this point there are not yet computational models that tackle problems of selecting optimal fixation points on the target object and obstacles. We consider that it would be tremendously useful to tackle this scientific problem. Recent work on active segmentation might offer the computational ground for tackling these problems (Mishra et al., 2009a,b).

6.2.2 Safety margin for obstacle avoidance

For the obstacle avoidance scheme we presented, we assumed a constant value for the safety margin between the arm's via point and the obstacle. The results of our human experiment, when the obstacle is moved along the midline of the desk, indicated that this safety distance was kept quasi-constant across subjects, and for all trials where the hand would have touched the obstacle if moving with the regular pattern of the motion. However, there is no reason to think that this safety margin is a constant, preset factor. Some studies showed that this safety margin was modulated by the speed of movement (e.g. faster prehensile movements are associated with a greater safety distance) (Tresilian, 1998; Mon-Williams et al., 2001) and "a variety of psychological factors related to the cost that a person attaches to a collision" (Tresilian, 1998). It would be of great importance, both for motor control science and robotic obstacle avoidance applications, to model this safety distance, rather than to consider it as a preset factor (Bendahan and Gorce, 2006). One approach to model this safety margin is to estimate it from the data recorded from human demonstrations by varying task conditions across trials (e.g. shape and size of an obstacle, relative positions of objects in the workspace, required speed of manipulation, task objectives, etc.), and then learn it by using suitable machine techniques.

6.2.3 More complex human obstacle avoidance studies

Robotic engineers have studied avoidance of multiple obstacles for a long time (Khatib, 1986; Lumelsky and Skewis, 1990; Simmons, 1996; Kavraki et al., 1996; Kuffner Jr and LaValle, 2000), but it is quite surprising that only a small number of studies in motor control, physiology and visual science studied human manipulation in tasks where several obstacles occupy the workspace. In their study, Mon-Williams et al. (2001) reported on the greater effect of two obstacles on the movement time, maximum grip aperture and peak speed compared to the one-obstacle case. Rothkopf and Ballard (2009), who studied human navigation in an immersed graphic environment, reported that subjects fixate the edges of obstacles for the purpose of planning a walking path for obstacle avoidance. Aivar et al. (2008) provided evidence that fast arm responses to the displacement of obstacles are triggered by a reaction to the retinal motion of moving obstacles. Many important questions still remain unanswered. Do humans assess multiple obstacles in a sequential manner, assigning priorities to obstacles according to the estimated risk of collision, or simultaneously? How are the eyes, arm and the hand coordinated when handling multiple obstacles in reaching and grasping tasks? How do the human visuomotor and planning systems react when one or several obstacles are perturbed in the workspace during prehensile tasks? Studying visuomotor coordination in natural prehensile tasks with several non-target objects in the workspace could provide more insights into these questions.

6.2.4 Flow of visuomotor coordination

In our robotic models, the flow of control signals is monodirectional, and it is oriented in the direction eyes → arm → hand. This is a current limitation, because there is a number of studies that have demonstrated that motor coupling and the reference frame transformations could be performed in the other direction, as well. A number of studies have demonstrated that the control signals also flow from the hand to the eyes (Vercher and Gauthier, 1988; Gauthier et al., 1988; Fisk and Goodale, 1985; Neggers and Bekkering, 2000), and from the hand to the arm (Timmann et al., 1996; Zackowski et al., 2002). Hence, it would be of primary interest to include this direction of visuomotor coordination, and assess its potential benefits over the monodirectional flow of control. Having the control signals flow in the opposite direction, hand → arm → eyes, could be useful, for instance, to trigger a reactive motion of the gaze and the arm when facing an unexpected displacement of the hand, such as when the hand inadvertently touches an obstacle.

6.2.5 More complex visuomotor coupling

Finally, in our modeling, we assumed that there is a single block that defines the coupling between each master and its corresponding slave effector. Considering the evidence from the psychological studies, this might be too restrictive. Several studies have shown that the profile of visuomotor coordination can be modulated depending on the task requirements (Vercher et al., 1994; Pelz et al., 2001; Hayhoe et al., 2003), which suggests the existence of either multiple coupling models or some parametric modulation inputs, descending from higher cortical areas such as the frontal lobe, that modulate motor coupling.

References

Abrams R, Meyer D, Kornblum S (1990) Eye-hand coordination: Oculomotor control in rapid aimed limb movements. Journal of Experimental Psychology: Human Perception and Performance 16(2):248

Abrams RA, Davoli CC, Du F, Knapp III WH, Paull D (2008) Altered vision near the hands. Cognition 107(3):1035–1047

Aivar M, Brenner E, Smeets J (2008) Avoiding moving obstacles. Experimental Brain Research 190(3):251–264

Alberts JL, Saling M, Stelmach GE (2002) Alterations in transport path differentially affect temporal and spatial movement parameters. Experimental Brain Research 143(4):417–425

Aloimonos J, Weiss I, Bandyopadhyay A (1988) Active vision. International Journal of Computer Vision 1(4):333–356

Andersen R, Asanuma C, Essick G, Siegel R (1990) Corticocortical connections of anatomically and physiologically defined subdivisions within the inferior parietal lobule. Journal of Comparative Neurology 296(1):65–113

Andersen RA, Buneo CA (2002) Intentional maps in posterior parietal cortex. Annual review of neuroscience 25(1):189–220

Andersen RA, Buneo CA (2003) Sensorimotor integration in posterior parietal cortex. Advances in Neurology 93:159–177

Andersen RA, Cui H (2009) Intention, action planning, and decision making in parietal-frontal circuits. Neuron 63(5):568–583

Andersen RA, Essick GK, Siegel RM (1985) Encoding of spatial location by posterior parietal neurons. Science 230(4724):456–458

Aryananda L (2006) Attending to learn and learning to attend for a social robot. In: IEEE-RAS International Conference on Humanoid Robots, IEEE, pp 618–623

Bajcsy R (1988) Active perception. Proceedings of the IEEE 76(8):966–1005

Bajcsy R, Campos M (1992) Active and exploratory perception. CVGIP: Image Understanding 56(1):31–40

Baldauf D, Deubel H (2008) Visual attention during the preparation of bimanual movements. Vision Research 48(4):549–563

Baldauf D, Deubel H (2009) Attentional selection of multiple goal positions before rapid hand movement sequences: An event-related potential study. Journal of Cognitive Neuroscience 21(1):18–29

Baldauf D, Deubel H (2010) Attentional landscapes in reaching and grasping. Vision Research 50(11):999–1013

Baldauf D, Wolf M, Deubel H (2006) Deployment of visual attention before sequences of goal-directed hand movements. Vision research 46(26):4355–4374

Ballard D (1991) Animate vision. Artificial Intelligence 48(1):57–86

Ballard DH, Hayhoe MM, Pelz JB (1995) Memory representations in natural tasks. Journal of Cognitive Neuroscience 7(1):66–80

Bastian A, Martin T, Keating J, Thach W, et al. (1996) Cerebellar ataxia: abnormal control of interaction torques across multiple joints. Journal of Neurophysiology 76(1):492–509

Batista AP, Buneo CA, Snyder LH, Andersen RA (1999) Reach plans in eye-centered coordinates. Science 285(5425):257–260

Bay H, Tuytelaars T, Van Gool L (2006) Surf: Speeded up robust features. Lecture Notes in Computer Science 3951:404–417

Becker W, Kunesch E, Freund H (1990) Coordination of a multi-joint movement in normal humans and in patients with cerebellar dysfunction. The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques 17(3):264–274

Becker W, Morrice B, Clark A, Lee R (1991) Multi-joint reaching movements and eye-hand tracking in cerebellar incoordination: investigation of a patient with complete loss of purkinje cells. The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques 18(4):476–487

Begum M, Karray F (2011) Visual attention for robotic cognition: a survey. IEEE Transactions on Autonomous Mental Development 3(1):92–105

Bekkering H, Adam JJ, van den Aarssen A, Kingma H, Whiting HJ (1995) Interference between saccadic eye and goal-directed hand movements. Experimental Brain Research 106(3):475–484

Bendahan P, Gorce P (2006) A neural network architecture to learn arm motion planning in grasping tasks with obstacle avoidance. Robotica 24(2):197–204

Bergeron A, Matsuo S, Guitton D (2003) Superior colliculus encodes distance to target, not saccade amplitude, in multi-step gaze shifts. Nature neuroscience 6(4):404–413

Bernardino A, Santos-Victor J (1999) Binocular tracking: integrating perception and control. IEEE Transactions on Robotics and Automation 15(6):1080–1094

Berthier NE, Clifton RK, Gullapalli V, McCall DD, Robin DJ (1996) Visual information and object size in the control of reaching. Journal of Motor Behavior 28(3):187–197

Beurze SM, Toni I, Pisella L, Medendorp WP (2010) Reference frames for reach planning in human parietofrontal cortex. Journal of neurophysiology 104(3):1736–1745

Bhattacharyya R, Musallam S, Andersen RA, et al. (2009) Parietal reach region encodes reach depth using retinal disparity and vergence angle signals. Journal of neurophysiology 102(2):805

Bishop C (2007) Pattern recognition and machine learning (information science and statistics). Pattern Recognition 4(2)

Blatt GJ, Andersen RA, Stoner GR (1990) Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area lip) in the macaque. Journal of Comparative Neurology 299(4):421–445

Blohm G, Khan A, Crawford J (2008) Spatial transformations for eye–hand coordination. New Encyclopedia of Neuroscience

Bonnefoi-Kyriacou B, Legallet E, Lee R, Trouche E (1998) Spatio-temporal and kinematic analysis of pointing movements performed by cerebellar patients with limb ataxia. Experimental brain research 119(4):460–466

Bowman M, Johannson R, Flanagan J (2009) Eye–hand coordination in a sequential target contact task. Experimental Brain Research 195(2):273–283

Bradski G, Kaehler A (2008) Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media, Incorporated

Breazeal C, Edsinger A, Fitzpatrick P, Scassellati B (2001) Active vision for sociable robots. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 31(5):443–453

Brochier T, Spinks RL, Umilta MA, Lemon RN (2004) Patterns of muscle activity underlying object-specific grasp by the macaque monkey. Journal of neurophysiology 92(3):1770–1782

Brouwer A, Franz V, Gegenfurtner K (2009) Differences in fixations between grasping and viewing objects. Journal of Vision 9(1)

Brown S, Kessler K, Hefter H, Cooke J, Freund HJ (1993) Role of the cerebellum in visuomotor coordination. Experimental Brain Research 94(3):478–488

Buneo CA, Andersen RA (2006) The posterior parietal cortex: sensorimotor interface for the planning and online control of visually guided movements. Neuropsychologia 44(13):2594–2606

Buneo CA, Jarvis MR, Batista AP, Andersen RA (2002) Direct visuomotor transformations for reaching. Nature 416(6881):632–636

Caminiti R, Johnson P, Galli C, Ferraina S, Burnod Y, Urbano A (1991) Making arm movements within different parts of space: the premotor and motor cortical representation of a coordinate system for reaching to visual targets. J Neurosci 11(5):1182–1197

Castiello U (2005) The neuroscience of grasping. Nature Reviews Neuroscience 6(9):726–736

Castiello U, Begliomini C (2008) The cortical control of visually guided grasping. The Neuroscientist 14(2):157–170

Castiello U, Bennett K, Mucignat C (1983) The reach to grasp movement of blind subjects. Experimental Brain Research 96(1):152–162

Castiello U, Bennett K, Stelmach G (1993) Reach to grasp: the natural response to perturbation of object size. Experimental Brain Research 94(1):163–178

Chaumette F, Hutchinson S (2008) Visual servoing and visual tracking. Handbook of Robotics pp 563–583

Chen-Harris H, Joiner WM, Ethier V, Zee DS, Shadmehr R (2008) Adaptive control of saccades via internal feedback. The Journal of Neuroscience 28(11):2804–2813

Churchland PS, Ramachandran V, Sejnowski TJ (1994) A critique of pure vision1. Large-scale neuronal theories of the brain p 23

Clower DM, Hoffman JM, Votaw JR, Faber TL, Woods RP, Alexander GE (1996) Role of posterior parietal cortex in the recalibration of visually guided reaching

Colby CL, Duhamel JR (1996) Spatial representations for action in parietal cortex. Cognitive Brain Research 5(1-2):105–115

Constantin A, Wang H, Martinez-Trujillo J, Crawford J (2007) Frames of reference for gaze saccades evoked during stimulation of lateral intraparietal cortex. Journal of neurophysiology 98(2):696–709

Constantin AG, Wang H, Crawford JD (2004) Role of superior colliculus in adaptive eye-head coordination during gaze shifts. Journal of neurophysiology 92(4):2168–2184

Cosman JD, Vecera SP (2010) Attention affects visual perceptual processing near the hand. Psychological Science 21(9):1254–1258

Cotti J, Vercher JL, Guillaume A (2011) Hand–eye coordination relies on extra-retinal signals: Evidence from reactive saccade adaptation. Behavioural brain research 218(1):248–252

Crawford JD, Medendorp WP, Marotta JJ (2004) Spatial transformations for eye-hand coordination. Journal of Neurophysiology 92:10–19

Crawford JD, Henriques DY, Medendorp WP (2011) Three-dimensional transformations for goal-directed action. Annual review of neuroscience 34:309–331

Culham JC, Cavina-Pratesi C, Singhal A (2006) The role of parietal cortex in visuomotor control: what have we learned from neuroimaging? Neuropsychologia 44(13):2668–2684

Dalton K, Nacewicz B, Johnstone T, Schaefer H, Gernsbacher M, Goldsmith H, Alexander A, Davidson R (2005) Gaze fixation and the neural circuitry of face processing in autism. Nature Neuroscience 8(4):519–526

Damas B, Santos-Victor J (2013) Online learning of single-and multivalued functions with an infinite mixture of linear experts. Neural computation 25(11):3044–3091

Dassonville P, Schlag J, Schlag-Rey M (1992) The frontal eye field provides the goal of saccadic eye movement. Experimental brain research 89(2):300–310

Davoli CC, Brockmole JR, Goujon A (2012) A bias to detail: how hand position modulates visual learning and visual memory. Memory & cognition 40(3):352–359

Dean J, Brüwer M (1994) Control of human arm movements in two dimensions: paths and joint control in avoiding simple linear obstacles. Experimental Brain Research 97(3):497–514

Desmurget M, Prablanc C (1997) Postural control of three-dimensional prehension movements. Journal of Neurophysiology 77(1):452–464

Desmurget M, Jordan M, Prablanc C, Jeannerod M, et al. (1997) Constrained and unconstrained movements involve different control strategies. Journal of Neurophysiology 77(3):1644–1650

Desmurget M, Pélisson D, Rossetti Y, Prablanc C (1998a) From eye to hand: planning goal-directed movements. Neuroscience & Biobehavioral Reviews 22(6):761–788

Desmurget M, Pélisson D, Urquizar C, Prablanc C, Alexander GE, Grafton ST (1998b) Functional anatomy of saccadic adaptation in humans. Nature neuroscience 1(6):524–8

Desmurget M, Pelisson D, Grethe J, Alexander G, Urquizar C, Prablanc C, Grafton S (2000) Functional adaptation of reactive saccades in humans: a pet study. Experimental brain research 132(2):243–259

DeSouza JF, Keith GP, Yan X, Blohm G, Wang H, Crawford JD (2011) Intrinsic reference frames of superior colliculus visuomotor receptive fields during head-unrestrained gaze shifts. The Journal of Neuroscience 31(50):18,313–18,326

Deubel H, Schneider WX (2004) Attentional selection in sequential manual movements, movements around an obstacle and in grasping. Attention in action pp 69–91

Deubel H, O'Regan K, Radach R (2000) Attention, information processing and eye movement control. Reading as a Perceptual Process pp 355–374

Doniec MW, Sun G, Scassellati B (2006) Active learning of joint attention. In: IEEE-RAS International Conference on Humanoid Robots, IEEE, pp 34–39

van Donkelaar P, Lee RG (1994) Interactions between the eye and hand motor systems: disruptions due to cerebellar dysfunction. Journal of Neurophysiology 72(4):1674–1685

Eimer M, Van Velzen J, Gherri E, Press C (2006) Manual response preparation and saccade programming are linked to attention shifts: Erp evidence for covert attentional orienting and spatially specific modulations of visual processing. Brain research 1105(1):7–19

Eriksen CW, James JDS (1986) Visual attention within and around the field of focal attention: A zoom lens model. Perception & Psychophysics 40(4):225–240

Espiau B, Chaumette F, Rives P (1992) A new approach to visual servoing in robotics. IEEE Transactions on Robotics and Automation 8(3):313–326

Fadiga L, Fogassi L, Gallese V, Rizzolatti G (2000) Visuomotor neurons: ambiguity of the discharge or 'motor' perception? International journal of psychophysiology 35(2):165–177

Fagg AH, Arbib MA (1998) Modeling parietal–premotor interactions in primate control of grasping. Neural Networks 11(7):1277–1303

Fernandez-Ruiz J, Goltz HC, DeSouza JF, Vilis T, Crawford JD (2007) Human parietal reach region primarily encodes intrinsic visual direction, not extrinsic movement direction, in a visual–motor dissociation task. Cerebral Cortex 17(10):2283–2292

Ferraina S, Paré M, Wurtz RH (2002) Comparison of cortico-cortical and cortico-collicular signals for the generation of saccadic eye movements. Journal of Neurophysiology 87(2):845–858

Findlay JM, Gilchrist ID (1998) Eye guidance and visual search. Eye Guidance in Reading and Scene Perception pp 295–312

Fisk J, Goodale M (1985) The organization of eye and limb movements during unrestricted reaching to targets in contralateral and ipsilateral visual space. Experimental Brain Research 60(1):159–178

Fogassi L, Gallese V, Di Pellegrino G, Fadiga L, Gentilucci M, Luppino G, Matelli M, Pedotti A, Rizzolatti G (1992) Space coding by premotor cortex. Experimental Brain Research 89(3):686–690

Freedman EG, Sparks DL (1997) Activity of cells in the deeper layers of the superior colliculus of the rhesus monkey: evidence for a gaze displacement command. Journal of neurophysiology 78(3):1669–1690

Frintrop S (2006) VOCUS: A visual attention system for object detection and goal-directed search, vol 3899. Springer

Galletti C, Kutz DF, Gamberini M, Breveglieri R, Fattori P (2003) Role of the medial parieto-occipital cortex in the control of reaching and grasping movements. Experimental Brain Research 153(2):158–170

Gauthier G, Vercher JL, Ivaldi FM, Marchetti E (1988) Oculo-manual tracking of visual targets: control learning, coordination control and coordination model. Experimental Brain Research 73(1):127–137

Geisler WS (2008) Visual perception and the statistical properties of natural scenes. Annu Rev Psychol 59:167–192

Gentilucci M, Toni I, Chieffi S, Pavesi G (1994) The role of proprioception in the control of prehension movements: a kinematic study in a peripherally deafferented patient and in normal subjects. Experimental Brain Research 99(3):483–500

Georgopoulos AP, Kettner RE, Schwartz AB (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. ii. coding of the direction of movement by a neuronal population. The Journal of Neuroscience 8(8):2928–2937

Gibson JJ (1950) The perception of the visual world. Houghton Mifflin

González-Alvarez C, Subramanian A, Pardhan S (2007) Reaching and grasping with restricted peripheral vision. Ophthalmic and Physiological Optics 27(3):265–274

Goodale MA (2011) Transforming vision into action. Vision Research 51(13):1567–1587

Goodale MA, Haffenden A (1998) Frames of reference for perception and action in the human visual system. Neuroscience & Biobehavioral Reviews 22(2):161–172

Grasso R, Prévost P, Ivanenko Y, Berthoz A, et al. (1998) Eye-head coordination for the steering of locomotion in humans: an anticipatory synergy. Neuroscience Letters 253(2):115–118

Graziano MS, Gross CG (1995) The representation of extrapersonal space: a possible role for bimodal, visual-tactile neurons. The cognitive neurosciences pp 1021–1034

Haggard P, Wing A (1991) Remote responses to perturbation in human prehension. Neuroscience Letters 122(1):103–108

Haggard P, Wing A (1995) Coordinated responses following mechanical perturbation of the arm during prehension. Experimental Brain Research 102(3):483–494

Haggard P, Wing A (1998) Coordination of hand aperture with the spatial path of hand transport. Experimental brain research 118(2):286–292

Hayhoe M, Ballard D (2005) Eye movements in natural behavior. Trends in Cognitive Sciences 9(4):188–194

Hayhoe M, Shrivastava A, Mruczek R, Pelz J (2003) Visual memory and motor planning in a natural task. Journal of Vision 3(1)

Haykin S (1998) Neural Networks: A Comprehensive Foundation (2nd Edition). Prentice Hall

He SQ, Dum RP, Strick PL (1993) Topographic organization of corticospinal projections from the frontal lobe: motor areas on the lateral surface of the hemisphere. The Journal of neuroscience 13(3):952–980

Henderson JM, Hollingworth A (1999) The role of fixation position in detecting scene changes across saccades. Psychological Science 10(5):438–443

Hesse C, Deubel H (2010) Effects of altered transport paths and intermediate movement goals on human grasp kinematics. Experimental Brain Research 201(1):93–109

Hesse C, Deubel H (2011) Efficient grasping requires attentional resources. Vision Research 51(11):1223–1231

Hicheur H, Berthoz A (2005) How do humans turn? head and body movements for the steering of locomotion. In: IEEE-RAS International Conference on Humanoid Robots (Humanoids), IEEE, pp 265–270

Hoffman JE, Subramaniam B (1995) The role of visual attention in saccadic eye movements. Perception & Psychophysics 57(6):787–795

Hoffmann H, Schenck W, Möller R (2005) Learning visuomotor transformations for gaze-control and grasping. Biological Cybernetics 93(2):119–130

Hulse M, McBrid S, Lee M (2009) Robotic hand-eye coordination without global reference: A biologically inspired learning scheme. In: IEEE International Conference on Development and Learning (ICDL), IEEE, pp 1–6

Inhoff A, Radach R (1998) Definition and computation of oculomotor measures in the study of cognitive processes. Eye Guidance in Reading and Scene Perception pp 29–54

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11):1254–1259

Jacob R, Karn K (2003) Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. Mind 2(3):4

Jakobson L, Goodale M (1991) Factors affecting higher-order movement planning: a kinematic analysis of human prehension. Experimental Brain Research 86(1):199–208

Jamone L, Natale L, Nori F, Metta G, Sandini G (2012) Autonomous online learning of reaching behavior in a humanoid robot. International Journal of Humanoid Robotics 9(03)

Jamone L, Damas B, Endo N, Santos-Victor J, Takanishi A (2013) Incremental development of multiple tool models for robotic reaching through autonomous exploration. Paladyn Journal of Behavioral Robotics pp 1–15

Javier Traver V, Bernardino A (2010) A review of log-polar imaging for visual perception in robotics. Robotics and Autonomous Systems 58(4):378–398

Jeannerod M (1984) The timing of natural prehension movements. Journal of Motor Behavior

Johansson R, Westling G, Bäckström A, Flanagan J (2001) Eye–hand coordination in object manipulation. The Journal of Neuroscience 21(17):6917–6932

Johansson RS, Flanagan JR, Johansson RS (2009) Sensory control of object manipulation. Sensorimotor Control of Grasping: Physiology and Pathophysiology Cambridge University Press, Cambridge pp 141–160

Johnson PB, Ferraina S, Bianchi L, Caminiti R (1996) Cortical networks for visual reaching: physiological and anatomical organization of frontal and parietal lobe arm regions. Cerebral Cortex 6(2):102–119

Jueptner M, Frith C, Brooks D, Frackowiak R, Passingham R (1997a) Anatomy of motor learning. ii. subcortical structures and learning by trial and error. Journal of neurophysiology 77(3):1325–1337

Jueptner M, Stephan K, Frith C, Brooks D, Frackowiak R, Passingham R (1997b) Anatomy of motor learning. i. frontal cortex and attention to action. Journal of neurophysiology 77(3):1313–1324

Kakei S, Hoffman DS, Strick PL (1999) Muscle and movement representations in the primary motor cortex. Science 285(5436):2136–2139

Kakei S, Hoffman DS, Strick PL (2001) Direction of action is represented in the ventral premotor cortex. Nature neuroscience 4(10):1020–1025

Kakei S, Hoffman DS, Strick PL (2003) Sensorimotor transformations in cortical motor areas. Neuroscience research 46(1):1–10

Kandel ER, Schwartz JH, Jessell TM, et al. (2000) Principles of neural science, vol 4. McGraw-Hill New York

Kavraki LE, Svestka P, Latombe JC, Overmars MH (1996) Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Transactions on Robotics and Automation 12(4):566–580

Kawato M (1999) Internal models for motor control and trajectory planning. Current opinion in neurobiology 9(6):718–727

Kawato M, Gomi H (1992a) The cerebellum and vor/okr learning models. Trends in neurosciences 15(11):445–453

Kawato M, Gomi H (1992b) A computational model of four regions of the cerebellum based on feedback-error learning. Biological cybernetics 68(2):95–103

Kestur S, Park MS, Sabarad J, Dantara D, Narayanan V, Chen Y, Khosla D (2012) Emulating mammalian vision on reconfigurable hardware. In: IEEE 20th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)

Kettner RE, Schwartz AB, Georgopoulos AP (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. iii. positional gradients and population coding of movement direction from various movement origins. The journal of Neuroscience 8(8):2938–2947

Khansari-Zadeh S, Billard A (2011) Learning stable nonlinear dynamical systems with gaussian mixture models. IEEE Transactions on Robotics 27(5):943–957

Khansari-Zadeh SM, Billard A (2012) A dynamical system approach to realtime obstacle avoidance. Autonomous Robots 32(4):433–454

Khatib O (1986) Real-time obstacle avoidance for manipulators and mobile robots. The International Journal of Robotics Research 5(1):90–98

Kim S, Shukla A, Billard A (2014) Catching objects in flight. IEEE Transactions on Robotics PP(99):1–17

Klier EM, Wang H, Crawford JD (2001) The superior colliculus encodes gaze commands in retinal coordinates. Nature neuroscience 4(6):627–632

Klier EM, Martinez-Trujillo JC, Pieter Medendorp W, Smith MA, Douglas Crawford J (2003a) Neural control of 3-d gaze shifts in the primate. Progress in brain research 142:109–124

Klier EM, Wang H, Crawford JD (2003b) Three-dimensional eye-head coordination is implemented downstream from the superior colliculus. Journal of Neurophysiology 89(5):2839–2853

Krauzlis RJ (2005) The control of voluntary eye movements: new perspectives. The Neuroscientist 11(2):124–137

Krauzlis RJ, Basso MA, Wurtz RH, et al. (2000) Discharge properties of neurons in the rostral superior colliculus of the monkey during smooth-pursuit eye movements. Journal of Neurophysiology 84(2):876–891

Krauzlis RJ, et al. (2004) Recasting the smooth pursuit eye movement system. Journal of neurophysiology 91(2):591–603

Kuffner Jr J, LaValle S (2000) Rrt-connect: An efficient approach to single-query path planning. In: IEEE International Conference on Robotics and Automation (ICRA), IEEE, vol 2, pp 995–1001

Kurata K (1991) Corticocortical inputs to the dorsal and ventral aspects of the premotor cortex of macaque monkeys. Neuroscience research 12(1):263–280

Kurata K, Hoffman DS (1994) Differential effects of muscimol microinjection into dorsal and ventral aspects of the premotor cortex of monkeys. Journal of Neurophysiology 71(3):1151–1164

Kurata K, Hoshi E (1999) Reacquisition deficits in prism adaptation after muscimol microinjection into the ventral premotor cortex of monkeys. Journal of neurophysiology 81(4):1927–1938

Lacquaniti F, Caminiti R (1998) Visuo-motor transformations for arm reaching. European Journal of Neuroscience 10:195–203

Lacquaniti F, Soechting J, Terzuolo S (1986) Path constraints on point-to-point arm movements in three-dimensional space. Neuroscience 17(2):313–324

Làdavas E (2002) Functional and dynamic properties of visual peripersonal space. Trends in cognitive sciences 6(1):17–22

Land M (1999) Motion and vision: why animals move their eyes. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology 185(4):341–352

Land M, Mennie N, Rusted J, et al. (1999) The roles of vision and eye movements in the control of activities of daily living. Perception 28(11):1311–1328

Land MF, Furneaux S (1997) The knowledge base of the oculomotor system. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 352(1358):1231–1239

Lang CE, Schieber MH (2003) Differential impairment of individuated finger movements in humans after damage to the motor cortex or the corticospinal tract. Journal of neurophysiology 90(2):1160–1170

Lazzari S, Vercher JL, Buizza A (1997) Manuo-ocular coordination in target tracking. i. a model simulating human performance. Biological cybernetics 77(4):257–266

Lefèvre P, Quaia C, Optican LM (1998) Distributed model of control of saccades by superior colliculus and cerebellum. Neural networks 11(7):1175–1190

Liversedge S, Findlay J (2000) Saccadic eye movements and cognition. Trends in Cognitive Sciences 4(1):6–14

Losier BJ, Klein RM (2004) Covert orienting within peripersonal and extrapersonal space: Young adults. Cognitive brain research 19(3):269–274

Lukic L, Santos-Victor J, Billard A (2012) Learning coupled dynamical systems from human demonstration for robotic eye-arm-hand coordination. In Proceedings of the IEEE-RAS International Conference on Humanoid Robots (Humanoids), Osaka, Japan

Lukic L, Santos-Victor J, Billard A (2014a) Learning robotic eye–arm–hand coordination from human demonstration: a coupled dynamical systems approach. Biological cybernetics 108(2):223–248

Lukic L, Santos-Victor J, Billard A (2014b) Learning robotic eye-arm-hand coordination from human demonstration: a coupled dynamical systems approach. Biological Cybernetics pp 1–26

Lumelsky V, Skewis T (1990) Incorporating range sensing in the robot navigation function. IEEE Transactions on Systems, Man and Cybernetics 20(5):1058–1069

Luppino G, Murata A, Govoni P, Matelli M (1999) Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas aip and vip) and the ventral premotor cortex (areas f5 and f4). Experimental Brain Research 128(1-2):181–187

Manfredi L, Maini ES, Dario P, Laschi C, Girard B, Tabareau N, Berthoz A (2006) Implementation of a neurophysiological model of saccadic eye movements on an anthropomorphic robotic head. In: IEEE-RAS International Conference on Humanoid Robots

Mansard N, Lopes M, Santos-Victor J, Chaumette F (2006) Jacobian learning methods for tasks sequencing in visual servoing. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 4284–4290

Maravita A, Iriki A (2004) Tools for the body (schema). Trends in cognitive sciences 8(2):79–86

Maringelli F, McCarthy J, Steed A, Slater M, Umilta C (2001) Shifting visuo-spatial attention in a virtual three-dimensional space. Cognitive Brain Research 10(3):317–322

Martinez-Trujillo JC, Medendorp WP, Wang H, Crawford JD (2004) Frames of reference for eye-head gaze commands in primate supplementary eye fields. Neuron 44(6):1057–1066

Mason CR, Miller LE, Baker JF, Houk JC (1998) Organization of reaching and grasping movements in the primate cerebellar nuclei as revealed by focal muscimol inactivations. Journal of neurophysiology 79(2):537–554

Matelli M, Luppino G (2001) Parietofrontal circuits for action and space perception in the macaque monkey. Neuroimage 14(1):S27–S32

Medendorp WP, Goltz HC, Vilis T, Crawford JD (2003) Gaze-centered updating of visual space in human parietal cortex. The journal of Neuroscience 23(15):6209–6214

Meeker D, Cao S, Burdick J, Andersen R (2002) Rapid plasticity in the parietal reach region demonstrated with a brain-computer interface. In: Soc Neurosci Abstr, vol 28

Metta G (2001) An attentional system for a humanoid robot exploiting space variant vision. Tech. rep., DTIC Document

Metta G, Gasteratos A, Sandini G (2004) Learning to track colored objects with log-polar vision. Mechatronics 14(9):989–1006

Metta G, Natale L, Nori F, Sandini G, Vernon D, Fadiga L, Von Hofsten C, Rosander K, Lopes M, Santos-Victor J, et al. (2010) The icub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks 23(8-9):1125–1134

Miall R, Reckess G (2002) The cerebellum and the timing of coordinated eye and hand tracking. Brain and cognition 48(1):212–226

Miall R, Weir D, Wolpert DM, Stein J (1993) Is the cerebellum a smith predictor? Journal of motor behavior 25(3):203–216

Miall R, Reckess G, Imamizu H (2001) The cerebellum coordinates eye and hand tracking movements. Nature neuroscience 4(6):638–644

Miall RC, Imamizu H, Miyauchi S (2000) Activation of the cerebellum in co-ordinated eye and hand tracking movements: an fmri study. Experimental Brain Research 135(1):22–33

Middleton FA, Strick PL (2000) Basal ganglia and cerebellar loops: motor and cognitive circuits. Brain Research Reviews 31(2):236–250

Mishra A, Aloimonos Y, Fah CL (2009a) Active segmentation with fixation. In: 12th International Conference on Computer Vision (ICCV), IEEE, pp 468–475

Mishra A, Aloimonos Y, Fermuller C (2009b) Active segmentation for robotics. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 3133–3139

Mon-Williams M, Tresilian J, Coppard V, Carson R (2001) The effect of obstacle position on reach-to-grasp movements. Experimental Brain Research 137(3):497–501

Monteon JA, Wang H, Martinez-Trujillo J, Crawford JD (2013) Frames of reference for eye–head gaze shifts evoked during frontal eye field stimulation. European Journal of Neuroscience 37(11):1754–1765

Montgomery DC, Runger GC (2010) Applied Statistics and Probability for Engineers. John Wiley & Sons

Murata A, Gallese V, Luppino G, Kaseda M, Sakata H (2000) Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip. Journal of neurophysiology 83(5):2580–2601

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press

Natale L, Metta G, Sandini G (2005) A developmental approach to grasping. In: Developmental Robotics AAAI Spring Symposium, vol 44

Natale L, Nori F, Sandini G, Metta G (2007) Learning precise 3d reaching in a humanoid robot. In: IEEE International Conference on Development and Learning (ICDL), IEEE, pp 324–329

Navalpakkam V, Itti L (2005) Modeling the influence of task on attention. Vision Research 45(2):205–231

Neggers S, Bekkering H (2000) Ocular gaze is anchored to the target of an ongoing pointing movement. Journal of Neurophysiology 83(2):639–651

Noris B, Keller J, Billard A (2010) A wearable gaze tracking system for children in unconstrained environments. Computer Vision and Image Understanding 115(4):476–486

Ogino M, Toichi H, Yoshikawa Y, Asada M (2006) Interaction rule learning with a human partner based on an imitation faculty with a simple visuo-motor mapping. Robotics and Autonomous Systems 54(5):414–418

Ohyama T, Nores WL, Murphy M, Mauk MD (2003) What the cerebellum computes. Trends in neurosciences 26(4):222–227

Olivier E, Davare M, Andres M, Fadiga L (2007) Precision grasping in humans: from motor control to cognition. Current opinion in neurobiology 17(6):644–648

Optican LM (2005) Sensorimotor transformation for visually guided saccades. Annals of the New York Academy of Sciences 1039(1):132–148

Orabona F, Metta G, Sandini G (2005) Object-based visual attention: a model for a behaving robot. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE

Osu R, Uno Y, Koike Y, Kawato M (1997) Possible explanations for trajectory curvature in multi-joint arm movements. Journal of Experimental Psychology: Human Perception and Performance 23(3):890

Paillard J (1982) The contribution of peripheral and central vision to visually guided reaching. Analysis of Visual Behavior, Ingle, Goodale, Mansfield (editors), Cambridge, MIT Press pp 367–385

Paré M, Wurtz RH, et al. (2001) Progression in neuronal processing for saccadic eye movements from parietal cortex area lip to superior colliculus. Journal of Neurophysiology 85(6):2545–2562

Pattacini U (2011) Modular cartesian controllers for humanoid robots: Design and implementation on the icub. PhD thesis, Ph. D. dissertation, RBCS, Italian Institute of Technology, Genova

Pattacini U, Nori F, Natale L, Metta G, Sandini G (2010) An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 1668–1674

Paulignan Y, Jeannerod M, MacKenzie C, Marteniuk R (1991a) Selective perturbation of visual input during prehension movements. 2. the effects of changing object size. Experimental Brain Research 87(2):407

Paulignan Y, Mackenzie C, Marteniuk R, Jeannerod M (1991b) Selective perturbation of visual input during prehension movements. 1. the effects of changing object position. Experimental Brain Research 83(3):502–512

Paulin MG (1993) The role of the cerebellum in motor control and perception. Brain, behavior and evolution 41(1):39–50

Pelisson D, Prablanc C, Goodale M, Jeannerod M (1986) Visual control of reaching movements without vision of the limb. Experimental Brain Research 62(2):303–311

Pelz J, Hayhoe M, Loeber R (2001) The coordination of eye, head, and hand movements in a natural task. Experimental Brain Research 139(3):266–277

Posner MI, Snyder CR, Davidson BJ, et al. (1980) Attention and the detection of signals. Journal of Experimental Psychology 109(2):160–174

Prablanc C, Echallier J, Komilis E, Jeannerod M (1979) Optimal response of eye and hand motor systems in pointing at a visual target. Biological Cybernetics 35(2):113–124

Previc FH (1998) The neuropsychology of 3-d space. Psychological bulletin 124(2):123

Purdy KA, Lederman SJ, Klatzky RL (1999) Manipulation with no or partial vision. Journal of Experimental Psychology: Human Perception and Performance 25(3):755

Quaia C, Lefèvre P, Optican LM (1999) Model of the control of saccades by superior colliculus and cerebellum. Journal of Neurophysiology 82(2):999–1018

Rand M, Shimansky Y, Stelmach G, Bracha V, Bloedel J (2000) Effects of accuracy constraints on reach-to-grasp movements in cerebellar patients. Experimental brain research 135(2):179–188

Raos V, Umiltá MA, Gallese V, Fogassi L (2004) Functional properties of grasping-related neurons in the dorsal premotor area f2 of the macaque monkey. Journal of neurophysiology 92(4):1990–2002

Rayner K (1998) Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124(3):372

Reed CL, Grubb JD, Steele C (2006) Hands up: attentional prioritization of space near the hand. Journal of Experimental Psychology: Human Perception and Performance 32(1):166

Reinagel P, Zador AM (1999) Natural scene statistics at the centre of gaze. Network: Computation in Neural Systems 10(4):341–350

Rizzolatti G, Craighero L (2010) Premotor theory of attention. Scholarpedia 5(1):6311

Rizzolatti G, Luppino G (2001) The cortical motor system. Neuron 31(6):889–901

Rizzolatti G, Camarda R, Fogassi L, Gentilucci M, Luppino G, Matelli M (1988) Functional organization of inferior area 6 in the macaque monkey. Experimental Brain Research 71(3):491–507

Rizzolatti G, Fogassi L, Gallese V (1997) Parietal cortex: from sight to action. Current Opinion in Neurobiology 7(4):562–567

Robinson FR (2000) Role of the cerebellar posterior interpositus nucleus in saccades i. effect of temporary lesions. Journal of Neurophysiology 84(3):1289–1302

Robinson FR, Fuchs AF (2001) The role of the cerebellum in voluntary eye movements. Annual review of neuroscience 24(1):981–1004

Rosenbaum DA, Loukopoulos LD, Meulenbroek RG, Vaughan J, Engelbrecht SE (1995) Planning reaches by evaluating stored postures. Psychological review 102(1):28

Rossetti Y, Stelmach G, Desmurget M, Prablanc C, Jeannerod M (1994) The effect of viewing the static hand prior to movement onset on pointing kinematics and variability. Experimental Brain Research 101(2):323–330

Rothkopf C, Ballard D (2009) Image statistics at the point of gaze during human navigation. Visual Neuroscience 26(01):81–92

Rothkopf C, Ballard D, Hayhoe M (2007) Task and context determine where you look. Journal of Vision 7(14)

Rushworth MF, Ellison A, Walsh V (2001) Complementary localization and lateralization of orienting and motor attention. Nature neuroscience 4(6):656–661

Russo G, Bruce C (1993) Effect of eye position within the orbit on electrically elicited saccadic eye movements: a comparison of the macaque monkey's frontal and supplementary eye fields. Journal of neurophysiology 69(3):800

Russo GS, Bruce CJ (1996) Neurons in the supplementary eye field of rhesus monkeys code visual targets and saccadic eye movements in an oculocentric coordinate system. Journal of neurophysiology 76(2):825–848

Russo GS, Bruce CJ, et al. (2000) Supplementary eye field: representation of saccades and relationship between neural response fields and elicited eye movements. Journal of neurophysiology 84(5):2605–2621

Saeb S, Weber C, Triesch J (2011) Learning the optimal control of coordinated eye and head movements. PLoS computational biology 7(11):e1002,253

Sahbani A, El-Khoury S, Bidaud P (2012) An overview of 3d object grasp synthesis algorithms. Robotics and Autonomous Systems 60(3):326–336

Sakata H, Taira M, Murata A, Mine S (1995) Neural mechanisms of visual guidance of hand action in the parietal cortex of the monkey. Cerebral Cortex 5(5):429–438

Saling M, Alberts J, Stelmach G, Bloedel J (1998) Reach-to-grasp movements during obstacle avoidance. Experimental Brain Research 118(2):251–258

Sandini G, Metta G, Vernon D (2004) Robotcub: An open framework for research in embodied cognition. In: IEEE/RAS International Conference on Humanoid Robots, IEEE, vol 1, pp 13–32

Sandini G, Metta G, Vernon D (2007) The icub cognitive humanoid robot: An open-system research platform for enactive cognition. In: 50 years of artificial intelligence, Springer Berlin Heidelberg, pp 358–369

Sasaki K, Gemba H (1986) Effects of premotor cortex cooling upon visually initiated hand movements in the monkey. Brain research 374(2):278–286

Schiegg A, Deubel H, Schneider W (2003) Attentional selection during preparation of prehension movements. Visual Cognition 10(4):409–431

Schwartz AB, Kettner RE, Georgopoulos AP (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. i. relations between single cell discharge and direction of movement. The Journal of Neuroscience 8(8):2913–2927

Scott SH (2003) The role of primary motor cortex in goal-directed movements: insights from neurophysiological studies on non-human primates. Current Opinion in Neurobiology 13(6):671–677

Scudder CA, Kaneko CR, Fuchs AF (2002) The brainstem burst generator for saccadic eye movements. Experimental Brain Research 142(4):439–462

Seara JF, Strobl KH, Schmidt G (2003) Path-dependent gaze control for obstacle avoidance in vision guided humanoid walking. In: IEEE International Conference on Robotics and Automation (ICRA), IEEE, vol 1, pp 887–892

Shadmehr R, Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. The Journal of Neuroscience 14(5):3208–3224

Shukla A, Billard A (2011) Coupled dynamical system based arm–hand grasping model for learning fast adaptation strategies. Robotics and Autonomous Systems 60(3):424–440

Simmons R (1996) The curvature-velocity method for local obstacle avoidance. In: IEEE International Conference on Robotics and Automation (ICRA), IEEE, vol 4, pp 3375–3382

Sivak B, MacKenzie CL (1990) Integration of visual information and motor output in reaching and grasping: the contributions of peripheral and central vision. Neuropsychologia 28(10):1095–1116

Soechting J, Lacquaniti F (1981) Invariant characteristics of a pointing movement in man. The Journal of Neuroscience 1(7):710–720

Soechting J, Lacquaniti F (1983) Modification of trajectory of a pointing movement in response to a change in target location. J Neurophysiol 49(2):548–564

Sparks D, Freedman E, Chen L, Gandhi N (2001) Cortical and subcortical contributions to coordinated eye and head movements. Vision research 41(25):3295–3305

Spijkers WA, Lochner P (1994) Partial visual feedback and spatial end-point accuracy of discrete aiming movements. Journal of Motor Behavior 26(3):283–295

Srinivasa SS, Berenson D, Cakmak M, Collet A, Dogar MR, Dragan AD, Knepper RA, Niemueller T, Strabala K, Vande Weghe M, et al. (2012) Herb 2.0: Lessons learned from developing a mobile manipulator for the home. Proceedings of the IEEE 100(8):2410–2428

Stein J (1986) Role of the cerebellum in the visual guidance of movement. Nature 323(6085):217–221

Sung HG (2004) Gaussian mixture regression and classification. PhD thesis, Rice University

Tatler BW, Hayhoe MM, Land MF, Ballard DH (2011) Eye guidance in natural vision: Reinterpreting salience. Journal of Vision 11(5)

Thach WT (1998a) A role for the cerebellum in learning movement coordination. Neurobiology of learning and memory 70(1-2):177–88

Thach WT (1998b) What is the role of the cerebellum in motor learning and cognition? Trends in cognitive sciences 2(9):331–337

Thach WT, Goodkin H, Keating J (1992) The cerebellum and the adaptive coordination of movement. Annual review of neuroscience 15(1):403–442

Tian J, Ethier V, Shadmehr R, Fujita M, Zee DS (2009) Some perspectives on saccade adaptation. Annals of the New York Academy of Sciences 1164(1):166–172

Timmann D, Stelmach G, Bloedel J (1996) Grasping component alterations and limb transport. Experimental Brain Research 108(3):486–492

Timmann D, Watts S, Hore J (1999) Failure of cerebellar patients to time finger opening precisely causes ball high-low inaccuracy in overarm throws. Journal of neurophysiology 82(1):103–114

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cognitive Psychology 12(1):97–136

Tresilian J (1998) Attention in action or obstruction of movement? a kinematic analysis of avoidance behavior in prehension. Experimental Brain Research 120(3):352–368

Triesch J, Ballard DH, Hayhoe MM, Sullivan BT (2003) What you see is what you need. Journal of Vision 3(1)

Tu TA, Keating EG (2000) Electrical stimulation of the frontal eye field in a monkey produces combined eye and head movements. Journal of neurophysiology 84(2):1103–1106

Tweed D (1997) Three-dimensional model of the human eye-head saccadic system. Journal of Neurophysiology 77(2):654–666

Vercher J, Magenes G, Prablanc C, Gauthier G (1994) Eye-head-hand coordination in pointing at visual targets: spatial and temporal analysis. Experimental Brain Research 99(3):507–523

Vercher JL, Gauthier G (1988) Cerebellar involvement in the coordination control of the oculo-manual tracking system: effects of cerebellar dentate nucleus lesion. Experimental Brain Research 73(1):155–166

Vernon D, Hofsten C, Fadiga L (2010) A roadmap for cognitive development in humanoid robots, vol 11. Springer

Vesia M, Crawford JD (2012) Specialization of reach function in human posterior parietal cortex. Experimental brain research 221(1):1–18

Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, vol 1, pp I–511

Weiss PH, Marshall JC, Wunderlich G, Tellmann L, Halligan PW, Freund HJ, Zilles K, Fink GR (2000) Neural consequences of acting in near versus far space: a physiological basis for clinical dissociations. Brain 123(12):2531–2541

Werner JS, Chalupa LM (2004) The visual neurosciences. Bradford Book

Wilson M (2002) Six views of embodied cognition. Psychonomic bulletin & review 9(4):625–636

Wolfe JM (1998) What can 1 million trials tell us about visual search? Psychological Science 9(1):33–39

Wolpert D, Miall R, Kawato M (1998) Internal models in the cerebellum. Trends in Cognitive Sciences 2(9):338–347

Wolpert D, Flanagan J, et al. (2001) Motor prediction. Current Biology 11(18):729

Wolpert DM, Ghahramani Z (2000) Computational principles of movement neuroscience. nature neuroscience 3:1212–1217

Xu-Wilson M, Chen-Harris H, Zee DS, Shadmehr R (2009) Cerebellar contributions to adaptive control of saccades in humans. The Journal of Neuroscience 29(41):12,930–12,939

Zackowski K, Thach Jr W, Bastian A (2002) Cerebellar subjects show impaired coupling of reach and grasp movements. Experimental brain research 146(4):511–522