Università degli Studi di Napoli
"Federico II"

Dottorato in
Fisica Fondamentale ed Applicata
20° ciclo

Sensorimotor Coupling: Eye Tracking and Computational Modeling in a Drawing Task

Ruben Coen Cagli

Coordinatore
Prof. Gennaro Miele

anno accademico 2006 – 2007

to my son Leon,
who is learning how to coordinate
his eyes with his hands.

Preface

The interdisciplinary nature of the framework in which this thesis was developed is evident already when reading the title. For a physicist, the aim is that of understanding the processes and mechanisms underlying natural phenomena; however the class of phenomena discussed in this book possess a 'special' feature, they pertain to living and intelligent organisms. Such phenomena are what is usually called *behaviors*.
This has been, historically, the reason why cybernetics was populated by experts of so many different disciplines, and its contemporary descendants still put together psychologists, mathematicians, engineers, physicists, neurologists, philosophers, and much more. This thesis makes no exception; I had the luck to interact with researchers from at least five disciplines, which made this project much more interesting than it could have been otherwise. Furthermore, much work was developed in tight collaboration with those people, and this (rather than just by mere convention) is the reason why in the rest of the book I will use the pronoun 'we' instead of 'I'.

Acknowledgements

First of all, I want to express my gratitude to Giuseppe Trautteur, who trusted me so much as to accept me as a doctoral student 3 years ago, when I had really little knowledge of the field (not to mention the possible contents of the thesis). The enlightening discussions we had, were far the most formative experience of my PhD program. A special thank is reserved to Paolo Coraggio who introduced me to Paolo Napoletano who introduced me to Giuseppe Boccignone who introduced me to Bayes (not in person, yet!). Working side by side with them produced the real essence of the thesis, as acknowledged

also by the publications we have together. The experimental work was made possible by Angelo Marcelli, to whom I express all my gratitude, who gave us the opportunity to work with the eye tracker. The most recent (and in progress yet) extensions of this project were made possible also by the important contribution of Agostino De Santis, who contributed with his experience in human–like robotic arm control. I also had to luck to discuss some of the ideas and problems connected to the subject of this thesis with, and got invaluable suggestions from, Guglielmo Tamburrini, Matteo Santoro, Roberto Casati and Alessandro Pignocchi, Paolo Viviani, John Tchalenko, Frederick Fol Le Mairie and Patrick Tresset, Aaron Kozbelt, Dana Ballard.

Ruben Coen Cagli
Napoli, 11.2007

Contents

Introduction

Cognitive science, and in particular the analysis of human vision and visual attention, have always payed some attention to the visual arts. In fact, this field has provided a rich source of images that are situated somewhere between *natural* images — such as pictures and videos of landscapes, animals, humans — and *synthetic* images — the kind of visual displays realized specifically for the purpose of testing some visual behavior. Visual artworks share some properties with both classes, because a) they possess some degree of artificiality, being images produced by humans and therefore possessing the kind of features that have been called *artifactual properties* [88]; and b) they are as common in our visual experience as natural images, since we are exposed to visual artworks very often in daily life (think of musems, books covers and illustrations, advertisement). Furthermore, drawing is an old practice (the oldest cave graffities dating back to about 30,000 years ago), and, orthogonally, they are present in almost all geographical and cultural areas.

The pioneeristic recordings of eye movements reported by Buswell [16] and Yarbus [136] used famous paintings as the test image; many later examples exist of analyses of the visual activity in response to paintings and drawings, and recently visual artworks have been used as well in neuroscientific studies of the visual brain. The groundbreaking work by Zeki [138] proposed even a step further: to not only use artworks to probe human vision, but, the other way around, to get an understanding of the aesthetic experience on the basis of our knowledge of the neural processes involved in perceiving artworks. Although such a reductionist approach has been strongly debated and criticized it was in part successful, and so much appealing that it opened an entire novel field of studies, now termed *neuroaesthetics*. Many extensions have been proposed, including the adoption of novel neurophysiological knowledge to better define this newborn branch of aesthetics (e.g. the recent proposal by Gallese and Freedberg [39]

that aesthetical experience would be based on the kind of empathetic responses enabled by mirror–like mechanisms [38]).

Although the work presented in this thesis is *not* directly related to neuroaesthetics, it took its first moves from a critical reading of Zeki's opus and related work, in particular from a methodological flaw that has been somehow overlooked: it has been argued in [139] that understanding the neural correlates of artwork perception could give not only some insight on aesthetic experience, but also a deep understanding of the brain processes involved in artwork creation — by comparing the work of the artist to that of the neurologist, whose ultimate aim is to induce specific neural activity (and therefore specific perceptual effects) in the viewer.
Our point here is that this 'inverse' approach cannot be just taken for granted, because of two reasons at least. First, the perceptual experience that the artist has of her own artwork is undoubtedly biased by her experience of the whole process that led to the result. Against this objection it could be argued that at least the so called *Early* visual analysis, which is thought to be cognitively *impenetrable* [93], should be common to the artist and the perceiver. It is well known however that even the perception of basic features such as orientation (which is processed in early visual steps) can be biased by the temporal context, leading to so called perceptual *after–effect* illusions [108]; the prolonged exposure that the artist has during all the intermediate stages of image creation could give rise to some kind of long term after–effects. Furthermore, perception of a visual scene is known to be determined also by how *overt* attention (namely eye movements) is deployed, and it is plausible to think that the attention of the artist will be directed to regions of the image that were critical during the creative process, and that are not necessarily the same that perceivers will attend to.
The second, more general reason, is that nothing guarantees that the *perceptual* process and the *creative* process share the same neural mechanisms. At least two different positions could be taken here. A) according to the most recent formulation of the *dual vison* theory [74], two separate pathways of visual processing exist, one implementing perceptual functions (e.g. object recognition) and ultimately delivering to us a coherent visual experience of the external world, and the second one subserving the control of motor actions. According to this view, it could be argued that mainly the *Vision for Perception* pathway is involved during artwork perception, while the *Vision for*

Action stream is the one upon which the visual creative process relies. B) following the *sensorimotor* approach to perceptual experience [85], and even more profoundly in view of the existence of mirror neurons [72], it could be argued that the perception of a visual artwork involves an internal simulation of the actions that produced that specific image, which would provide a common ground to artwork creation and artwork perception.
In the present thesis we do not commit to any of the above mentioned positions; however, we believe that a grounded, 'direct' analysis of the creative process itself would be a fundamental contribution to the scientific debate on visual creativity. In addition, creative processes can be regarded, from the vantage point of Cognitive science, as a goal-directed activity involving several human skills and abilities: sensorimotor coordination, evaluation and decision, memory and emotion. In this perspective, we surmise that the analysis of the creative process by scientific means can prove itself a powerful methodology for the understanding of human capabilities such as those mentioned above, at least as much as the analysis of visual artwork perception has proven fruitful for the understanding of human vision.

In order to narrow down our field of analysis, we have focused on sensorimotor coordination, namely the problem of how sensory and motor resources are integrated to give rise to efficient behaviors for the solution of specific tasks. In particular, as explained below, our analysis has concentrated on eye–hand coordination in the task of performing a *realistic drawing from life*, namely copying an *original image* on an initially blank *canvas*, trying to reproduce image contours as faithfully as possible.

In **chapter 1** we propose an extensive survey of the existing literature on eye movements, visuomotor coordination and motor control. The problem of eye-hand coordination in performing a given task is considered [6] a paradigmatic one with respect to the more general question of sensorimotor integration, which in turn is reputed to be a crucial issue both for designing situated artificial agents and for the investigation about the underlying cognitive mechanisms in biological agents. Recent approaches to sensorimotor coordination in primates claim that motor preparation has a direct influence on subsequent eye movements [110], sometimes turning coordination into competition. Complementary, eye movements come into play in generating motor plans, as suggested by the existence of *look ahead* fixations in

many natural tasks [63].
Differently from the problem of modeling eye movements in purely visual tasks, dealing with visuomotor tasks requires a shift of perspective: the main difference in such cases is that eye movements should not be treated as entirely independent from movements of other parts of the body. In fact, it is the basic tenet of Active Vision [5] that eye movements depend on the task at hand, and if the task is a sensorimotor one, it is reasonable to expect a dependence on body movements as well.

On these premises, in **chapter 2** we first outline a functional account of the kind of sensorimotor processing involved in a generic visuomanual task, overtly inspired by the functional organization of the primate brain areas involved in sensory and motor processing. Then, with the aim of providing in a principled way a computational theory of the underlying processes, we conjecture that such model could be formalized in terms of a novel type of Dynamic Bayesian Network [77] (DBN), which we denoted the *Input–Output Coupled Hidden Markov Model* (IOCHMM). It is worth remarking that considerations about noise in motor and perceptual neural signals form the main reason for the widespread diffusion of probabilistic techniques in modeling sensorimotor behaviors in humans and animals. Furthermore, probabilistic graphical models together with Bayesian Decision Theory are a rich tool not only for modeling biological systems [58] (the inverse problem, fitting the data), but also for controlling artificial agents [4] (the direct problem, generating/simulating the data).
With respect to previous work, the proposed IOCHMM provides a general high level mechanism for the dynamic integration of eye and hand motor plans, and enables the use of information coming from multiple sensory modalities. It also accounts for the task–dependency of eye and hand plans, by learning a sensorimotor mapping that is suitable for the given task.
Chapter 3 concludes with a detailed mathematical account of how inference, decision, and learning can be implemented in the IOCHMM; such an account provides an extension of existing algorithms for HMM's to a special case that, to the best of our knowledge, had never been treated before in the literature on sensorimotor behaviors.

The following **chapters 3** and **4** specialize the present thesis to our case study: sensorimotor coordination in the drawing task. The choice of the drawing task proved helpful: since copying an original image

on a white canvas requires a quite regular alternation of eye and hand movements, this task provides a good example of the re–entrant influence between active vision and motor planning/control; at the same time, it produces quite regular observable behaviors, and this intuition is at the basis of both our experiments and the implementation of the proposed computational model.

In **chapter 3**, as a starting point to characterize the visuomotor strategies adopted in the drawing task, we propose some hypotheses that try to capture the essential features that distinguish drawing from other tasks, both with respect to the a priori requirements and the observed behavior. The rest of the chapter is a detailed presentation of our eye–tracking experiments, whose aims were to test the correctness of our hypotheses, as well as their implications for the observable sensorimotor behavior. The data collected in the experiments also informed the implementation of our computational model. Although many complex, natural activities have been studied in the framework of Active Vision, to the best of our knowledge only very few experiments on drawing tasks have been reported in the existing literature, with just one experimental team focusing on gaze behavior [123, 46]; therefore the experiments presented here, and the principled analysis paralleled with a computational model, represent a novelty in the panorama of eye–hand movement research.

Eventually, in **chapter 4** we discuss the details of the implementation of our computational model in the drawing task, and provide simulation results along with qualitative and quantitative comparisons with behavioral data. Notwithstanding the simplifying assumptions that are at the basis of the first implementation presented here, our results have proved successful in modeling the observed oculomotor behavior (specially when compared with other existing influential models); furthermore, the results obtained indicate that the approach proposed here represents a promising perspective for the sensorimotor control of a situated artificial agent.

Chapter 1

Active Vision

1.1 Passive Vision

Understanding of human vision has been dominated in the last decades by the approach originally proposed by David Marr [68], whose fundamental idea is that the visual brain's function is to build an accurate internal *representation* of the visual scene, much in the vein of traditional Artificial Intelligence. This approach is usually termed *Passive Vision*, as it underemphasizes, or even discards, the contribution of active eye movements to the collection of processes that we call Vision.

This paradigm can be recognized at many different level of the analysis. Most psychophysical methods rely on determining sensitivity thresholds, and to this end adopt tachistoscopic displays that do not leave sufficient time for eye movement; mathematical descriptions of the retinal stimulation have been provided in terms of Fourier series of sine–wave patterns; physiological studies have concentrated on determining the specific contribution of single cells to the internal representation; and computer science has provided computational models of human vision based on parallel processes operated on a static image, to produce the final inner representation upon which higher order processes run.

In the rest of this section the results of this approach are analyzed, giving an overview of the main anatomical and physiological discoveries about the visual brain, and the resulting *standard* computational model of the early stages of human vision.

1.1.1 Early Vision: Neuroanatomy and the 'standard' computational model

The early stages of the primate visual system, from the retinal activity up to projections to the visual cortex, have been studied long since with different techniques. The "wiring" can be studied using stains that carry from cell to cell; the response of individual cells can be studied by displaying patterns and recording the electrical behavior of the cell, and typically, neurons in the visual pathway are discussed in terms of their receptive field (RF), namely a record of the spatial distribution of the effect of illumination on the neuron's output; finally, some structural information can be elicited using psychophysical experiments.

Evidence from all the levels of experimental analysis indicates that the visual pathway can be profitably schematized in subsequent stages as follows [37]:

- The *retina*, where photoreceptive cells transduce impinging light to electrical spikes. These signals are processed by subsequent layers of cells, with the retinal ganglion cells connecting to the final layer.

- The *optic nerve* consists of the fibers of the *retinal ganglion cells*, and connects the retina to the brain through the the *optic chiasma* where the left-hand side of each retina is connected to the left half of the brain, and the right-hand side to the right half.

- Two main pathways can be identified at this stage; most connections go to the *Lateral Geniculate Nucleus* (LGN), but there are several secondary paths among which the predominant one projects to the *Superior Colliculus* (SC, see section 1.4.3).

- The LGN is connected to the *Visual Cortex*, one of the regions most studied in the primate brain. The visual cortex consists of a series of quite well defined layers, which carry out specialized computations. Much of what is commonly called early vision occurs in this structure, whose outputs are interpreted as a large selection of different representations of an image.

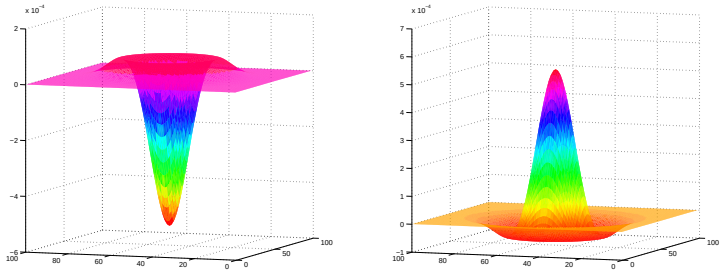
- Visual information leaves the visual cortex for the *Posterior Parietal Cortex* (PPC) and the *Infero–Temporal cortex* (IT).

The functionalities implemented along these two pathways are the object of a long standing debate, and are discussed in more detail in section 1.3.2.

Understanding the functional specialization of visual brain areas was a major scientific achievement [138]. The hierarchy itemized above is primarily motivated as a descriptive classification of increasingly complex functions. The main experimental support for this hierarchical structure is the well studied systematics of laminar connections [30]. Complementary, cells' response recordings show a general tendency of the RF's to increase their size along the hierarchy, consistently with the specialization of function; the most striking example are cells tuned to respond almost only to specific visual categories (e.g. faces, letters, animals, ...) irrespective of their location in the retinal image [59].

Retinal Cells and the Lateral Geniculate Nucleus

Retinal Ganglion Cells collect responses from the retinal photoreceptors. They are usually classified as *on–center, off–surround*, when their response (defined as the mean firing rate) is increased by light impinging within the RF and decreased by light falling off the RF; in the opposite case, they are called *off–center, on–surround*. Fig. 1.1 depicts a model of the spatial response as a Difference of Gaussians (DoG). Following the main visual path, signals from the retinal cells



(a) (b)

Figure 1.1: The response of a retinal ganglion cell can be predicted by adding the temporal response of the center to the temporal response of the surround. The model of the spatial response is a Difference of Gaussian model — there is a center field that has a spatial sensitivity of the form of a narrow Gaussian, and a surround field that has the form of a broad Gaussian. One field excites, the other inhibits.

travel through the optic nerve up to the LGN, whose neurons display similar receptive field effects to retinal cells. The LGN is a layered structure, with feedforward and feedback connections to and from higher level areas, including visual cortex. The functional role, apart from being a hub for distributing signals upwards in the hierarchy, is generally unclear; however LGN shows a feature that is often taken as a preliminary indication of the dual nature of the visual system (discussed in section 1.3.2): it is formed by two main classes of layers (Magnocellular and Parvocellular) whose cells are characterized by different body sizes, different behaviors, and different areas of projection.
Input to the LGN from the retina keeps a retinotopic mapping, meaning that nearby regions on the retina end up near one another in the layer; it is usual to think of each layer as representing some form of feature map.

The Visual Cortex

Most visual signals leaving LGN arrive at the primary visual cortex, usually called area **V1**, where they are processed in a fairly well known way, and then routed to parietal and temporal areas of the cortex. The visual cortex is far the most studied, and most extended, portion of the cortex; and this understanding has led to a widely accepted computational model of early vision, described in section 1.1.1. In this section we briefly review the three main facts about early vision: there exist separate areas, each one processing a specific feature of the visual input; processing of a given feature takes place concurrently at several spatial locations and scales; there is a distinction between *simple* cells with smaller receptive fields, and *complex* cells that receive projections from pools of simple cells.

Similarly to lower areas, the cortex is retinotopically mapped, and cells are arranged so that their receptive fields move smoothly from the center to the periphery of the visual field. The early stages of cortical processing are usually classified according to their response properties, and this part of the visual brain is interpreted as an *atlas* composed by many separate retinotopic maps [138]: each map represents, i.e. consists in, the responses to a specific feature of the visual input, such as intensity, orientation and color (**V1**,**V2**), motion (**V5**), directed motion of oriented edges (**V3**,**V5**), oriented color (**V4**), etc. Furthermore, psychophysical experiments on adaptation to specific

spatial frequencies (e.g. [13]) have shown that contrast thresholds after a short time of adaptation, are elevated only for a limited range of spatial frequencies close to the adapting frequency. This suggests that the visual cortex is sensitive to several spatial frequency channels; the contrast sensitivity function can be seen as a superposition of several contrast sensitivity functions, one for each channel. Generalization of these results led to the idea that each specific feature is processed at several spatial scales.
Finally, along the hierarchy, cells are classified also as being *simple* or *complex*: simple cells have smaller receptive fields, and respond locally to a specific feature, while complex cells have larger receptive fields, receive input from a pool of simple cells and are thus selective for the given feature (or combination of features) at a more global level. The most basic example is provided by cells selective for a given orientation: simple cells of this type respond most strongly to *edges* with the given orientation, while complex cells are tuned to *bars* of that same orientation.

A Model of Early Vision

The current best model of human early vision, based on the evidence described in sections 1.1.1 and 1.1.1, is that the visual signal is split into several spatial frequency bands, and each band is then subjected to a set of linear filters. After this point, different routes can be followed: if the task is to obtain recognition abilities, then the responses of these filters are subjected to a non–linearity followed by noise; this is schematized in Fig. 1.2 and described in the rest of this section. Otherwise, when the task is modeling bottom–up attentional processes, linear multiscale responses are combined across scales and across features to obtain a so called *saliency map*, upon which the attentional process is deployed (see section 1.4.2 for details). The model depicted in Fig. 1.2 is the multiresolution model of early vision, reduced to the case where the only basic feature processed is orientation, and the ability to be reproduced is pattern sensitivity.
The stimulus is first smoothed with a Gaussian Derivative and subsampled to obtain the Pyramid, i.e. the collection of copies of the input image at different scales. Then the image at each scale is convolved with linear filters at a variety of orientations. Linear filters are chosen because spatial simple cells act as edge detectors; this response

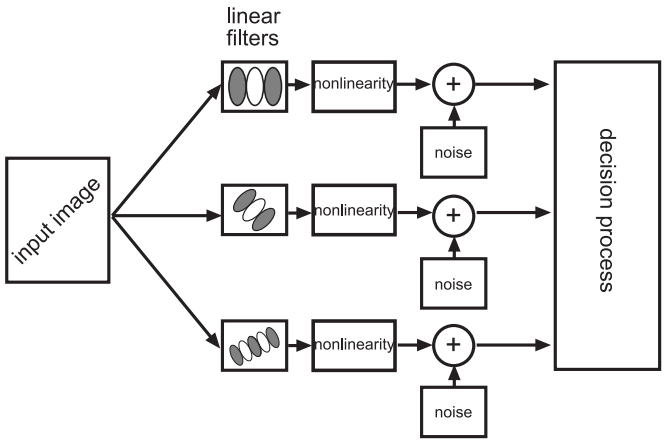


Figure 1.2: An overview of a multiresolution model of human pattern sensitivity. The stimulus is convolved with linear filters at a variety of scales and orientations — as an illustration we show three scales, one orientation per scale — and then subjected to a nonlinearity. The results have noise added, and are passed to a decision process.

is well modeled, and generalized, by Gabor filter pairs (see Fig.1.3):

$$G_{sym} = cos(k_x x + k_y y)\, exp - \left(\frac{x^2 + y^2}{2\sigma^2}\right) \quad (1.1)$$

$$G_{asym} = sin(k_x x + k_y y)\, exp - \left(\frac{x^2 + y^2}{2\sigma^2}\right) \quad (1.2)$$

The characteristic behavior of filters is that they are *similar* to the pattern they can detect: as an example, the first filter in Fig.1.3 looks like a vertical light blob next to a vertical dark blob, which is similar to what happens in proximity of a vertical edge. In general Gabor filters resemble groups of oriented bars.
After convolution with several Gabor filters, at several orientations, frequencies and scales, the resulting retinotopic maps are subjected to some form of non–linearity, to get the output of the specific complex cells. The results have noise added, and are passed to a decision process.

1.1.2 Limitations

The passive approach to vision has proven successful in many respects, and has led to efficient algorithms to solve a number of visual tasks. Nonetheless, many fundamentally problematic issues can be found in this approach.

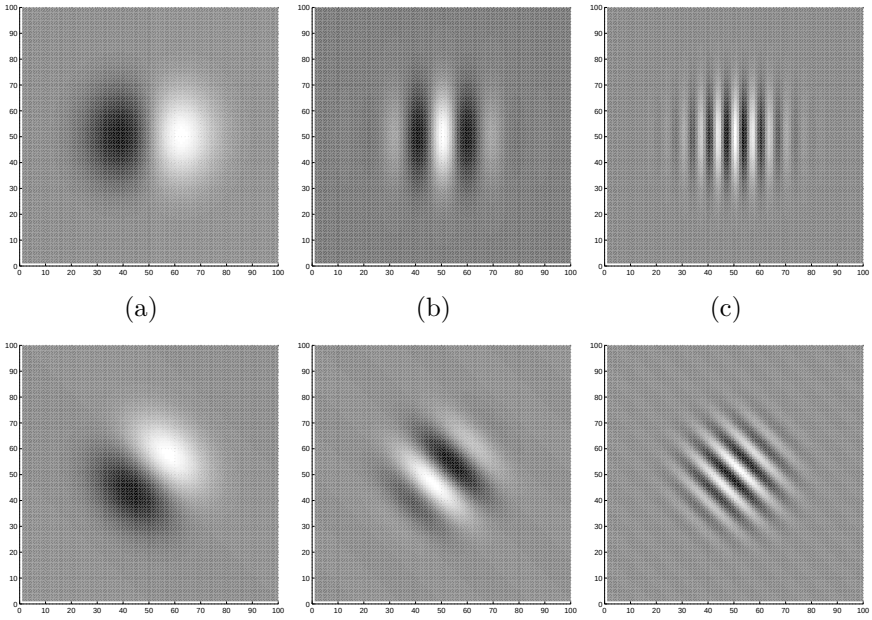


Figure 1.3: Gabor filter kernels are the product of a symmetric Gaussian with an oriented sinusoid; the form of the kernels is given in the text. The images show Gabor filter kernels as images, with mid-grey values representing zero, darker values representing negative numbers and lighter values representing positive numbers. The top row shows the vertical antisymmetric component, and the bottom row shows the diagonal antisymmetric. Spatial frequency increases from left to right.

In the first place, nowadays it is largely accepted that the function of the visual system cannot be reduced to that of building an internal representation of the outside world, intended as a processed version of the retinal image, and there is much debate about even the existence — leave alone the usefulness — of such representation.
Problems arising from this approach include the issue of *transaccadic integration*, namely how the supposed internal representation produced by passive vision might be maintained when the eyes are moved; and the well known *binding problem* [126], questioning the mechanisms by means of which different features processed in separate modules are then integrated in a veridical way. This issue is usually solved introducing the concept of visual attention: however, different perspectives on visual attention can be taken, and there is much evidence that the *mental spotlight* metaphor advocated by passive vision is misleading (see section 1.2.1).
Finally, the passive approach underemphasizes the inhomogeneity of

the retina and visual projections (see section 1.2.2); it has however been pointed out recently that this is one of the most fundamental features of the visual system, e.g. because it is the best way to make the computational resources available in the brain sufficient to process the incoming information.

1.2 Active Vision And Visual Attention

The ability to move the eyes is a feature common to humans, most other vertebrates and some invertebrates. This section analyzes the strands of research on human and computational vision and visual attention, that have led to recognizing the pervasive importance of eye movements in perceptual and action–related functions. The main critiques to traditional, passive approaches, and some milestones in research on eye movements will be presented in section 1.2.1; this discussion provides an introduction to the reasons and main themes of the so called *Active Vision* approach, and its tight relation with studies on visual attention. It follows, in section 1.2.2, a brief overview of the anatomical background to active vision.

1.2.1 The Active Vision approach and its relations to Visual Attention

The expression *Active Vision* has been coined in relation to studies in the computer vision community, that tried to overcome the computational difficulties raised in the passive vision framework [1, 5, 34]. These works tried to reduce the computational demand of creating a complete, detailed representation of the visual scene, by adopting animal–like visual sensors with a central high–resolution region that could be redirected to different locations of the visual field. Although this approach underemphasizes the perceptual role of vision, not only such systems were shown to be effective in the visual guidance of agents' behaviors, but also to allow the deployment of pointing, or *deictic*, properties that can be used as an interface with the cognitive activity of a situated agent [7, 93].
In a well–known paper by O'Regan and Noe [85] dealing with the issue of visual consciousness, the basic idea of the Active Vision paradigm was effectively synthesized as follows:

> Instead of assuming that vision consists in the creation of an internal representation of the outside world whose activation somehow generates visual experience, we propose to treat vision as an exploratory activity. [...] The central idea of our new approach is that vision is a mode of exploration of the world ...

The proposal that eye movements provide an essential means of exploring the world, is in line with a critical analysis of classical ex-

periments in psychology. Such experiments, following the idea of a static, fully detailed representation in the visual brain, used tachistoscopic stimulus presentation techniques, with stimuli displayed for times shorter than the saccadic latency period of about 150 *msec* required for an eye movement to occur. Although it can be shown that subjects under this condition are able to see and solve tasks such as recognition of familiar images, it probably cannot be said that observers are seeing the pictures in the normal sense of the word [81]. Several experiments have shown in fact that when the viewers are not allowed to move their eyes [105], or when the image is modified as soon as a saccade is executed [80], then their ability is highly reduced in solving tasks such as learning and recognition of abstract patterns, or item counting [3].
The active vision approach has moved the emphasis towards the sequential nature of information intake in vision. As a consequence of the shift from the passive to the active vision paradigm, starting in the 1980's a renewed interest in eye tracking studies has flourished, and a number of experimental tests have been designed in order to specifically analyze the role of eye movements in both purely visual (see section 1.4.1) and visuomotor (section 1.5.1) tasks. Similarly, a number of mathematical and computational models of eye movements have been developed, some of which will be discussed in sections 1.4.3 and 1.5.2.

A similar course can be recognized as well in the literature on *Visual Attention* (see [33] for a review). This phenomenon, or rather collection of phenomena, has been investigated diffusely in the cognitive sciences; as a reflex of the predominant passive vision approach, one of the most frequently emphasized facts concerning visual attention, is the ability to attend *covertly* to a location in the visual field without directing the eyes to that location.
This result, already known to Helmholtz, has been commonly interpreted in terms of a *mental spotlight*. The spotlight metaphor envisaged covert attention as the process of moving the spotlight towards a specific region of an internal representation, thus assigning enhanced processing to that region at the expense of others. Once again, this interpretation relied on the idea of a fully detailed mental representation of the visual scene. When considering the relation to overt attention, this approach implies that the spotlight is moved *before* each saccade in order to select the locus of the next fixation; this requires that covert attention be reallocated much more rapidly than

saccade latency, which is highly implausible. Furthermore, the conceptual weakness of the traditional approach is that it postulates the existence of a central, supramodal mechanism of attention subserved by anatomical circuits separated from those involved in sensorimotor processing.
However it can been noticed from more recent literature on attention that a distinction has been traced between *selective* visual attention for object recognition [24] and *spatial* visual attention for action control [99] (a similar classification of the functional areas of the visual system is discussed in section 1.3.2). While the mechanism proposed for object attention seems to be related to those for object analysis [75], the mechanism for spatial attention appears to be related to processes responsible for the organization of movements in space [100]. Thus according to this view the programming of eye movements — and generally speaking of motor responses — is the primary spatial attentional process, among whose consequences are the effects identified as covert attention. This idea builds on the long recognized link between covert attention and eye movement programming. Several experiments have demonstrated that during the preparatory period of a voluntary eye movement, responses to an attentional probe are faster at the destination location of the eye movement [111]. In particular, the *premotor theory of attention* [110] proposes that covert attention to a visual location results from the activation of the same neural circuits employed for the programming of saccades and actions in space, while at the same time inhibiting the actual motor response.

The work presented in this thesis follows the active vision approach, and tries to extend it further by generalizing the fundamental ideas of the premotor theory of attention; as detailed in the following chapters, the focus of the thesis is in fact the extension of current models of active vision to account not only for eye movements in visual tasks, but also for the effects of the tight interaction of motor and oculomotor programs in tasks that require some bodily action.

1.2.2 Anatomical need for eye movements

The first stage of visual processing happens in the so called *cones*, photoreceptive cells on the retina that transduce the impinging light in electrochemical signals. Apart from cones, other cells are distributed on the flat retinal surface. However a depression with a diameter about 1500 μm can be observed in the central region; it is in this re-

gion that cones are most densely present while other cells are absent. This region corresponds to about 5 degrees of the visual field, where the highest optical quality (resolution) of the image is achieved; correspondingly, psychophysics shows that many visual functions exhibit gradually decreasing ability as the stimuli are placed at increasing distance from the visual axis (an important exception being the detection of motion or temporal changes): for descriptive convenience, the region within 1 degree of the visual field is called the *fovea* and it is indicated as the region of highest visual acuity, the region extending from 1 to 5 degrees is called *parafoveal*, and the *peripheral* region encompasses the remaining visual field.
Signals coming out of the cones are then transmitted through the optical nerve and ganglion cells, and distributed along different pathways: the main pathway is the one comprising the Lateral Geniculate Nuclei (LGN) of the Thalamus, up to the Visual Cortex (see section 1.1.1). Many secondary pathways exist, the most relevant to active vision being the one that terminates in the Superior Colliculus (SC, see section 1.4.3).
A general characteristic of the signals projected along early stages of vision, namely from the retina up to the visual cortex, is that a topographic mapping is used, whereby spatial relations in the activation map of retinal cells are maintained in all subsequent layers; however, as the signal proceeds, an increasing proportion of representation is assigned to central regions, following an empirical transformation law proposed in [107]:

$$\begin{aligned} u(r,\phi) &= log(r) \\ v(r,\phi) &= \phi \quad . \end{aligned}$$

Here r and ϕ define a point in peripheral vision using radial coordinates, while u, v are the cartesian coordinates of the corresponding point in the given cortical map.

As a consequence of the retinal anatomy and the central magnification effect of visual projections, visual acuity declines from the center to the periphery, although with quantitative effects that depend on the given task and context (see [34] for a detailed discussion).
These facts are the basis for the active vision approach, for two main reasons: first, for any given task, eye movements should be considered as necessary to acquire the detailed information required to solve the task, since the information acquired from the visual periphery is poor

and degrades during subsequent elaborations; and second, peripheral vision should not be considered simply as a low resolution version of foveal vision, serving the same purposes, but rather as functional to providing *preview* cues useful to reorient the gaze direction.

1.3 Dichotomies In The Visual Brain

1.3.1 Bottom–Up *vs* Top–Down processes in visual attention

A seminal paper by Pylyshyn [91] points out that although the study of visual perception has made more progress in the past 40 years than any other area of cognitive science, there remain major disagreements as to how closely vision is tied to cognition:

> [...] the question of why we see things the way we do in large measure still eludes us: Is it only because of the particular stimulation we receive at our eyes, together with our hard-wired visual system? or is it also because those are the things we expect to see or are prepared to assimilate in our mind? There have been, and continue to be, major disagreements as to how closely perception is linked to cognition - disagreements that go back to the 19th century. At one extreme some see perception essentially as building larger and larger structures from elementary retinal or sensory features. Others accept this hierarchical picture but allow centripetal or top–down influences within a circumscribed part of vision. Then there is "unconscious inference" first proposed by von Helmholtz and rehabilitated in modern times in Bruner's New Look movement in American psychology [15]. According to this view, the perceptual process is like science itself; it consists in finding partial clues (either from the world or from one's knowledge and expectations), formulating a hypothesis about what the stimulus is, checking the data for verification, and then either accepting the hypothesis or reformulating it and trying again in a continual cycle of hypothesize-and-test.

The main claim of Pylyshyn's paper is that there are undoubtedly top–down cognitive influences on visual perception, but also that the visual system itself is composed of different functional areas, and some of these areas are not accessed by cognitive factors. In the words of the author, "vision as a whole is cognitively penetrable" but the "early vision system is encapsulated from cognition, or to use the terms we prefer, it is cognitively impenetrable" [91].
A further distinction introduced in the above mentioned paper is a

classification of different types of top–down influences: first, there are well known top–down signals running through descending projections in the hierarchy of cortical areas assimilated to early vision; however, these are not related to cognitive factors, such as beliefs or expectations, rather this is a fine grain description of how the early visual analysis is carried out to eventually provide higher functional areas with a bottom–up contribution. Secondly, there exist genuinely cognitive top–down factors that influence vision, but these contribute in determining the nature of perception at only two loci. In other words, the influence of cognition upon vision is constrained in how and where it can operate. These two loci are: 1) In the decisions involved in recognizing and identifying patterns after the operation of early vision. Such a stage may (or in some cases must) access background knowledge as it pertains to the interpretation of a particular stimulus; 2) In the allocation of attention to certain locations or certain properties prior to the operation of early vision.
Thus, for what concerns our discussion on visual attention, Pylyshyn's work helps us define the kind of *top–down* processes which contribute to overt attention allocation.

On the other hand, it is known from everyday experience that eye movements can as well be entirely automatic and independent on cognitive factors, e.g. saccades in response to abrupt changes in the visual periphery. As a matter of fact, many models (some of which are discussed in section 1.4.2) have tried to account for the statistical properties of eye movements (mainly the distribution of fixation locations) on the basis of image properties alone. Most part of such models, that in the perspective of the present discussion account for *bottom–up* contributions to visual attention, rely on the concept of a *saliency map*: not only does the hierarchy of primary visual areas (as described in section 1.1.1) produce a detailed representation of the external world, but it also produces the so called *saliency map*, namely a topological representation of the saliency, or conspicuity, of image locations; according to these accounts, foveal attention would be driven towards the most salient locations, thus relying entirely on the properties of the visual stimulus.

1.3.2 Action *vs* Perception, and the debate on consciousness

As anticipated in section 1.1.1, the physiology of early visual stages reveals the existence of two classes of layers, namely Magnocellular and Parvocellular, whose cells are different in body size, response and outgoing projections.
This fact is often interpreted, although with some controversy, as a signal of a much more general property of the visual system: the existence of two different streams of processing, whose physiological and functional characteristics are significantly different. Although the interconnections among cortical areas are multiple, two main routes were first clearly identified in primate's brain [130]: a *ventral stream* projecting onto temporal areas and a *dorsal stream* running from the primary visual cortex to parietal areas. The authors suggested that the kind of visual processes carried out by the two streams were related respectively to visual recognition, and visuospatial awareness.
Further investigation of this duality has led to the classical view of the streams as *what* (recognition of object) and *where* (spatial location of objects). More recently however, Milner and Goodale [74] proposed a different interpretation: rather than emphasizing differences in the visual information handled by the two streams, their account has instead focused on the difference in the requirements of the output systems that each stream of processing serves. The view is now largely accepted that the functions subserved by the two streams are respectively *recognition* and the *control of action*, thus leading to the following dichotomy: Vision for Perception vs Vision for Action (see Fig. 1.4). The fundamental idea of this account is that the visual system for perception and the visual system for action carry out very different transformations on the same visual input, because the requirements of the output systems they subserve are different. Information about attributes of objects — such as size, shape, orientation, location — are processed by both streams, but the nature of that processing is different; as an example, only the relative value of the above mentioned attributes is of concern to perception, and perceptual representations deliver inaccurate, rough metric information. By contrast, actions must be fine tuned to the real metrics of the world. The following table summarizes some of the main differences between the two streams, as indicated by the work of Milner and colleagues.

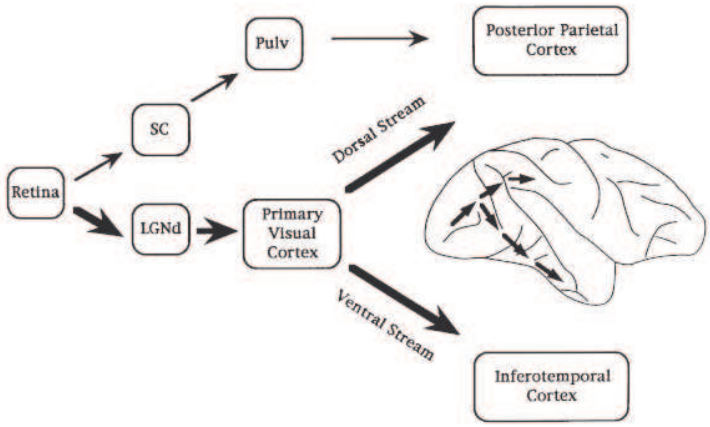


Figure 1.4: Simplified scheme of the main visual pathways, with emphasis on the *Dorsal* and *Ventral* systems, after [74].

Ventral	Dorsal
Scene–based reference frame	Effector–based reference frame
–	–
Relative metrics	Absolute metrics
–	–
Representation can enter memory	Moment–to–moment computations
–	–
Visually conscious	Unconscious

The main experimental evidence for this interpretation of the dual vision system, come from observation of patients with brain damages. In particular patients affected by Optic Ataxia in consequence of lesions to the PPC are able to recognize familiar objects, but show a difficulty or complete inability to reach and grasp; clearly, depending on the precise region of the lesions, several specific deficits appear, as different subregions of PPC support transformations related to different motor output. An example of particular interest to the present thesis is that of ataxic patient D.F., who can recognize objects but when asked to recall them from memory and represent them by drawing, can only produce unrecognizable traces[1]. At the opposite, patients

[1]As a side remark, it should be said however that drawing from memory (and drawing in general) should not be considered as a purely action–related task, since recalling the visual appearance of objects from memory could as well involve perceptual operations: drawing is the kind of task where perception and action are hardly dissociated.

affected by Visual Agnosia due to lesions to the ventrolateral region of the Occipital Cortex are unable to recognize familiar objects and, in the most severe cases, even faces of relatives and friends; yet, they are able to execute consistent motor actions directed towards those same objects, and in same cases can recognize the objects in another, non–visual sensory modality.

Thus, all this stream of evidence not only confirms that the two main pathways of the visual system support different functions, and that such functions are related to perception and action, respectively. It also, at least in the perspective of Goodale and Milner, supports the idea that perceptual experience in the visual modality is almost entirely enabled by ventral modules, while the dorsal stream subserves action–control functions that can run semi-autonomously, or that at least carry on computations whose results are not accessed directly by awareness.
Although out of the scope of this thesis, it is interesting to remark that clearly this view is not undebated, and an influential account of perceptual experience and consciousness has been proposed by O'Regan and Noë [85] to oppose the dual vision theory. The sensorimotor, or *enactive*, approach advocated by those authors regards action as constitutive to perceptual experience. In particular reference to vision, recalling the fundamental role of eye movements, the enactive approach states that the human brain is able to learn so called *sensorimotor contingencies*, namely sets of rules describing the transformation of the sensory input upon movement of the sensory apparatus: it is exactly the mastery and usage of these contingencies that enables perceptual experience, thus once again highlighting the fundamental role of the *exploratory* movements of active sensors.

1.4 Bridging The Gaps: Eye Movements

Notwithstanding the lively debate regarding the dichotomies mentioned in section 1.3, there is widespread agreement about the picture of the visual and visuomotor brain as an integrated system that delivers to us a coherent perceptual experience of our relations to the world. In other words, if any dichotomies exist, it seems to be out of discussion that some mechanisms must exist that facilitate the integration and coordination of the processes corresponding to the opposite poles of each dichotomy.
Intuitively, the ability to move the eyes and redirect the gaze is one of such mechanisms, probably the most apparent, and eye movement recordings have been for long time a valuable source of insight on the kind of strategies that are adopted in solving visual and visuomotor tasks. This approach, namely that of inferring the relevant strategic and cognitive factors from eye movement data, will be discussed in detail in section 1.4.1 where we consider the most common experimental paradigms in active vision: text reading, scene perception and visual search (more recent experimental trends, aimed at analyzing eye movements in complex visuomotor tasks, will be discussed separately in section 1.5).
In support of the above mentioned intuition, neurophysiological evidence shows that eye movements are the result of the cooperation/competition of multiple brain areas, each related to a different function. In section 1.4.2 we discuss the contribution of Early visual areas, as captured by computational models that have been successful in accounting mainly for the bottom–up processes that guide visual attention. Subsequently, in section 1.4.3, we analyze a different stream of models, based on probabilistic techniques, that have recently proven successful in modeling the main brain centers for gaze control thus accounting also for top–down effects.
Eventually, in section 1.4.4, we briefly review the neural paths of eye movement generation, and submit that those paths involve processes that span all of the poles of the dichotomies recalled in section 1.3.

1.4.1 Classical paradigms in eye movement research

Classical experimental paradigms for studying eye movements and their relation to attention and cognition are text reading, scene perception and visual search.

Visual search, namely the task to locate a *target* among a number of *distractor* items on the basis of some visual properties, is an activity where cognitive influences can be minimized, and it is thus the most suitable to focus on the basic properties of saccadic movements, namely the *latency* and the *metric properties* (the partial independence of the processes regulating such properties is also at the basis of one of the first models of complex 2d eye movements, described in section 1.4.2). Furthermore, it turns out there to be a close correspondence between the findings in such behavioral studies and those shown in experiments of primate brain physiology (see [132] for a review).
The main question related to visual search is the following: are the items processed within each fixation dealt with in parallel, or is there a process of covert attentional scanning that scans each item serially? Typically the *Reaction Time* (RT) to indicate the presence or absence of the target is measured, and its plot against the number of display elements is termed *search function*; a flat search function is thought to reflect *parallel search*, whereas a function that denotes an increase in RT with the number of elements involves *serial search*.
A dominant tradition in visual search was initiated with a seminal paper by Treisman and Gelade [127]. They argued that some primary visual properties allow a search in parallel across large displays. In such cases the target appears to 'pop out' of the display. More recent and detailed studies of eye movements in search reinforce this view and exclude a model in which covert attention scans around the display prior to any overt eye movement, thus favoring the parallel–processing model according to which visual search involves serial overt scanning with eye movements and parallel processing of few items during a given fixation.
Work in the area of visual search thus produces convergent findings with those discussed in section 1.2.1, and closely related to the ideas concerning the *premotor theory* of attention: during a fixation, visual information is processed in parallel with enhanced processing at the area of the following saccade location. There is no support for sequential intra–fixational scanning by covert attention.

Another historical area of behavioral experimentation in eye movements is *text reading*. This task involves a tight interaction between perceptual and cognitive functions, while at the same time posing strict constraints on eye movements of the reader that make it easier

to classify saccades in few classes: a reader typically advances along a line of text making a sequence of fixations and *forward* saccadic movements. Occasionally within–word *refixations* or inter–word *regressive* movements in the reverse direction breaks the forward sequence of the scanning. Again, measured quantities are usually saccade length and fixation duration.
Further experiments, based on *gaze–contingent methodologies* [71], allowed to study the *perceptual span* and the role of parafoveal preview. Various different manipulations of this technique have been developed, among which the *moving window* manipulation is perhaps the most widely used. In this technique, as applied to reading, the subject sees normal text within a window of predefined size but outside this window the text is masked in some way, for example each text character may be replaced by an "x". Each time the eye moves, the display is changed so that the unmodified material is always presented precisely where the gaze is directed. If the manipulation does not affect the speed of reading, it is reasonable to assume that the material outside the window is playing no part in the normal process. Conversely, when the window size is reduced so that reading speed is affected adversely, then it can be deduced that material from regions outside the boundary must normally be processed. This has allowed measurements of the perceptual span, the region from which visual information is extracted. Summarising many results, the perceptual span is found to extend no more than 15 characters to the right of fixation and only 3-4 characters to the left, and the asymmetry of the span is an immediate indication that premotor attentional effects are involved.
Thanks to the restricted set of movements allowed in reading, this area was the one where the first mathematical and computational tractable models of eye movements emerged, to reproduce the statistics of saccade length and fixation timing (see e.g. [70, 83, 96], and section 1.4.3).

Although much less intensively investigated, some of the principles emerging from studies of active vision in reading are also at work during the more general situation of viewing visual material such as *natural objects and scenes*. Early studies in picture viewing, such as the famous works by Buswell [16] and Yarbus [136], already pointed out that wide individual differences were present, and most important, that even more substantial differences could be found in the same subject, looking at the same image, when asked to solve different visual tasks.

An important breakthrough in gaze analysis was the concept of *scanpath*, introduced by Noton and Stark [82] who claimed that when a particular *visual pattern* is viewed, a particular *sequence of eye movements* (the scanpath) is executed. Although this direct relation of the scanpath to visual content has been rejected soon, the general position that eye movement patterns can reveal much about visual perception has been retained in subsequent research, and a number of workers have developed techniques to capture statistical regularities in the pattern of eye scanning. The simplest form of sequential dependence in basic gaze parameters is the Markov process, in which the properties of the immediately preceding saccade constrain the probabilities of the one currently programmed; these kind of models were soon extended to account also for properties of the image that can make a region more *salient* than others, and additional cognitive factors that make a region more *informative* than others (see section 1.4.3 for extensive discussion).

Other attempts to carry out quantitative measures in this framework, address the *useful field of view* for picture perception using a gaze-contingent window technique in a manner similar to that used in studies of reading; but in the case of pictures [101, 112] viewing time and recognition scores are impaired unless the window is large enough for about half the display to be visible, and even very low–resolution detail in the periphery (very well below the acuity limit at the peripheral location) aids performance considerably.

However, although fundamental, the role of peripheral vision in picture viewing, according to most influential recent studies, remains quite different from what envisaged in the framework of passive vision. A particularly striking example is the phenomenon called *Change Blindness*: under many experimental conditions, viewers tend to not notice even significant changes in an image, for example if a white frame (a 'flicker') is shown for about 80 *msec* between the two images depicted in Fig. 1.5(a) and 1.5(b). In [84] Change Blindness is interpreted as a proof of the fact that our internal representations of the outside world, instead of being very detailed and rich, are actually rather sparse. In this view, to get the impression of richness, there's actually no need for the richness to be in the head; rather, what has to be in the head is merely effective procedures for *getting at* the information in the world. Such algorithms we have, in the form of movements of the eyes or shifts of attention. If we're interested in some detail of the visual scene, we simply need to move our eyes or our attention to that detail, and it is immediately available.

The above mentioned account of Change Blindness can be summarized by saying that in the active vision approach the world is treated as an *external memory*.

(a)

(b)

Figure 1.5: An image and its modified version, used in a Change Blindness demonstration. The change in the image is hardly detected by a human observer if a white frame is interposed between the two images for about 80 *msec*.

1.4.2 Saliency–based models

Several models for stimulus–driven (or bottom–up) overt eye movements have been designed; here we limit the presentation to three recent, influential models, that are based on the pervasive concept of a *saliency map*, and also give some hints on how to include top–down factors. It is worth recalling that biological plausibility of the saliency map is still largely debated, and a review of candidate brain areas that could contribute to such map, is presented in [113].

Saccadic movements can be studied according to two basic properties, namely the *latency* and the *metric properties*. The partial separation between such two properties was the principal feature of the functional model proposed by Findlay and Walker [35]: based on the emerging physiological knowledge of the brain pathways involved in visual orienting, this model separates two pathways controlling *Where* and *When* information. In Fig. 1.6 this separation corresponds to the two vertical streams. The *When* stream is envisaged as a single individual signal whose activity level varies. The *Where* stream is a set of interconnected activity maps, resulting in a *salience map* from which the saccadic target location is selected. The horizontal bands represent processing levels that become progressively less automatic moving in the hierarchy from bottom to top. Interaction between the two streams occurs at the lowest levels in terms of reciprocal competitive inhibition.
It should be noticed that such model presents a precise, physiologically motivated, account only of its lowest levels, those corresponding to automatic, or preattentive, orienting processes. Top–down influences are just loosely designated.

Furthermore, the concept of *saliency map* was first proposed by Koch and Ullman [56], as the key part of a model for implementing parallel search in biological systems. A saliency map is a topographical map encoding the 'saliency' of each point in the visual input; its essential feature is that it pools the outputs of different feature maps. In this way, the final level of salience at any point is indiscriminate with regard to its origins in color/brightness, form, motion properties, and should be additive across features.
Up to date, the most successful computational implementation of a preattentive selection mechanism based on the architecture of the primate early visual system, was presented recently by Itti and Koch [51],

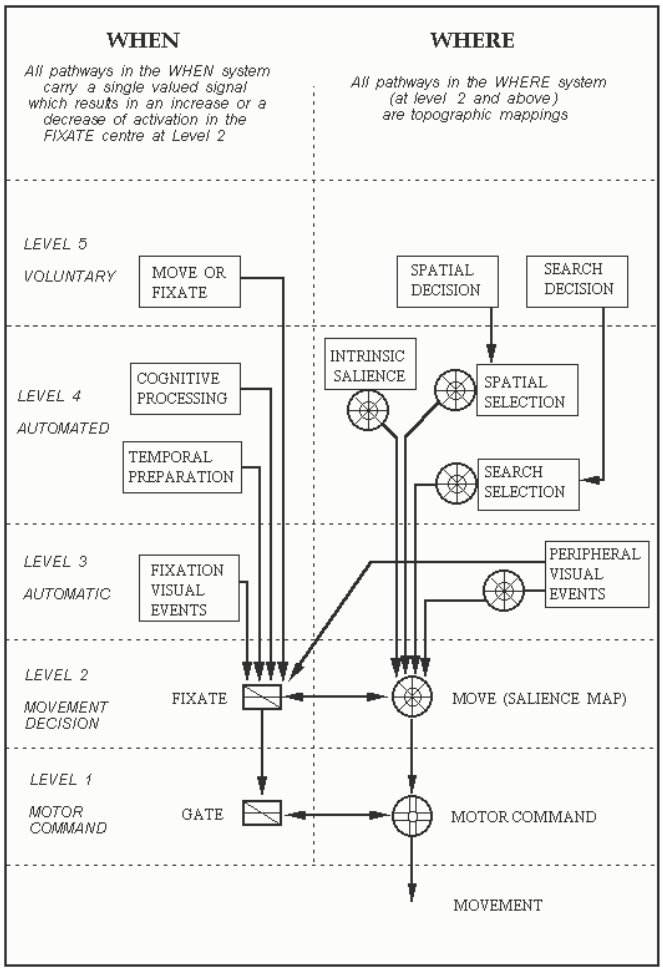


Figure 1.6: The framework for saccadic eye movement generation presented by Findlay and Walker [35].

and it relies essentially on the saliency map. In this architecture (see Fig. 1.7), low–level visual features (color, intensity and orientation) are extracted in parallel from several spatial scales, using a biological center-surround architecture (see section 1.1.1). The resulting feature maps are combined to yield three *conspicuity* maps for color, intensity and orientation. These, in turn, feed into a single saliency map, consisting of a 2D layer of *integrate–and–fire* neurons. Finally a *winner–take–all* network shifts the focus of attention to the currently most salient image location. Feedback inhibition (also called inhibition-of-return, *IOR*) then transiently suppresses the currently attended location, causing the focus of attention to shift to the next most salient image location.

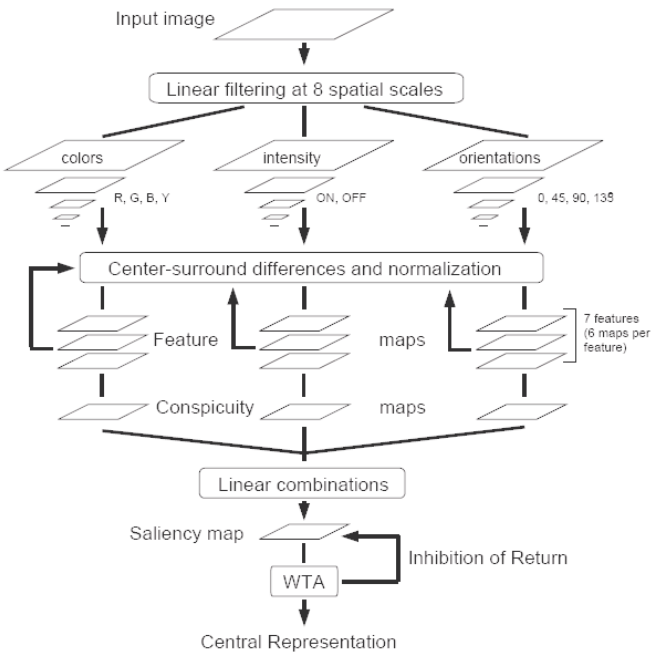


Figure 1.7: General architecture of the saliency based model by Itti and Koch [51].

A third model that is worth mentioning here, is the 'DPZ' model [22], implementing the proposal appeared in [26] later refined by neurophysiological evidence[24]. This so called 'biased competition' theory of visual attention includes two forms of Top–Down influence: the first is related to signals coming from higher ventral area **IT**, and accounts for a bias in the computation of feature maps. A sort of saliency map, located in this model in Posterior Parietal (**PP**) areas, is the source for the second type of top–down signals; the feature maps and **PP** are reciprocally linked, and it is through their iterative interaction that the model gradually converges to a single winning–item location. Attention is thus a dynamic emergent property of the modeled system, and in the converged state the selected location is active in all feature maps – even if, for a given map, the target item is of low salience.

1.4.3 Probabilistic models that include top–down factors

In this section we present a survey of the considerable body of literature from experimental psychology and artificial intelligence that led to the adoption of probabilistic tools in modeling eye movements

in *purely visual* tasks; these class of models gave rise to most recent computational accounts for top–down influences in gaze shifting. The models presented here will be recalled then in chapter 3, where a novel functional model for gaze control in *visuomotor* tasks is presented.

Hidden Markov Models for gaze control

Eye movements are usually analyzed on the basis of the distinction between *saccades*, i.e. ballistic movements, and *fixations*, corresponding to the periods during which the displacement of the gaze point remains below a given threshold. Under this conceptualization, the oculomotor behavior is described primarily by the *scanpath*, i.e. the temporal sequence of fixations executed by the subject.
The sequential nature of gaze allocation, and the observation that saccades are driven by neural signals that are inherently noisy, suggest that scanpaths are best described by stochastic processes. Early attempts to model eye movements this way adopted zero–order and first–order Markov chains [27, 119, 47], while Hidden Markov Models (HMM) where then introduced [98, 69, 28] to account for the two stages of saccade generation, namely planning and execution; more recent technical advances in this direction were described in [67], with the adoption of a Bayesian framework for statistical inference on HMM's that allows to infer the hidden *cognitive* state from the observed eye movements (inverse inference).
A problem with the models above is that standard HMM's for visual attention conceptualize attention as an *autonomous* random process that is not affected by the visual information perceived during fixations. Opposite to this schematization, it stands the obvious fact that the variability in observable quantities, e.g. fixation duration and saccade length, reflects not only random fluctuations in the system but also factors such as moment–to–moment changes in the visual input, cognitive influences, and the state of the oculomotor system. In order to account for those additional factors, previous models have been recently extended with the adoption of variables that explicitly model top–down contributions to saccade planning.
Feng, in [31, 32], has extensively described a suitable computational strategy to capture the relationship between cognitive (hidden) processes and (output) eye movements, accounting at the same time also for the contribution of the visual input at each fixation. This model of eye movements in reading adopts the Input–Output HMM (IOHMM, see [11]) which is a special case of a class of probabilistic graphical

models called Dynamic Bayesian Networks [77]. In the IOHMM the temporal evolution of the cognitive/attentional state (or hidden) variable is described as a Markov process and inferred from the observed output variable, but is also conditioned on an additional observed (input) variable (see Fig. 1.8). More formally, at each time step t, define

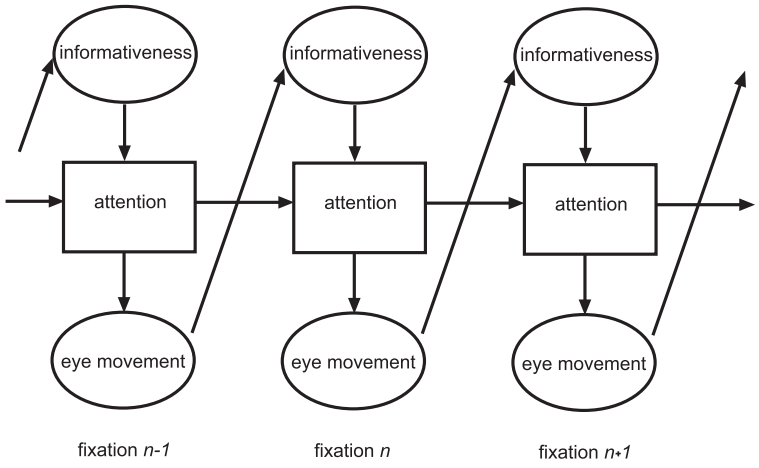


Figure 1.8: Diagram of the Input-Output Hidden Markov Model (IOHMM) for eye movements in a generic visual task, proposed in [31].

an observable variable i_t that accounts for the *informativeness* of the fixated region of the image, a hidden variable d_t for the cognitive (attentional) state of the system, and an observable variable x_t denoting the location of the fixation. In order to design a suitable graphical model, some assumptions must be made on the dependencies of different variables; a particular example would be that the cognitive state depends only the observed input and previous cognitive state according to:

$$p(d_t|d_{t-1}, i_t) \quad ; \tag{1.3}$$

and eye position depends on the cognitive state:

$$p(x_t|d_t) \quad . \tag{1.4}$$

Under these assumptions, the graphical model becomes the one depicted in Fig. 1.8, upon which inverse inference of the cognitive state can be realized:

$$p(d_t|x_{1:t}, i_{1:t}) = \frac{p(x_{1:t}|d_t, i_{1:t})p(d_t|i_{1:t})}{p(x_{1:t}|i_{1:t})} = \frac{p(x_{1:t}|d_t)p(d_t|i_t)}{p(x_{1:t})} = \frac{p(d_t|x_{1:t})p(d_t|i_t)}{p(d_t)} \tag{1.5}$$

where the first and last passages are obtained after application of the Bayes rule, and simplifications in the third term correspond to missing connections in the independence graph of Fig. 1.8. Notice that in equation (1.5) the posterior probability of a cognitive state depends separately on the observations (eye movements) and the visual input. This model has been implemented in a reading task, and after learning model parameters from experimental data, it has proven effective in fitting the statistical properties of reading eye movements [32].

Bayesian models of cortical pathways for gaze control

We move now to another influential model of eye movements based on Bayesian techniques, that has been proposed recently, as an extension of Bayesian models of visual cortical connections.
As explained in section 1.1.1, processing of visual information in the visual cortex is usually schematized as the operation of a sequence of modules, each one taking as input the output of the previous one; this can be called a *feedforward* model, as it accounts for the feedforward connections among cortical areas; the role of forward ('driving') connections is thought as favoring certain input patterns over others, leading to a progressive evolution of response selectivity in the ascending direction. It is well known however that *feedback* connections also exist and, although their role is more subtle, they constitute a significant part of visual processing in primates, especially because cortical feedback enables top–down factors to influence visual attention.
The key idea in modeling feedback connections is that feedback acts to select certain ascending signals in preference to others, culminating in a steady state resonance, in which the feedback and feedforward activity is mutually reinforcing over several hierarchical levels (see [76], and [114] for a review).
The model proposed in [65] captures the structure of feedback/feedforward cortical interactions in the case of vision, schematized in Fig. 1.9(a), with the elegant probabilistic model summarized in Fig. 1.9(b). Here different visual areas are schematized as boxes implementing stochastic variables, and linked as a Markov Chain. The activity in **V1**, x_1 , is influenced by the bottom-up feedforward data x_0 and the probabilistic priors $P(x_1|x_2)$ fed back from **V2**, and so on.

The approach described above was originally proposed for image analysis; it has been then extended to model the reciprocal interactions of brain areas along the whole visual system, including the main centers for gaze control, to account for eye movements in visual tasks

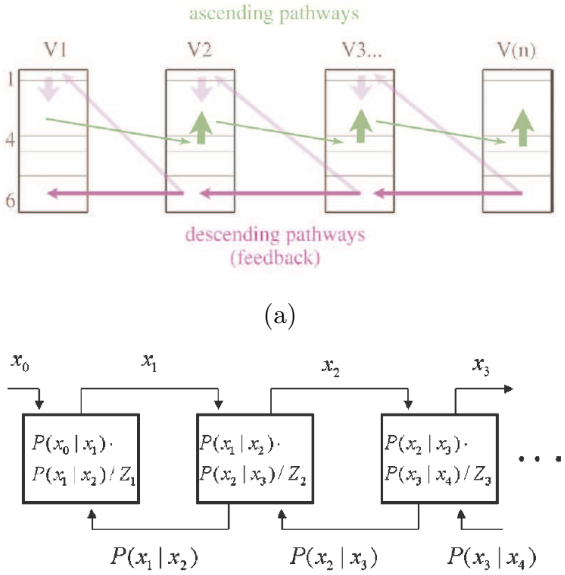
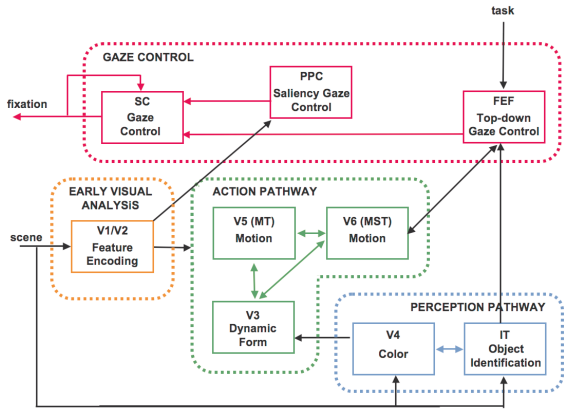


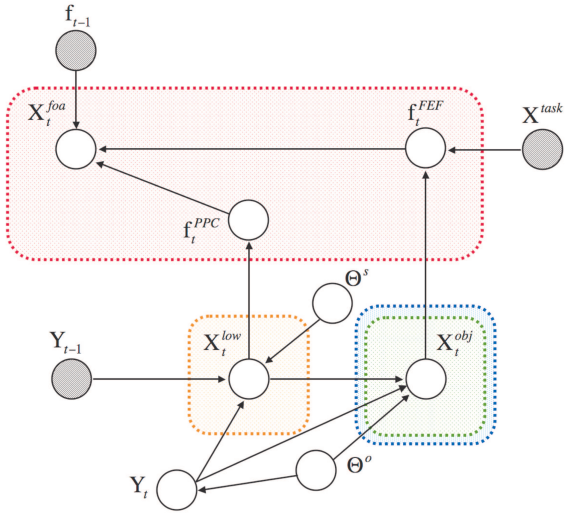
Figure 1.9: 1.9(a) Diagram illustrating how ascending and descending streams may operate semi-autonomously within a serial pathway, after [114]. 1.9(b) Schematic of the corresponding hierarchical Bayesian inference framework in the cortex, after [65].

[14].

A schematic overview of the brain centers for saccade generation was given in [104].The model in [14], schematized in Fig. 1.10(b), reproduces the interdependence of only the main brain areas involved in gaze control, but proposes as well a detailed account of the connection of such oculomotor areas to those related to the analysis of visual input (Fig. 1.10(a)). Each area is represented as the process that computes the probability distribution of a stochastic variable, and gaze-related variables are then computed via the *Maximum A Posteriori* (MAP) rule. In particular, this model is able to include three *levels* of gaze control, namely the *Superior Colliculus* (SC) at the lowest level, *Posterior Parietal Cortex* (PPC) at the middle level, and *Frontal Eye Fields* (FEF) at the highest level. Furthermore, it shows how the two visual streams related to Action processing and Perception contribute to gaze control separately in PPC and FEF respectively, to eventually cooperate/compete for gaze control in SC. On one hand, the **PPC** module, given the low level information X_t^{low} , computes a candidate fixation f_t^{PPC} via MAP rule on $P(f_t^{PPC}|X_t^{low})$.



(a)



(b)

Figure 1.10: 1.10(a): interconnections among the main brain areas involved in saccade generation. 1.10(b): graphical structure of the corresponding Bayesian Network. Adapted from [14] and [79]

On the other hand, the **FEF** module takes into account information about the task, X^{task}, giving a saliency to the objects detected in the scene X_t^{obj} (e.g., faces), and selects a location f_t^{FEF} via MAP rule on $P(f_t^{FEF}|X_t^{obj}, X^{task})$. Eventually the saliency's and object's parameters are denoted by Θ^s, Θ^o and regulate the extraction mechanism of the specific features.
The target f_t where the focus of attention (FOA), X_t^{foa} is set, is eventually selected by taking into account the probability that the stimulus

in the response field is the target from prior information, through a posteriori analysis of the sensory cue to target information, and that a location at time $t-1$, f_{t-1}, was explored. In other words, at time t, the target position is estimated in terms of the MAP probability of focusing a location X_t^{foa}, then,

$$f_t = argmax P(X_t^{foa} | f_{t-1}, f_t^{PPC}, f_t^{FEF}) \quad . \tag{1.6}$$

1.4.4 Bridging the gaps

Although the real neural architecture of the areas that contribute to eye movement generation is highly complex, the survey of existing models of gaze control have highlighted three main regions, namely **SC**, **PPC** and **FEF**. The joint operation of these three regions, together with their feedforward/feedback connections from/to visual areas, constitute a possible mechanisms that facilitate the integration and coordination of the processes corresponding to the opposite poles of the dichotomies described in section 1.3.
In particular, **FEF** — the highest level of gaze control, in terms of number of 'steps' that separate it from the brainstem where motor signals are forwarded to the muscles — projects to SC accounting for top–down influences, both action–related (resulting from input from the dorsal stream to **FEF**) and perception–related (resulting from input from **IT** to **FEF**). **PPC** — the middle level of gaze control — contributes mostly in terms of reflexive saccades towards non–anticipated targets; thus, it accounts for an action–related contribution to eye movements that incorporates both bottom–up (saliency) signals and behavioral goals (something that can be situated between purely stimulus driven gaze shifts and those generated purely by cognitive activity). Finally, **SC** acts as the lowest–level hub for gaze control, also mediating a strategic mechanism such as *Inhibition Of Return*.
Fig. 1.11 gives a schematic positioning of the three gaze centers along the dichotomic axes discussed in section 1.3, showing that all the poles are covered.

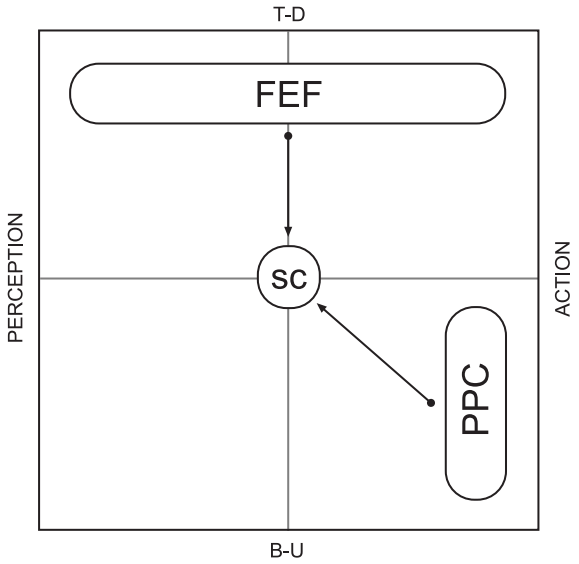


Figure 1.11: A schematic depiction of the three gaze centers, positioned in a metaphoric reference frame defined by the dichotomic axes discussed in section 1.3. The processes that guide eye movements span all the poles.

1.5 Eye Movements In The Guidance Of Action

Differently from studies of eye movements in purely visual tasks, dealing with *visuomotor* tasks requires a shift of perspective: the main difference in such cases is that eye movements should not be treated as entirely independent from movements of other parts of the body. Situations in which the viewer is also engaged in carrying out some action are the most common in daily life, and in these cases the pattern of eye scanning must be integrated in an overall action sequence.

1.5.1 Novel experimental paradigms: natural complex actions

A famous visuomotor task involving manipulation of objects was originally devised by Ballard and colleagues [6, 7] on a mouse–controlled computer screen, later on reproduced with real world objects [87]. The *block–copying* task consisted in copying a certain disposition of colored blocks (the *model*) onto a *workspace* initially blank, by selecting and picking up colored blocks from a *resource*.
The main observation in this experiment is the regularity in the rhyth-

mic pattern of eye–hand movements, within each component subtask. More specifically, it appears that temporal coordination is controlled by the availability of the eye, with hand movement delayed with respect to the relevant fixation.
The interpretation given by the authors is usually referred to as the theory of *deictic pointers*: according to this view, eye fixations are used deictically (i.e. as pointers) to visual information that is relevant to the current subtask. In this way, eye movements are guided by task demands, and become a means to select objects for subsequent actions, thus allowing a minimization of the memory and computational loads, and definitely eliminating the need for a detailed internal representation.

A number of studies carried on in sports [61, 62, 63] and natural visuomanual actions [64, 86], support the view that eye movements in most cases are not simply attracted towards visually salient locations, but rather are determined by the demands of the motor task. These experiments have shown that the eye tend to fixate locations of the scene where task–relevant information is found, and that not only fixations often anticipate motor actions (*lookahead fixations*), but also, as in the case of table–tennis or cricket, they predict the locations that relevant objects will occupy in the near future.
Land's analysis of natural actions introduced the concept of *object–related actions* (ORA). ORA's are most often carried out sequentially, and involve engagement of all sensorimotor activity on the relevant object, including a number of fixations and subsequent manual actions. Furthermore, Land provides a tentative taxonomy of fixation's functions in manipulative tasks, using the following categories: *locate* an object to be used; *direct* fixation to the object that is going to be manipulated; *guide* relative motion of two objects that must be put in contact; *check* the state of an object.

Another influential work in this field was presented by Johansson and colleagues [53]: they recorded eye and hand movements in a task that involved picking an object, bringing it to a target and avoiding an intermediate obstacle, and bringing it back to the initial position. The aim of this experiment was to explore the precise spatial and temporal relation between gaze fixations and ORA's. In accordance with the above mentioned studies, Johansson and colleagues found that fixations are almost always directed to critical landmarks for action control, and that gaze usually leads hand movements. Furthermore,

they made the hypothesis that fixations are used to obtain *spatial information* for controlling manipulatory actions, and indeed in their experiment subjects always fixated contact points before moving the hand there; often, this was also the case with fixations directed to the intermediate obstacle, thus supporting he idea that fixations play the role of *spatial keypoints* for hand/object trajectory.
Regarding temporal relations, the timing of saccades away from a landmark is found to be determined by critical kinematic events happening at that landmark (e.g. grasp contact), supporting the idea of a discrete event–driven sensory control.
A last point worth mentioning here is that it is largely accepted that motor output during dexterous manipulation largely relies on predictive control mechanisms, the formation and updating of which depend on correlations between motor output signals and their sensory consequences as established by experience (this point is largely discussed in chapter 2). The results presented in [53] suggest that the anchoring of gaze–hand coordination to contact points constitutes a mechanism for managing correlations between visual and somatosensory information, and efferent copy signals required for predictive motor control.

1.5.2 Modeling: Visual routines and the allocation of visual resources

Following the experimental evidence presented in section 1.5.1, Ballard and colleagues are developing a computational model that tries to capture the role of eye movements in visuomotor tasks [118, 49]. Before describing the model, it should be made clear that such model aims at capturing *not* the spatial targeting of saccades, but rather the temporal scheduling of movements related to different subtasks.
The starting point for this approach is the observation that in a complex visuomotor task, the eyes are to be considered as a physical resource that can only be allocated *sequentially* to support the resolution of different subtasks. In other words, concurrent pieces of the task may compete, at a given time, for getting control of sensory resources and obtaining the required information.
From a computational standpoint, this fact can be associated with the concept of *visual routines* [129]: visual routines can be described as sensorimotor programs that keep control of the eye, implement all the necessary processing of the visual input, and extract relevant information for a given task (perceptual or motor). In the framework

proposed in [118, 49] fixations are considered as part of visual routines, and the model aims at understanding how different routines associated with different tasks are managed in time, rather than describing the spatial oculomotor behavior. Thus, any routine, or *behavior*, at a given time t has the ability to:

- direct the eye
- perform appropriate visual processing
- choose an appropriate action course

The solution envisaged by the authors to the problem of managing concurrent behaviors is based on the idea that eye movements serve to reduce uncertainty about task–relevant environmental variables, and borrows techniques from the theory of *reinforcement learning* [121]: a value is assigned to a behavior by estimating the expected reward of taking the related action, and the expected cost of the uncertainty that will result if the related eye movement is not made; then at any time step, the behavior with highest expected value is chosen.
To see this in practice, the authors model behaviors as Partially Observable Markov Decision Processes (POMDP); each behavior is a 4-tuple (S, A, T, R), where:

$$S = \textit{state space} \tag{1.7}$$
$$A = \textit{action space} \tag{1.8}$$
$$T(s, a, s') = \textit{transition probability from } s \textit{ to } s', \textit{ through action } a \tag{1.9}$$
$$R(s, a) = \textit{expected payoff for taking action } a \textit{ in state } s \tag{1.10}$$

Furthermore visual information regarding the state of the agent should be considered as noisy, and this is modeled by estimating the state with a noisy Kalman filter according to the system dynamics; this allows to keep track of the increase in uncertainty due to a missing Kalman update (which happens when sensory resources are allocated to another behavior).
Then, the goal is to find the *optimal policy*

$$\Pi^{\star} : S \longrightarrow A \tag{1.11}$$

so as to maximize *discounted long term reward* [121]. Standard learning techniques allow to discover the *optimal value function* $Q(s, a)$, namely the expected discounted reward if action a is taken in state

s, and the optimal policy is followed thereafter; then the agent, given its state estimate s, can behave optimally by always choosing:

$$a^{\star} = argmax\, Q(s, a)\,. \tag{1.12}$$

Finally, uncertainty is taken into account by evaluating the cost (or *loss function*) associated to an action that is optimal with respect to the state estimate s, but suboptimal with respect to the actual state of the environment.

1.6 Motor Control

According to [54] "The study of motor control is fundamentally the study of sensorimotor transformations." Such a position stresses the crucial role of sensory information in motor control: Wiener's original idea of feedback control as the basis for intelligent behavior, is still central in most recent explanations of biological movement. In this section we try to analyze how sensory feedback is actually used in existing models, and which features of biological systems are missing in such models.

1.6.1 The Optimal Control Framework

Most successful motor control models are those based on Optimal Control, that have yield accurate descriptions of observed phenomena; this class of models finds an element of biological plausibility in the fact that the sensorimotor system is shaped by processes, such as evolution, adaptation and learning, that act to improve behavioral performance.
Given a task, the problem of motor control is that of generating/selecting motor signals that produce the appropriate movement, in terms of kinematics and dynamics, among the infinite possible movements that biomechanical redundancy allows for. In Optimal Control models this is cast as an optimization problem, where the appropriate movement is the one that satisfies a performance criterion, i.e. minimizes a *cost function*.
Formally, the system is described as a dynamical system with state variable(s) $x \in R^m$ and control signal $u \in R^n$, where

$$\dot{x}(t) = f(x, u, t) \quad . \tag{1.13}$$

The performance criterion, or cost function, takes the general form of an integral over the time interval of the movement to account for the fact that cost is accumulated during execution:

$$J_0 = \Phi(x(T), T) + \int_0^T L(x(t), u(t), t)dt \quad . \tag{1.14}$$

On the right side, T represents the terminal time, and the first term is the terminal cost, which is a function only of the terminal (i.e., final) state, $x(T)$. Minimization of this functional is related to the minimization of action in Lagrangian mechanics, and so $L(x, u, t)$ is

also called the Lagrangian here.
The problem of optimal control is to find the control law $u^\star(t)$ that leads the system along a trajectory $x^\star(t)$ in the state space such that the functional (1.14), usually called cost function, is minimized.
It is immediately clear that a critical point is the choice, and the theoretical justification, of a particular cost function; indeed, the many optimal motor control models are named after the chosen cost function (minimum jerk model, minimum torque model, minimum variance model, ...). As a matter of facts, more recent models tend to adopt a cost function that combines different criteria, in order to be able to model a wider class of observed phenomena.
However a more fundamental characterization for existing models concerns the type of control law, which leads to the distinction between Open Loop and Closed Loop optimal control.

Open Loop Control

In Open Loop, or feed–forward, models the optimality criterion is applied to plan the best trajectory in the state space, while feedback from the sensory apparatus is involved only in the execution phase to correct deviations from the desired trajectory.
In this section we briefly review some of the main open loop models for biological movement.
In the analysis of point–to–point hand movements, the first well–known optimal control model was based on minimization of *jerk* [50, 36], a quantity related to the temporal derivative of the acceleration. Such a model produced smooth movements by using a *kinematic* cost, thus predicting an optimal trajectory in the cartesian space, as well in the space of velocity and higher order temporal derivatives. Its success consisted mostly in the ability to reproduce the observed bell–shaped velocity profile, local isochrony [133], and the phenomenological two–thirds power law [60].
Subsequently, models based on *dynamic* cost were introduced (e.g. [131]), that used dynamic variables such as joint torque change or muscle tension. The main difference with kinematic models is that in that case computation of the optimal movement involves the explicit positions and velocities as a function of time, thus neatly dividing the planning stage and the execution stage. Conversely, the solution to dynamic models are the motor commands required to achieve the movement and therefore planning and execution are no longer separate processes.

Although the above mentioned models were able to yield good predictions in some specific class of movements, they all suffered the lack of a principles justification for the choice of the cost function: there is no evident reason why the CNS should choose to optimize such quantities as jerk or torque change, neither is it clear how the CNS could estimate such quantities and integrate them over time.
An effective way to circumvent those limitations was proposed for eye and hand point–to–point movements [48], based on the physiological observation that motor noise is signal dependent, since neural control signals are corrupted by noise whose variance increases with the size of the control signal [120]. The *minimum variance* model chooses the sequence of muscle activations so as to minimize the variance in the final position. Not only such choice appears more biologically motivated and the involved costs are directly available to the CNS so that the optimal trajectories could be learnt from experience of repeated movements; but it also naturally implies smooth movements, thus giving a principled motivation to minimum jerk and torque change models.
Finally, another class of 'ecological' models based on a dynamic cost is common in biomechanics and locomotion, where most models minimize *energy* used by the muscles [89, 2], through cost functions that increase supra–linearly with muscle activation.

Closed Loop Control

While open loop models have been able to account for behavioral observations averaged over multiple trials of a task, they revealed poor capabilities in modeling inter–trial variability.
Indeed more recent motor control models [124] adopt techniques of stochastic optimal closed loop (or feedback) control, and use online feedback in a simultaneous planning / execution stage.
In such models feedback is not used by a predefined control law that corrects deviations from the optimized trajectory; rather, performance is optimized over the family of all possible feedback control laws. The (time–varying) best possible transformation from states of the body and the environment into control signals is constructed by the feedback controller.
Temporal difference reinforcement learning techniques are usually adopted: the optimal feedback law is approximated as the one that minimizes the total expected cost, given the current state, and assuming that the optimal control law is applied until the end of the movement.

The function expressing total expected cost is usually called *optimal cost–to–go*, and it includes a term encoding the error in task completion and a term penalizing effort to account for signal dependent motor noise. This corresponds to a mix of the minimum energy and minimum variance models discussed in section 1.6.1, and allows to approximate the best control scheme for the given task.
Indeed, the most striking result is that the optimal feedback controller obtained this way obeys the *minimal intervention principle* [125]: only errors in task–relevant directions are corrected, while allowing variability in the redundant (task–irrelevant) subspace of the state space, the so called *uncontrolled manifold*.

However sensory feedback in biological systems is known to be corrupted by noise and delays. An important point in the closed loop approach is the way it deals with such noise.
As indicated in figure 1.12, the controller must rely on a state estimate that integrates all available information from sensory data, recent control signals, knowledge of body dynamics, as well as earlier outputs. It has been shown that the resulting controls are optimal when the estimator is also optimal, that is Bayesian (see section 1.6.2).

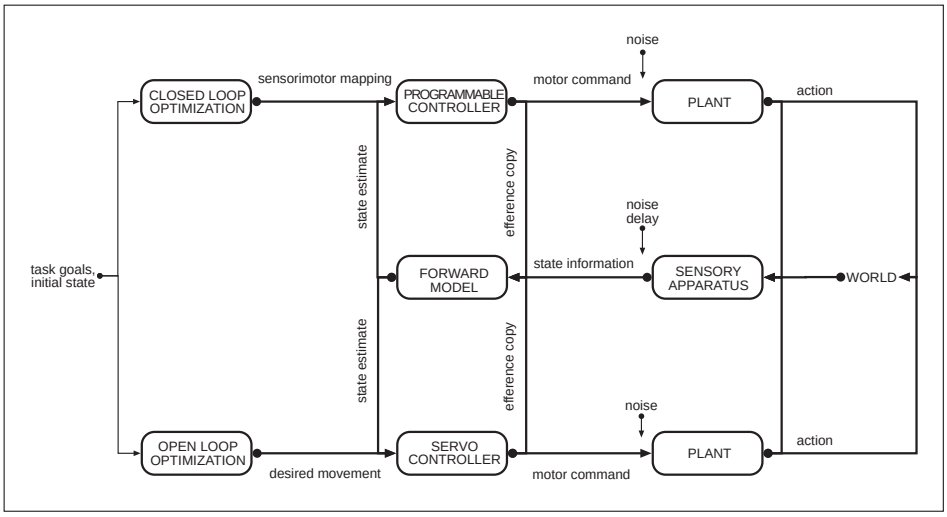


Figure 1.12: Direct comparison of Open Loop (above) and Closed Loop (below) architectures for motor control. Notice that the main difference is the kind of controller resulting from the optimization process: in Open Loop, this is a fixed servo controller, while in Closed Loop it is programmed online.

1.6.2 Internal Models and State Estimation

Internal models can be thought of as a form of knowledge (representation) that the system possesses about the sensorimotor transformations that it realizes.
Inverse models describe how, given the current state of the system, a desired final state is transformed into appropriate motor commands. Since such models produce the motor commands necessary to obtain a result, they have a natural use as controllers.
In fact, usually in Open Loop control the feedforward controller plans the desired trajectory in the state space, and an inverse model generates the sequence of control signals that will drive the system along that trajectory.
However, it has been argued [124] that in Closed Loop control the concept of inverse model is unnecessary, since the job of the optimal feedback controller is exactly to construct the sensorimotor mapping, namely transforming task goals into motor command.

Conversely a *forward model*, given an initial state, is the mapping of a motor command to the next state (dynamical forward model) or to sensory feedback that will result from the observation of such next state (output forward model).
Forward models have been used in different ways, both in Open Loop and Closed Loop control. In any case, however, their function is to provide an estimate of the state of the dynamical plant, or of its sensory consequences. The theoretical justification for the introduction of such an estimator, is that feedback control in biological systems is subject to potential instabilities, especially in fast movements, due to noise and delay of sensory feedback; then an estimate must be used, alone or combined with raw sensory data, to maintain stability.
In [73] an output forward model is the essential component of a so called Smith predictor. The forward model provides internal feedback of the predicted outcome of an action before sensory feedback is available. The internal feedback is used by the controller in an inner loop to correct predicted deviations from the desired trajectory, before sending new motor commands to the plant; then when sensory feedback becomes available, the previously predicted feedback is subtracted, and only the unpredictable component of the error is corrected by the feedback controller.
Moving to dynamical forward models, in [134] a modular Open Loop control architecture is presented, which adopts several couples of in-

verse and forward models as a mechanism for context selection. In each module, the inverse model provides motor commands to get to the desired state, while the forward model generates a predicted final state; such state is compared with the desired state, and the module (or subset of modules) which produced the closest final state is selected as the most appropriate one for the given context.
On the other hand, in Closed Loop control, the state estimates provided by forward models have been used online to select the optimal feedback controller (i.e. the sensorimotor mapping) [125]. Notice that in this context the forward model must satisfy a precise constraint: in fact, for the resulting controller to be optimal, also the state estimator must be optimal, i.e. Bayesian [57]. In practice, the Kalman filter is often used because it is optimal under the assumption of linear dynamics and gaussian noise.
As a last remark, it should be noticed that the adoption of a state estimator in optimal motor control schemes, not only allows to combine feedback coming from internal (the efference copy of motor commands) and external (sensorial apparatus) sources, but also leaves open the possibility that data from multiple sensory modalities contribute to state estimation. What specific type of sensory data are used, and how such data can be integrated, are the issues discussed in section 1.6.3.

1.6.3 Characterizations of the Sensory Feedback

In normal conditions, information from multiple sensory modalities is available to all higher organisms, raising the question of how such signals are combined to guide behavior.
This question comprises several issues that can be analyzed separately, although they do not necessarily correspond to separate neural processes. First, within a single modality suitable features must be extracted from the input stream. Second, signals from different modalities must be converted to a common representation (the *coordinate transformation problem*). And third, signals in the common representation must be fused using some criterion.
Although the role of sensory feedback in motor control has been extensively studied, only recent experimental work has addressed the first problem, i.e. which features are relevant to the control task. In the case of online visual feedback on hand movement, [103] have shown that position and motion information are both used by humans, and they are combined in a manner consistent with the reliability of

the signals. Notice that neurophysiological studies (e.g. [41]) have shown that such quantities like hand motion direction and position are actually encoded in the activity of large populations of neurons in primates.
The problem of coordinate transformation has been cast historically in the framework of classical information theory, suggesting that mutual information, i.e. the information common to two transformed signals, can be used as the basis for an optimization algorithm that extracts information from multiple input streams [9]. Such model can also be augmented to incorporate topological order [42].
The third problem, i.e. multimodal integration, is the one that has attracted most attention, and the most successful approach follows an optimization paradigm that is similar to the Open Loop control schemes: integration is realized according to the minimization of a biologically motivated cost function. It has been shown experimentally that humans integrate visual feedback with auditory [43] or haptic [29] cues in a way that can be predicted by a maximum likelihood estimate, which corresponds to minimization of uncertainty (or variance) in the final estimate. Further experimental work [10] has shown that also visual and proprioceptive feedback on hand position (in a plane) are integrated according to a weighting scheme related to the direction–dependent variance in the unimodal information.
Extending this approach, further explorations on the integration of vision and proprioception [117] have shown that the input statistics alone is not sufficient to explain the following observation [116]: during reach movements, the relative weighting is variable, with a planning stage mostly relying on visual signals, and the computation of intrinsic motor commands relying more on proprioception. An explanation is given reconsidering the coordinate transformation problem: it appears that the transformation of sensory signals between coordinate frames is inherently noisy, and sensory integration follows as a unifying principle the minimization of such noise.

1.6.4 Fundamental Limitations of the Existing Models

Optimal motor control models have been devoted much attention, and in the last decade have reached a significant degree of sophistication, especially with the introduction of internal models.
In particular, the use of forward models allows one to deal with the

problem of noise, which is a fundamental one because both sensory and motor control signals are known to be affected by signal dependent biological noise, and furthermore sensory information is usually processed by the CNS with significant delays with respect to the motor signals that should be controlled.
Recently, output forward models have been reconsidered as a mechanism for endowing robotic systems with Expected Perception (EP) capabilities [21, 20], namely the ability to predict the sensory consequences of future motor actions, on the basis of previous perceptions, efference copies of motor commands, proprioception, and forward models; in this context, EP can be seen as an attentive process aimed at reducing the computational load that would result from elaborating the whole sensorial input stream.
That said, it should be noticed however that all motor control models discussed above always consider the sensory apparatus as passive. Even in the sophisticate framework of forward models and EP, where the perceptual system is somehow 'active' thanks to the ability to estimate and predict, the possibility that motor commands are issued in order to orient the sensors towards (task–)relevant information, is never considered. Only recently such possibility is becoming object of research [115], with the introduction of the distinction between acts of exploration (i.e. motor signals issued in order to gather information) and acts of exploitation (motor commands that use the information to pursue the task objectives). However in [115] the active exploration is just made of random movements, while the focus is on optimizing the timing and relative weight of the two kind of movements.
We argue that the 'intelligent' (i.e. not purely random) use of active exploration is a fundamental ingredient of intelligent motor behavior in biological systems, and we argue that it is an important missing component in most existing models of motor control. In the present chapter we have discussed existing approaches to study this specific issue, focusing on the visual modality: those constitute the so called *Active Vision* framework, i.e. attempts to understand the biological basis and to model the mechanisms of intelligent, overt allocation of visual resources in complex tasks. In this context however the problem of motor planning and control is usually not considered.
In chapter 3 we will focus on the eye–hand system, and discuss a new model that accounts both for how the dynamics of the motor actions and the incoming sensory information could be integrated to guide eye movements, and how the behavior and output of the active vision system could contribute to action control. We surmise that such model

represents a first step towards reconciling the active vision approach and the optimal motor control approach.

Chapter 2

Sensorimotor Coupling. A Bayesian Framework For Eye–Hand Coordination

2.1 Introduction

The general issue of sensorimotor coordination is usually treated only in one direction: either from the point of view of active perception, or in the framework of motor control with or without online feedback (see chapter 1).

In the case of motor control theories, existing models usually reflect the functional architecture of the primate cortico-cerebellar system [95] . Most acclaimed computational models cast the issue of movement planning and execution as an optimization problem [124, 125]. Optimality means minimization of a scalar function that depends on control signals as well as on the current state of the musculo–skeletal system and environment: such function can be e.g. jerk [50, 36], energy [89, 2], or variance [48]. Closed loop models are those where optimization is performed online, on the basis of sensory feedback [8], possibly integrating information from multiple sources [103, 9] and, in case of noisy or delayed sensors, integrating it with state estimates [10, 29] through internal models [134, 21]. In such framework the sensory apparatus is always considered as *passive*.

On the other hand, in the case of active perception, the object of study is the overt attentional process, namely how sensory resources

are allocated — e.g. via eye movements — so as to facilitate the accomplishment of a given task. Here we focus essentially on vision.
In the case of purely visual tasks, several computational models have been proposed that reflect the functional organization of the primate visual system, and generate saccadic movements on the basis of the image properties alone [51] or combined with top–down cognitive influences [92, 67, 31, 14]. Eye tracking research in complex visuomotor tasks [87, 61, 63] has highlighted the reciprocal influence of the motor task and the oculomotor behavior, showing that most fixations are targeted to extract information that is relevant to the execution of the task [49]. As opposed to research in motor control, the active vision approach indicates a strong relation between active vision and action.

It is worth noting also that active vision subserving action can be considered as a form of spatial attention, and according to the premotor theory of attention [100] spatial attention is the consequence of motor preparation. Gaze–shifts can be considered as the motor realization of overt shifts of attention, which in turn — according to most recent theories [110] — arises from the activation of those same circuits that process sensory and motor data. In particular, selective attention for spatial locations [99] is related to the action pathway of the (dorsal) visual system, which is mainly devoted to trigger prompt actions in response to environmental varying conditions, and mostly relies upon motion analysis and gaze–shift control at low and intermediate levels [74]; selective attention for objects [24] derives from activation of ventral cortical areas involved in the perception pathway, responsible for object property processing, such as the analysis of form in association with color, and the visual perception/identification of objects (e.g., faces), with tight integration to high–level, cognitive tasks of frontal areas [45].
Clearly, the two pathways are not segregated but cooperate/compete to provide a coherent picture of the world. Gaze control, is the ultimate product of such integration effort: the target where the fixation is eventually set, is selected by taking into account the probability that the stimulus in the response field is the target from prior information, through a posteriori analysis of the sensory cue to target information [14].

Yet, we lack a well defined framework for integrating active vision models with motor control strategies. We are faced with two comple-

mentary problems: how the dynamics of the motor actions and the incoming sensory information could be integrated to guide eye movements, and how the behavior and output of the active vision system, including eye movements, contribute to action control.

In the present chapter we discuss a functional model of sensorimotor coupling that, in perspective, could be considered as a first step towards reconciling the active vision approach and feedback motor control approach. We will focus specifically on eye–hand coupling, and in the following chapters we will implement the model in a drawing task (chapter 4), and compare the results with eye–tracking analysis in the same task (discussed in chapter 3).
The model extends a previous one [19, 18], whose aim was to simulate the scanpath of the draughtsman. In section 2.2 we discuss the overall functional model, and analyze the modules related to sensory processing and to the control of the desired eye–hand movements. The core of the model however is the module that realizes a premotor coupling between eye and hand: this module, discussed extensively in section 2.3, collects inputs from the external sensory modules, and feeds its outputs, namely premotor information, to the subsequent modules responsible for the detailed control of eye and hand motor signals.
It is formulated in terms of a Bayesian generative model and its corresponding graphical model. The rationale behind the adoption of a probabilistic, Bayesian framework grounds in the fact that signals in sensory and motor systems are corrupted by variability and noise, and the nervous system needs to estimate these states [57]. This overall uncertainty places the problem of estimating the state of the world and the control of the motor system within a statistical framework.

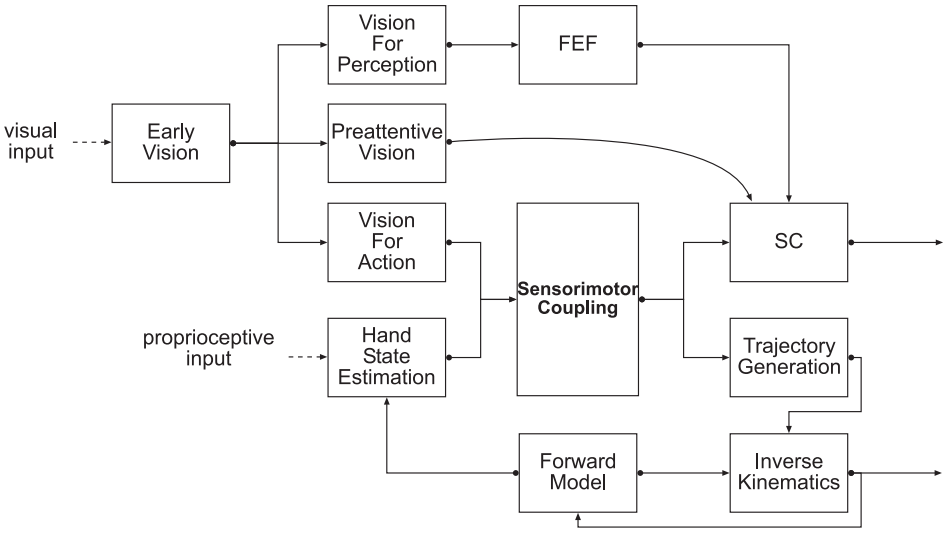


Figure 2.1: A schematic model for sensorimotor coordination of a robot involved in visuomanual tasks (see text for explanation). Each module roughly corresponds to a functional area of the primate brain; in particular, FEF (standing for Frontal Eye Fields) and SC (Superior Colliculus) refer to two of the main areas involved in gaze control.

2.2 Aims and overall functional model

The functional model depicted in Fig. 2.1 proposes a general organization of visual, motor and visuomotor modules for the coordination of an artificial agent in a visuomanual task, on the basis of existing pieces of knowledge and models of the corresponding functional areas of the human brain, as discussed in previous chapters.
The first aim of the model is to provide a modular system that — once the processes taking place in each module are specified — can generate hand and eye motor signals and control the actual movements (corresponding to the modules called *SC*, standing for the low–level gaze control functions usually attributed to the Superior Colliculus, and *Trajectory Generation* and *Inverse Kinematics*, functions that are commonly attributed to motor and cerebellar areas). As suggested by most recent theories of motor preparation [110], a premotor stage should be considered as well, where hand–related and eye–related signals are coupled (the core module called *Sensorimotor Coupling*).

Furthermore, it aims to provide a general mechanism to combine bottom–up strategies (accounted for by the *Preattentive Vision* module, discussed at length e.g. in [51]) and top–down strategies, cor-

responding to the *FEF* module (FEF stands for Frontal Eye Fields, which are known to be related to high level gaze control) which receives inputs from the *Vision For Perception* module (responsible for object property processing, such as the analysis of form in association with color, and the visual perception/identification of objects (e.g., faces), with tight integration to high–level, cognitive tasks of frontal areas).
Orthogonally, the model aims as well to combine Action-related and Perception-related visual processes (respectively the *Vision For Action* and *Vision For Perception* modules), whose outputs are fed, through intermediate modules, to the *SC* where the integration is performed.

Eventually, the central module (*Sensorimotor Coupling*) allows also to combine multimodal sensory information, i.e. the results of visual processing in the *Early Vision* and *Vision For Action* modules, and proprioceptive information, integrated in the *Hand State* module with the efferent copies of motor commands through the *Forward Model*.

We will not discuss in this chapter the details of each module, which will be explained in chapter 4 in the specific context of a drawing task. Rather, here we propose a mathematical formulation of the core module, namely *Sensorimotor Coupling*, which in our opinion could provide a first step towards a general mathematical framework for targeting the aims itemized above, and for reconciling the active vision approach and the feedback motor control approach.

2.3 A Dynamic Bayesian Network for eye–hand coupling

In any real world implementation, the processes corresponding to the modules of the architecture discussed in section 2.2 should be treated as inherently noisy, due to errors in the sensory and motor apparatuses as well as to signal–dependent noise in neural signals [48] in the case of biological systems.
Thus an agent adopting the above architecture should also be able to deal with uncertainty; to this end, we resort to a probabilistic Bayesian framework and consider the values of the relevant variables as realizations of corresponding random variables. This way we can map the core of the functional model outlined in Fig. 2.1 into the graphical model shown in Fig.2.2, where nodes denote random variables, and arrows, conditional dependencies.

Since we are dealing with a process unfolding in time, the network we designed is in the form of a Dynamical Bayesian Network (DBN [77]) and the graph depicted in Fig. 2.2 pictures two temporal slices. The process corresponding to the temporal evolution of the eye plan is modeled in our network as a discrete–time I*nput–Output Hidden Markov Model* (IOHMM), as suggested e.g. in [31]; in the IOHMM we have three layers of variables, i.e. input variables u^e, u^h related to vision and proprioception respectively, and the hidden and out-

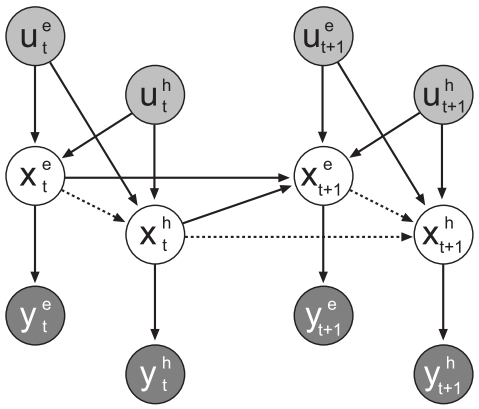


Figure 2.2: The IOCHMM's for combined eye and hand movements. The gray circles denote the input ($\boldsymbol{u}$) and output ($\boldsymbol{y}$) variables. Continuous connections in the hidden layer denote the core process relating hand movements to previous eye movements, while dotted connections highlight the subgraph that represent the complementary process.

put variables x^e, y^e related to eye movements. Similar considerations hold for the hand movement, where the inputs are the same while the hidden and output variables are denoted x^h, y^h. The inputs are collected from external sensory modules, and the DBN feeds its outputs, namely premotor information, to the subsequent modules responsible for the detailed control of eye and hand motor signals.

The very point to be remarked here is that, following the discussion of section 2.2, the two processes should not be considered as independent, but rather as two coupled chains: this results in a graphical model that unifies the IOHMM Model and another DBN known in the literature as the Coupled HMM [77, 140, 97]. Thus we call the DBN represented in figure 2.2 an *Input–Output Coupled Hidden Markov Model* (IOCHMM). To the best of our knowledge this kind of architecture has never been exploited in the literature, and in particular for sensorimotor modelling.
Notice that the coupling, that in Fig. 2.1 is implicit in the module we called *Sensorimotor Coupling*, here unfolds in time in such a way that at any time step the hand plan depends on the current eye plan, while the eye plan depends on the previous hand plan.

This way, the process that accounts for the probability of the current eye movement information conditional on previous eye and hand signals, can be formally modeled as the probability distribution $p(x^e_{t+1}|u^e_{t+1}, u^h_{t+1}, x^e_t, x^h_t)$. Similarly, we can write the probability of the current hand signals given the current eye signals and the previous hand signals, as $p(x^h_{t+1}|u^e_{t+1}, u^h_{t+1}, x^e_{t+1}, x^h_t)$. In the vocabulary of HMM's, the above functions denote *state–transition* probabilities, and represent respectively the (premotor) influence of hand motor preparation on subsequent eye movements, and the role of active vision in the guidance of subsequent hand actions.

By considering again the dependencies in the graphical model, we can write the statistical dependence of the eye output signal on the corresponding state variable as the distribution $p(y^e_{t+1}|x^e_{t+1})$; similarly for the output hand movement, we can write the density $p(y^h_{t+1}|\, x^h_{t+1})$. Both represent *emission* probability distributions. Notice also that the input nodes have no probability distribution associated, as their values are provided by endogenous processes that are not modeled within this network.

Eventually, by generalizing the time slice snapshot of Fig. 2.2 to a time interval $[1, T]$ we can write the joint distribution of the state

and output variables, conditioned on the input variables as:

$$\begin{aligned} p(\bar{x}_{1:T}, \bar{y}_{1:T} |\, \bar{u}_{1:T}) \;=\; & p(x_1^e |\, u_1^e, u_1^h) p(y_1^e |\, x_1^e) p(x_1^h |\, u_1^e, u_1^h, x_1^e) p(y_1^h |\, x_1^h) \\ & \cdot \prod_{t=1}^{T-1} \Big[p(x_{t+1}^e |\, u_{t+1}^e, u_{t+1}^h, x_t^e, x_t^h) p(y_{t+1}^e |\, x_{t+1}^e) \\ & \cdot p(x_{t+1}^h |\, u_{t+1}^e, u_{t+1}^h, x_{t+1}^e, x_t^h) p(y_{t+1}^h |\, x_{t+1}^h) \Big] , \end{aligned} \tag{2.1}$$

where $\bar{u}_{1:T}$ denotes the input sequence from $t = 1$ to T, $\bar{x}_{1:T}$ denotes the pair of state sequences $(x_{1:T}^e, x_{1:T}^h)$, and similarly for $\bar{y}_{1:T}$.

2.4 Movement selection as Bayesian inference and decision

To use the IOCHMM as a control device for an artificial agent, we must contend with three problems: learn the parameters of the model; use the model for inference (i.e., to compute the expected hidden states for each time slice), and exploit inferences to make decisions. In order to do this, we first recall how Bayesian inference is formulated in a standard HMM, and then extend the results to IOHMM's and to our IOCHMM; we then outline the decision criterion adopted, and the learning stage.

2.4.1 Inference

Consider a HMM with hidden and observed variables denoted by X_t and Y_t respectively. In the following, we discuss only the case of discrete variables. Generally speaking, inference consists in evaluating the probability distribution of the hidden state conditioned on the observations. Two special cases are of interest here, namely *prediction* and *filtering*.

Prediction is the term reserved to inference of the *future* hidden state, given all past and current observations, namely evaluating $p(X_{t+h} \mid y_{1:t})$, where $h > 0$ is how far we want to look-ahead. We can without loss of generality set $h = 1$, then by simple application of the rule of total probability, this can be written as:

$$p(X_{t+1} \mid y_{1:t}) = \sum_{x_t} p(X_{t+1} \mid x_t)\, p(x_t \mid y_{1:t})\,. \tag{2.2}$$

In the r.h.s. the first term corresponds to the *transition model* and the second term is the so called *belief* state, denoting knowledge of the current hidden state based on all past and current observations. Once the distribution written in eq. 2.2 has been evaluated, we can easily convert this into a prediction about the future observations by marginalizing out X_{t+1}:

$$p(Y_{t+1} \mid y_{1:t}) = \sum_{x_{t+1}} p(Y_{t+1} \mid x_{t+1})\, p(x_{t+1} \mid y_{1:t})\,. \tag{2.3}$$

Notice that the solution to eq. 2.2 requires the current belief state, that can be evaluated by the filtering process, discussed below.

Filtering is the most common inference problem in online analysis,

and it consists in producing an estimate of the belief state using Bayes' rule:

$$p(X_t \mid y_{1:t}) \propto p(y_t \mid X_t, y_{1:t-1})\, p(X_t \mid y_{1:t-1}) \quad (2.4)$$
$$= p(y_t \mid X_t)\, p(X_t \mid y_{1:t-1}) \quad (2.5)$$

where the second line is obtained by applying zero-th order Markov assumption on Y. This equation is sometimes called the *update* equation of Bayesian filtering, and it is the basic equation in many probabilistic problems including Kalman Filtering, Dynamic Belief Networks, probabilistic localization in robotics.
Notice that the filtering equation involves the prediction of the hidden state, which in turn requires filtering; this fact allows to design recursive algorithms for filtering, the most basic of which are the *forwards–backwards* algorithm for HMM's [94] and the *junction–tree* for generic graphical models [12]. Thus recursive estimation for online filtering consists of two main steps: *predict* and *update*. In the following we will not discuss in depth any inference algorithm, but rather try to clarify what is the inference problem that is relevant to the adoption of the IOCHMM as a sensorimotor control mechanism for an artificial agent.

The formulation of the inference problem for standard HMM's can be readily extended to IOHMM's as well as to our IOCHMM.
In the case of the IOHMM, denoting input variables with U, (one time step) prediction amounts to computing $p(X_{t+1} \mid y_{1:t}, u_{1:t+1})$ or $p(Y_{t+1} \mid y_{1:t}, u_{1:t+1})$, while filtering requires evaluation of $p(X_t \mid y_{1:t}, u_{1:t})$. After repeated application of Bayes' rule, total probability rule, and Markovianity, one finds that:

$$p(X_{t+1} \mid u_{1:t+1}, y_{1:t}) = \sum_{x_t} p(X_{t+1} \mid u_{t+1}, x_t)\, p(x_t \mid u_{1:t}, y_{1:t}) \quad (2.6)$$
$$p(Y_{t+1} \mid u_{1:t+1}, y_{1:t}) = \sum_{x_{t+1}} p(Y_{t+1} \mid u_{t+1}, x_{t+1})\, p(x_{t+1} \mid u_{1:t+1}, y_{1:t}) \quad (2.7)$$
$$p(X_t \mid u_{1:t}, y_{1:t}) \propto p(y_t \mid u_t, X_t)\, p(X_t \mid u_{1:t}, y_{1:t-1})\,. \quad (2.8)$$

So far, we can write down the corresponding formulas for prediction of hidden and output states in the IOCHMM presented in section 2.3. The explicit computation, which involves simply repeated application of Bayes' rule under Markov assumption, while using the conditional independencies expressed by the graph structure, leads to:

$$p(\bar{X}_{t+1} \mid \bar{u}_{1:t+1}, \bar{y}_{1:t}) = \sum_{x_t^e} \sum_{x_t^h} p(X_{t+1}^e, X_{t+1}^h \mid u_{t+1}^e, u_{t+1}^h, x_t^e, x_t^h)$$

$$\cdot \quad p(x^e_t, x^h_t \mid u^e_{1:t}, u^h_{1:t}, y^e_{1:t}, y^h_{1:t}) \tag{2.9}$$

$$p(\bar{Y}_{t+1} \mid \bar{u}_{1:t+1}, \bar{y}_{1:t}) = \sum_{x^e_{t+1}} \sum_{x^h_{t+1}} p(Y^e_{t+1} \mid u^e_{t+1}, x^e_{t+1})\, p(Y^h_{t+1} \mid u^h_{t+1}, x^h_{t+1})$$

$$\cdot \quad p(x^e_{t+1}, x^h_{t+1} \mid u^e_{1:t+1}, u^h_{1:t+1}, y^e_{1:t}, y^h_{1:t}) \tag{2.10}$$

$$p(\bar{X}_t \mid \bar{u}_{1:t}, \bar{y}_{1:t}) \propto p(\bar{y}_t \mid \bar{u}_t, \bar{X}_t)\, p(X^e_t, X^h_t \mid u^e_{1:t}, u^h_{1:t}, y^e_{1:t-1}, y^h_{1:t-1}) \tag{2.11}$$

where, in the l.h.s., $\bar{X}_{t+1}$ stands for X^e_{t+1}, X^h_{t+1}, and similarly for the other variables.

Notice that if the hidden variables are taken to be discrete, then the entity expressed by equation 2.9 is a matrix. If in addition the output probability distributions are Gaussian, then equation 2.10 corresponds to a Mixture of Gaussians, with mixing weights expressed by equation 2.9.

2.4.2 Learning

The problem of learning the parameters associated with the DBN described above is the following: each node has a conditional probability distribution, which describes the probability of taking each particular value *given* the values of all its parent nodes. Such probability distributions depend on some parameters, that are unknown, and should therefore be estimated on the basis of some examples (the *training data*).
There exist several techniques to solve this problem, and two main classes can be identified: classical (or *frequentist*) approaches are aimed just at evaluating the parameters, while a fully Bayesian approach aims at learning a probability distribution on the parameters, which allows the model to express uncertainty on the actual parameters, due for example to a small training set.
In the work presented here, and more specifically in chapter 4, we will follow the classical approach. In the case of partial observability (due to the presence of hidden nodes), classical learning is based on the *Maximum Likelihood Estimate* (MLE) of the parameters (i.e. finding the values of the parameters that maximize the likelihood of training data), which in turn can be achieved using *Gradient Ascent* techniques, or the standard *Expectation Maximization* algorithm (EM [23]) for exact inference.
Since learning relies on inference, several ad hoc techniques have been

developed for the case of approximate inference, but we will not discuss this issue further.
We will focus instead on learning in standard HMM's under exact inference, which is usually done by a suitable variant of EM, namely the *Baum Welch* [8] algorithm; here we briefly recall how it works, and then extend the results to the IOHMM and the IOCHMM.

MLE learning consists in finding the values for the parameters θ that maximize the *log–likelihood* of the training data Y (for simplicity of notation, we consider just one sequence in the training set):

$$\theta^\star = \arg\max \, logP(Y|\theta)\,; \tag{2.12}$$

this problem is simplified by considering the hidden variables as *missing data*, introducing a free probability distribution Q. Then the Jensen's inequality, which holds for concave functions, can be applied to the logarithm, and then a lower bound for the log–likelihood can be set as follows:

$$\begin{aligned}
\log(P(Y\,|\,\theta)) &= \log \sum_X Q(X) \frac{P(Y,X\,|\,\theta)}{Q(X)} \\
&\geq \sum_X Q(X) \log(\frac{P(Y,X\,|\,\theta)}{Q(X)}) \\
&= \sum_X Q(X) \log P(Y,X\,|\,\theta) - \sum_X Q(X) \log Q(X) \qquad (2.13) \\
&= F(Q,\theta)\,. \qquad (2.14)
\end{aligned}$$

The quantity $\log P(Y,X\,|\,\theta)$ is usually called the *complete–data log–likelihood*. The negative of the functional F is known in statistical physics as the *Free energy*. The EM algorithm maximizes such Free energy with respect to Q and θ alternatively, starting from a given value θ_0, until convergence is reached (i.e. the log–likelihood does not increase significantly anymore); the k-th step is given by:

$$E - step: \quad Q_{k+1} \leftarrow \arg\max_Q F(Q,\theta_k) \tag{2.15}$$

$$M - step: \quad \theta_{k+1} \leftarrow \arg\max_\theta F(Q_k,\theta)\,. \tag{2.16}$$

After simple calculations, it can be shown that the E–step is obtained by setting $Q_{k+1} = P(X\,|\,Y,\theta_k)$, while the M–step reduces to finding θ_{k+1} that maximizes $\sum_X P(X\,|\,Y,\theta_k) \log P(Y,X\,|\,\theta)$.

The quantity computed in the E–step is obtained in principle by inference on the graph; in practical cases it is usually not necessary to compute it in its entirety, rather one only needs to compute the so called *Expected Sufficient Statistics* (ESS), which plays an analogous role to the sufficient statistics in the case of complete data. Recalling that the expected value of a function $f(x)$ under the probability distribution $p(x)$ is defined as:

$$\langle f(x) \rangle_{p(x)} = \int f(x) p(x) dx \,, \tag{2.17}$$

the ESS usually corresponds to the expected values (found by inference) of suitable combinations of the hidden variables, computed under Q_{k+1}. In order to define the ESS, and write the parameters as a function of it, some additional calculations are required. This complete derivation in the case of standard HMM's can be found in many textbooks [12]; we limit ourselves here to show it for the IOHMM, and eventually for our IOCHMM, in the case that all variables are discrete.

In the IOHMM, the EM algorithm requires a straightforward generalization of the above formulas, to account for the fact that the training set includes input sequences as well. Given a single training sequence $\{u_{1:T}, y_{1:T}\}$, and using the graph (in)dependence structure in the IOHMM, the complete–data log–likelihood is rewritten as:

$$\begin{aligned} L_C \doteq \log p(\bar{y}_{1:T}, \bar{X}_{1:T} \,|\, \bar{u}_{1:T}, \theta) &= \log p(x_1 \,|\, u_1, \theta) \\ &+ \sum_{t=1}^{T-1} \log p(x_{t+1} \,|\, u_{t+1}, x_t, \theta) \\ &+ \sum_{t=1}^{T-1} \log p(y_{t+1} \,|\, x_{t+1}, \theta) \,. \end{aligned} \tag{2.18}$$

We will specialize now to the case where all variables are discrete, therefore the associated distributions are matrices (usually called Conditional Probability Tables, CPT), whose entries are the parameters that must be learnt.
In the discrete case, computations are simplified by encoding the states of each variable as unit vectors in the canonical basis of the associated state space; e.g. if $x_t \in \{x^1, \ldots, x^H\}$, then $x^1 = (1, 0, \ldots, 0)$ and so forth. This approach allows us to rewrite the probability distributions in eq. 2.18 as follows:

$$p(x_1 \,|\, u_1) = \Pi_{i=1}^{I} \Pi_{j=1}^{H} (\Phi_{ij})^{u_{1,i} x_{1,j}} \tag{2.19}$$

$$p(x_{t+1} \mid u_{t+1}, x_t) = \Pi_{i=1}^{I}\Pi_{j=1}^{H}\Pi_{k=1}^{H}(T_{ijk})^{u_{t+1,i}x_{t,j}x_{t+1,k}} \quad (2.20)$$

$$p(y_{t+1} \mid x_{t+1}) = \Pi_{i=1}^{H}\Pi_{j=1}^{O}(E_{ij})^{x_{t+1,i}y_{t+1,j}} \quad (2.21)$$

where I, H, O denote the cardinality of the input, hidden and output state spaces, Φ, T, E are respectively the *initial state*, *transition* and *emission* probability matrices that we want to learn.
Taking the logarithms and writing the summation in matrix form, we get the following form for the complete–data log–likelihood:

$$\begin{aligned} L_C \doteq \log p(\bar{y}_{1:T}, \bar{X}_{1:T} \mid \bar{u}_{1:T}) &= x_1^{\perp}\log(\Phi)u_1 \\ &+ \sum_{t=1}^{T-1} x_{t+1}^{\perp}\log(T)x_t u_{t+1} \\ &+ \sum_{t=1}^{T-1} y_{t+1}^{\perp}\log(E)x_{t+1} \end{aligned} \quad (2.22)$$

which is written in terms of the training data (u, y), the parameters (Φ, T, E), and the *missing data* (x). Now the E–step consists in evaluating the ESS, namely

$$\gamma_{t,i} \doteq \langle X_{t,i}\rangle = p(X_t = x^i \mid u_{1:T}, y_{1,T}) \quad (2.23)$$

$$\xi_{t,ij} \doteq \langle X_{t,i}, X_{t-1,j}^{\perp}\rangle = p(X_t = x^i, X_{t-1} = x^j \mid u_{1:T}, y_{1,T}) \,, \quad (2.24)$$

given the previous values of the parameters. Then the M–step is straightforward: we take the derivatives of eq. 2.22 with respect to the parameters, set to zero and solve under the constraint that transition, emission and initial state probabilities sum to one. This gives the following form for the parameters, similar to that for standard HMM's:

$$\Phi_{ij} = \gamma_{1,i}u_{1,j} \quad (2.25)$$

$$T_{ijk} = \frac{\sum_{t=2}^{T}\xi_{t,ij}u_{t,k}}{\sum_{t=2}^{T}\gamma_{t,j}u_{t,k}} \quad (2.26)$$

$$E_{ij} = \frac{\sum_{t=2}^{T} y_{t,i}\gamma_{t,j}}{\sum_{t=2}^{T}\gamma_{t,j}} \,. \quad (2.27)$$

Eventually, γ and ξ are found via the standard forward–backward inference algorithm [12], which is an instance of belief propagation [137] applied to HMM's.

The derivation described above is readily extended to the IOCHMM

case[1]. Again the log–likelihood can be expressed in terms of the complete data log–likelihood, called here L_c, which in this case is

$$\begin{aligned} L_c &= \log p(x_1^e|u_1^e,u_1^h) + \log p(x_1^h|u_1^e,u_1^h,x_1^e) \\ &+ \sum_{t=1}^{T-1} \log p(x_{t+1}^e|u_{t+1}^e,u_{t+1}^h,x_t^e,x_t^h) + \sum_{t=1}^{T-1} \log p(x_{t+1}^h|u_{t+1}^e,u_{t+1}^h,x_{t+1}^e,x_t^h) \\ &+ \sum_{t=0}^{T-1} \log p(y_{t+1}^e|x_{t+1}^e) + \sum_{t=0}^{T-1} \log p(y_{t+1}^h|x_{t+1}^h)\,. \end{aligned} \tag{2.28}$$

Define with M, N, L, K the dimensionality of the hand and eye movement hidden and input space respectively. Again we encode discrete variables in the canonical basis, e.g. if $x^e \in \{x^{e,1} \dots x^{e,M}\}$, then we have $x^{e,1} = (1, 0 \dots 0)$ and so on [12]. With this choice, the pdf's in the log–likelihood become (notice that we will not discuss further the emission distributions, as they are exactly the same as in the IOHMM case.):

$$p(x_1^e|u_1^e,u_1^h) = \prod_{i=1}^{M}\prod_{j=1}^{L}\prod_{p=1}^{K}(\Phi_{ijp}^e)^{x_{1,i}^e u_{1,j}^e u_{1,p}^h} \tag{2.29}$$

$$p(x_1^h|u_1^e,u_1^h,x_1^e) = \prod_{i=1}^{N}\prod_{j=1}^{L}\prod_{p=1}^{K}\prod_{r=1}^{M}(\Phi_{ijpr}^h)^{x_{1,i}^h u_{1,j}^e u_{1,p}^h x_{1,r}^e} \tag{2.30}$$

$$p(x_{t+1}^e|u_{t+1}^e,u_{t+1}^h,x_t^e,x_t^h) = \prod_{i=1}^{M}\prod_{j=1}^{L}\prod_{p=1}^{K}\prod_{r=1}^{M}\prod_{s=1}^{N}(T_{ijprs}^e)^{x_{t+1,i}^e u_{t+1,j}^e u_{t+1,p}^h x_{t,r}^e x_{t,s}^h} \tag{2.31}$$

$$p(x_{t+1}^h|u_{t+1}^e,u_{t+1}^h,x_{t+1}^e,x_t^h) = \prod_{i=1}^{N}\prod_{j=1}^{L}\prod_{p=1}^{K}\prod_{r=1}^{M}\prod_{s=1}^{N}(T_{ijprs}^h)^{x_{t+1,i}^e u_{t+1,j}^e u_{t+1,p}^h x_{t+1,r}^e x_{t,s}^h} \tag{2.32}$$

where, again, Φ, T are respectively the *initial state* and *transition* probability matrices that we want to learn. The log–likelihood (apart from the emission terms, as announced above) can be finally written in matrix form as:

$$\begin{aligned} L_c &= x_1^{e\perp}\log(\Phi^e)u_1^e u_1^h + x_1^{h\perp}\log(\Phi^h)u_1^e u_h^h x_1^e \\ &+ x_{t+1}^{e\perp}\log(T^e)u_{t+1}^e u_{t+1}^h x_t^e x_t^h + x_{t+1}^{h\perp}\log(T^h)u_{t+1}^e u_{t+1}^h x_{t+1}^e x_t^h\,. \end{aligned} \tag{2.33}$$

[1]Notice however that some care should be exerted in the forward–backwards algorithm, since in the case of coupled HMM's the forward operator cannot be decoupled, as explained in [97].

Again, the M–step is straightforward: we take the derivatives of eq. 2.33 with respect to the parameters, set to zero and solve under the constraint that transition and initial state probabilities sum to one. After defining the following quantities that make part of the ESS:

$$\gamma^e_{t,i} \doteq \langle X^e_{t,i} \rangle \tag{2.34}$$
$$\gamma^h_{t,i} \doteq \langle X^h_{t,i} \rangle \tag{2.35}$$
$$\gamma^{eh}_{t,i} \doteq \langle X^e_{t,i}, X^h_{t,i} \rangle \tag{2.36}$$
$$\xi^{e,h}_{t,ij} \doteq \langle X^e_{t,i}, X^h_{t-1,j} \rangle \tag{2.37}$$
$$\xi^{e,eh}_{t,ij} \doteq \langle X^e_{t,i}, X^e_{t-1,j}, X^h_{t-1,j} \rangle \tag{2.38}$$
$$\xi^{eh,h}_{t,ij} \doteq \langle X^e_{t,i}, X^h_{t,i}, X^h_{t-1,j} \rangle \tag{2.39}$$

we can rewrite the parameters as:

$$\Phi^e_{ijk} = \gamma^e_{1,i} u^e_{1,j} u^h_{1,k} \tag{2.40}$$
$$\Phi^h_{ijkl} = \gamma^{eh}_{1,il} u^e_{1,j} u^h_{1,k} \tag{2.41}$$
$$T^e_{ijklm} = \frac{\sum_{t=2}^{T} \xi^{e,eh}_{t,ilm} u^e_{t,j} u^h_{t,k}}{\sum_{t=2}^{T} \gamma^{eh}_{t,lm} u^e_{t,j} u^h_{t,k}} \tag{2.42}$$
$$T^h_{ijklm} = \frac{\sum_{t=2}^{T} \xi^{eh,h}_{t,ilm} u^e_{t,j} u^h_{t,k}}{\sum_{t=2}^{T} \xi^{e,h}_{t,lm} u^e_{t,j} u^h_{t,k}} \,. \tag{2.43}$$

Eventually, the γ and ξ terms are found via the forward–backward inference algorithm. This is done by first introducing the following quantities:

$$\alpha^{eh}_t \doteq p(x^e_t, x^h_t, \bar{y}_{1:t} \,|\, \bar{u}_{1:T}) \tag{2.44}$$
$$\beta^{eh}_t \doteq p(\bar{y}_{t+1:T} \,|\, x^e_t, x^h_t, \bar{u}_{1:T}) \tag{2.45}$$

usually called *forward* and *backward* operators; then after simple calculations γ and ξ can be expressed in terms of α^{eh}, β^{eh}:

$$\begin{cases} \gamma^e_{t,i} = \frac{\alpha^e_{t,i}\beta^e_{t,i}}{p(\bar{y}_{1:T} \,|\, \bar{u}_{1:T})} \\ \gamma^h_{t,i} = \frac{\alpha^h_{t,i}\beta^h_{t,i}}{p(\bar{y}_{1:T} \,|\, \bar{u}_{1:T})} \\ \gamma^{eh}_{t,ij} = \frac{\alpha^{eh}_{t,ij}\beta^{eh}_{t,ij}}{p(\bar{y}_{1:T} \,|\, \bar{u}_{1:T})} \\ \xi^{e,h}_{t,ij} = \frac{p(\bar{y}_t \,|\, x^e_{t,i}, \bar{u}_{1:T})\, \alpha^h_{t-1,j}\, p(x^e_{t,i} \,|\, x^h_{t-1,j}, \bar{u}_{1:T})\, \beta^e_{t,i}}{p(\bar{y}_{1:T} \,|\, \bar{u}_{1:T})} \\ \xi^{e,eh}_{t,ijk} = \frac{p(\bar{y}_t \,|\, x^e_{t,i}, \bar{u}_{1:T})\, \alpha^{eh}_{t-1,jk}\, p(x^e_{t,i} \,|\, x^e_{t-1,j}, x^h_{t-1,k}, \bar{u}_{1:T})\, \beta^e_{t,i}}{p(\bar{y}_{1:T} \,|\, \bar{u}_{1:T})} \\ \xi^{eh,h}_{t,ijk} = \frac{p(\bar{y}_t \,|\, x^e_{t,i}, x^h_{t,j})\, \alpha^h_{t-1,k}\, p(x^e_{t,i}, x^h_{t,j} \,|\, x^h_{t-1,k}, \bar{u}_{1:T})\, \beta^{eh}_{t,ij}}{p(\bar{y}_{1:T} \,|\, \bar{u}_{1:T})} \end{cases} \tag{2.46}$$

where α^e is obtained by marginalizing α^{eh} with respect to h, and similarly for the other quantities. Eventually, α^{eh}, β^{eh} and the normalization term are evaluated recursively:

$$\alpha_t^{eh} = p(\bar{y}_t \mid x_t^e, x_t^h) \sum_{x_{t-1}^e} \sum_{x_{t-1}^h} p(x_t^e, x_t^h \mid x_{t-1}^e, x_{t-1}^h, \bar{u}_t) \alpha_{t-1}^{eh} \quad (2.47)$$

$$\beta_t^{eh} = \sum_{x_{t+1}^e} \sum_{x_{t+1}^h} p(\bar{y}_{t+1} \mid x_{t+1}^e, x_{t+1}^h) p(x_{t+1}^e, x_{t+1}^h \mid x_t^e, x_t^h, \bar{u}_{t+1}) \beta_{t-1}^{eh} \quad (2.48)$$

$$p(\bar{y}_{1:T} \mid \bar{u}_{1:T}) = \sum_{x_T^e} \sum_{x_T^h} \alpha_T^{eh} \quad . \quad (2.49)$$

Following the extension of the Baum–Welch algorithm outlined above, the distributions of initial state, transition and emission probabilities can be inferred from a number of suitable *example* sequences that show how the inputs and outputs (observed nodes) are related. We will discuss in chapter 4 the choice of the training set, and the results of the learning stage.

2.4.3 Decision

The next step to solve the problem of movement selection, is to apply a decision process on the results of inference. Here we could follow two different routes, depending if we start from equation 2.9 or 2.10. In the first case, we use an inference algorithm to compute the probability distribution for the next *hidden* state, then apply the decision process to select from this distribution a particular value for the next hidden state, and finally generate the value for the next *output* state by sampling the corresponding output distribution. We assume that this procedure is the one that should be used to generate the agent's actions, because the hidden states model the internal dynamics of the agent itself, while the output distributions account for noisy execution.
On the other hand, the second case corresponds to the predictions that an observer could make on the agent's behavior; in fact the observer does not have access to the agent's hidden states, but can only observe the outputs. Then the observer can resort to equation 2.10 to infer the probability distribution of the next output given previous observations, and then apply a decision process to select the value for the next *output* state.

Notice in passing that although the observer does not have access to the agent's internal dynamics, however it has its own internal dynamics for the same sensorimotor task (in terms of the learned transition and emission probability distributions; these in principle are different from the distributions learned by the agent, since the agent and the observer could have been trained with different data sets). In the evaluation of equation 2.10, the observer is in fact playing what has been called an *embodied simulation* of the agent's actions [38], thus implementing a mirror–like mechanism [72].

Having chosen the first strategy for action selection, we now have to define the decision procedure. According to Bayesian Decision Theory, if the agent is given a set of observation data d and formulates some hypotheses h, a decision rule is a function $\alpha(d)$ that associates the data with an hypothesis. A loss function $L(h, \alpha(d))$ can be defined to quantify the cost of choosing a wrong hypothesis. Then, the agent will take its decision so as to minimize the risk, namely a functional that quantifies the cost of the decision weighted by the joint probability of the data and the hypothesis:

$$r(\alpha) = \sum_{h,d} L(h, \alpha(d)) P(h, d) \quad . \tag{2.50}$$

Different choices can be made for the cost function, and the decision rule that minimizes the risk changes accordingly. In the simulations presented in this thesis we used the most basic cost function:

$$L(h, \alpha(d)) = \begin{cases} 0 & \textit{if } \alpha(d) \neq h \\ 1 & \textit{otherwise} \end{cases} \quad . \tag{2.51}$$

With this choice the risk function is minimized when the decision rule just selects the hypothesis that maximizes the conditional probability $P(h|d)$.
Interestingly enough, the use of Bayesian inference jointly exploited with Decision Theory has gained some currency in recent sensorimotor modelling generalizing to Bayesian accounts of biological agent evolution (e.g., [40], [128])

Coming back to our IOCHMM, we can then select the next eye–hand movement as the couple that maximizes the probability in equation 2.9:

$$(x_{t+1}^{e\star}, x_{t+1}^{h\star}) = \arg\max \left[p(\bar{X}_{t+1} \mid \bar{u}_{1:t+1}, \bar{y}_{1:t}) \right] \quad . \tag{2.52}$$

We move now to analyze a simple special case, that will be considered later in chapter 4 where we present simulation results.

If we assume as a first approximation, that the agent's movement execution is perfect, we have

$$p(y_t^i \mid x_t^i) = \delta_{y_t^i, x_t^i} \tag{2.53}$$

where $i = e, h$. Under this assumption, also the update equation (2.11) reduces to a product of *delta* functions:

$$\begin{aligned} p(\bar{X}_t \mid \bar{u}_{1:t}, \bar{y}_{1:t}) &= \eta\, p(y_t^e \mid x_t^e)\, p(y_t^h \mid x_t^h)\, p(X_t^e, X_t^h \mid u_{1:t}^e, u_{1:t}^h, y_{1:t}^e, y_{1:t}^h) \\ &= \delta_{x_t^e, y_t^e}\, \delta_{x_t^h, y_t^h} \end{aligned} \tag{2.54}$$

and this in turn implies that the prediction equation (2.9) for hidden states reduces to

$$\begin{aligned} p(\bar{X}_{t+1} \mid \bar{u}_{1:t+1}, \bar{y}_{1:t}) &= \sum_{x_t^e} \sum_{x_t^h} p(X_{t+1}^e, X_{t+1}^h \mid u_{t+1}^e, u_{t+1}^h, x_t^e, x_t^h)\, \delta_{y_t^e, x_t^e}\, \delta_{y_t^h, x_t^h} \\ &= p(X_{t+1}^e, X_{t+1}^h \mid u_{t+1}^e, u_{t+1}^h, x_t^e = y_t^e, x_t^h = y_t^h)\,. \end{aligned} \tag{2.55}$$

This quantity is the matrix that expresses the transition probability distribution, in the case that hidden variables at the previous time step are clamped to the previous observed values.
Thus, in this simplified case, once the network has been trained, it is no more necessary to resort to inference algorithms, since no marginal distribution needs to be computed before applying the decision process. In other words, in this case we just train the network, store the transition distribution (in the discrete case, a matrix called Conditional Probability Table), and then select the next eye–hand movement according to:

$$(x_{t+1}^{e\star}, x_{t+1}^{h\star}) = \arg\max \left[p(X_{t+1}^e, X_{t+1}^h \mid u_{t+1}^e, u_{t+1}^h, x_t^e = y_t^e, x_t^h = y_t^h) \right]. \tag{2.56}$$

Finally, notice also that in this case the prediction equation 2.10 for observations takes exactly the same form as equation 2.55. This means that in this case, the action selected by the agent coincides with the action predicted by an observer, since in the case of perfect execution the observer gets full knowledge of the agent's hidden states just by observing the outputs.

Chapter 3

The Drawing Task. Eye–Tracking Experiments

3.1 Introduction

Visual creation is a specifically human activity, with a long history and multiple uses. From the perspective of cognitive sciences, the process of carrying out a visual creation can be seen as a goal-directed activity involving several human skills and abilities: visuomotor coordination, evaluation and decision, memory and emotion.

Here we focus our attention on realistic drawing; our motivation for studying this task is that the behavior adopted in this case can be considered as a building block of the visual creative behavior in a broader sense, yet it allows to concentrate the analysis on the physical aspects of the creative process.
Realistic drawing is not considered a 'common' visuomanual activity like driving [61], washing one's hands [86], or making a sandwich [6], neither is it considered a 'common' visual task such as the recognition of a face or a specific object in the scene; in fact, drawing requires a better precision of hand movements and a higher degree of voluntary attentional control of fixations. Making a realistic portrait of a visual scene is a very specific task, and it imposes rigid constraints on eye–hand coordination.

Among the few existing scientific studies on the drawing process, the earliest were focused on the motor component. Recordings of the arm movements during curve tracing revealed a correlation between the

curvature of the trajectory and the speed profile of the drawing hand; this was expressed in mathematical form by the empirical $\frac{2}{3}$ *power law* [60, 133], which states that the absolute value of the hand velocity is proportional to the $\frac{2}{3}$ power of the inverse curvature radius.
Later on, cortical recordings in behaving *rhesus monkey* proved the correctness of the power law at the neural level, and found in addition that populations of neurons in the motor cortex encode the direction of hand movement with about 100 msec of anticipation [106].
Only recently however, the coordination of eye and hand movements was brought to the focus of the attention by a series of studies realized by C. Miall and colleagues [123, 46]. Interestingly, a visuomotor strategy can be clearly observed on subjects involved with drawing tasks, although the strategy can vary significantly among different subjects. Indeed, the above mentioned eye tracking experiments on draughtsmen at work provide evidence of two nested execution cycles: the longer, external cycle is an oscillation between periods when the hand is not drawing and globally distributed eye movements can be observed, and periods when the hand is *tracing*; within the tracing period a shorter nested cycle can be noticed, with eye movements localized alternately in small parts of the scene and the canvas. The shorter cycle can be schematized as follows: *fixation on the original image – saccade – fixation(s) on the canvas – saccade – fixation on the original image.*

Further analysis [19, 18] indicates that four main subtasks should be distinguished:

1. *Segmentation* of the original scene;
2. *Evaluation* of the emerging result;
3. Feature extraction for *motion planning and generation*;
4. *Visual feedback* for online motion control.

The oscillation between local and global scanpaths may be understood by considering gaze–shifts as the motor realization of overt shifts of attention. Visual attention arises from the activation of those same circuits that process sensory and motor data [110]. In particular, selective attention for spatial locations is related to the dorsal visual stream that has been named *action pathway* after Goodale and Humprey [74], and is mainly devoted to trigger prompt actions in response to environmental varying conditions (*Vision for Action*). On

the contrary, selective attention for objects derives from activation of ventral cortical areas involved in the *perception pathway*, responsible for object recognition, with tight integration to high–level, cognitive tasks of frontal areas (*Vision for Perception*). Clearly, the two pathways are not segregated but cooperate/compete to provide a coherent picture of the world and gaze control is the ultimate product of such integration.
In this framework behaviors 1 and 2, that require globally distributed eye movements, could be associated to the Vision for Perception stream, while 3 and 4 produce localized eye movements related to the Vision for Action stream. Thus, the oscillation can be seen as a part of a high level strategy, which takes advantage of the functional architecture of the human visual system to keep separate two classes of visual behaviors, the first of which is global in nature and perceptual in purpose, while the second is local and pragmatic, sub-serving a precise hand movement.

In the present work we focus mainly on behaviors 1 and 3, namely eye movements related to the segmentation of the image in separate objects, and to the extraction of visual features that are required for hand motion planning and control. We leave instead for future work the analysis of oculomotor behavior associated to the feedback control of hand movements, and to the more general issue of the evaluation of emerging results.

3.2 Three basic hypotheses

As a starting point to characterize the visuomotor strategies adopted in the drawing task, we make some hypotheses that try to capture the essential features that distinguish drawing from other tasks, both with respect to the *a priori* requirements and the observed behavior. Three assumptions can be introduced, in reference to a drawing task described as copying an *original image* on an initially blank *canvas*. [18].

1. *An object in the original image becomes relevant (almost) only during the time that it is being copied. Therefore (almost) all fixations on an object are executed within a time interval in which no fixations occur on other objects.*

2. *Fixations are distributed among the original objects according to the number of salient points on each object, and on each single object following the distribution of most salient points.*

3. *The sequence of fixations on the original scene is constrained to maximize continuity of tracing hand movements.*

The first assumption states that a peculiar feature of the drawing behavior is that the gaze does not move back and forth among different objects, but proceeds sequentially. Gaze is directed to an object only when it becomes relevant to the task, i.e. during the time that it is being copied.
Salient points can be defined as those with local intensity and orientation contrast [51] above a given threshold and the second assumption requires the draughtsman to move the gaze towards all salient points. This implies a segmentation which is finer than the initial object-based segmentation and is directly related to pragmatic sensorimotor control.
Third assumption implies that feedback information on hand motion plays an important role in determining the actual scanpath. One possible implication is that the scanpath on the original scene should resemble a coarse–grained edge following along the contours of the objects, which has never been observed in the eye–tracking literature to the best of our knowledge.

In the following sections we present our eye–tracking experiments, whose aims were to test the correctness of our hypotheses, as well as

their implications for the observable sensorimotor behavior.
In the experimental sessions we recorded eye and hand movements, while the subjects were copying an original image on a blank sheet (the canvas); we call this the *realistic drawing* task.

3.3 Methods

3.3.1 Participants

Two experimental sessions were realized. In the first, preliminary session, eye scan records were obtained from three right–handed individuals, aged between 27 and 33 years; in the second, main session 29 subjects were used, 5 of which were left–handed.
All subjects had normal or corrected to normal vision; none of them had specific previous training in drawing or painting.

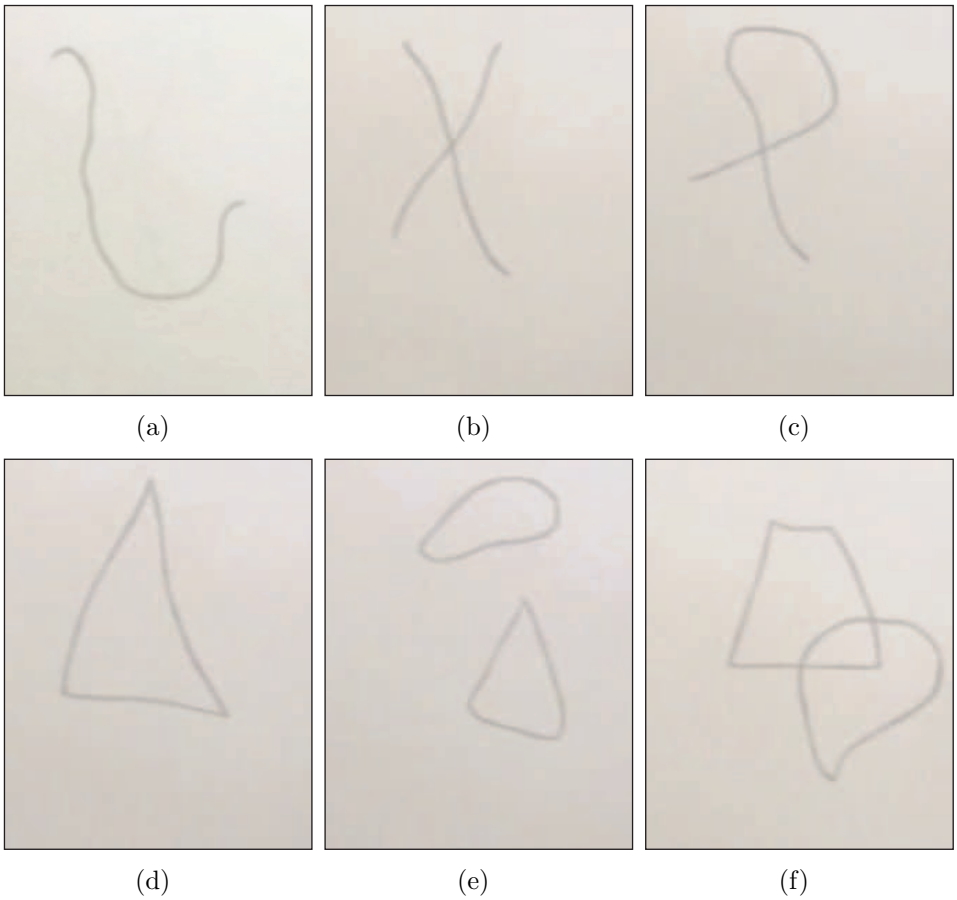


Figure 3.1: The original images adopted in the eye tracking experiments; human subjects were instructed to copy these images on an initially blank canvas.

3.3.2 Displays and Instructions

The experimental setup is shown in Fig. 3.2. Subjects were presented with a rectangular, vertical tablet 40 cm × 30 cm, viewed binocularly from a distance ranging from 35 cm to 45 cm depending on the subject's arm length. In the left half of the tablet hand–drawn images were displayed, while the right half was initially occupied by a white sheet. The original images, shown in Fig. 3.1, represented simple contours drawn by hand with a black pencil on white paper, that occupied an area of approximately 15 cm × 15 cm.
One image per trial was shown, and the subjects were instructed to copy its contours as faithfully as possible, drawing on the right hand sheet. These instructions did not make specific mention of eye movements and did not give constraints on the execution time.
Each subject carried out six trials, one per image, and the images were presented always in the same order.

3.3.3 Eye movement recording

The subject's left eye movements were recorded with a remote eye tracker (ASL 5000) with the aid of a magnetic head tracker (Ascension *Flock of Birds*), with the eye position sampled at the rate of 60 Hz[1]. The instrument can integrate eye and head data in real time and can deliver a record with an accuracy of less than 1 deg in optimal light conditions.
The spatial configuration of the experiment is shown schematically in Fig. 3.2. Due to the wide field of view (about 30 deg × 23 deg), and the fact that the eye camera was on the bottom left margin of the field of view, only the records corresponding to the left hemifield showed a good accuracy; thus, information on eye position in the right hemifield, namely fixations on the drawing hand, were discarded[2].

The eye camera focuses on the left eye of the subject; it is an active camera that automatically pans and tilts, in response to head movements reported by the magnetic sensor, in order to keep the image centered on the subject's eye. The image thus captured is processed by custom hardware, in order to extract the center of the pupil and of

[1] All experimental facilities were kindly provided by the *Natural Computation Lab* of the University of Salerno.

[2] This problem will be solved in the future by adopting a *head–mounted* eye tracker.

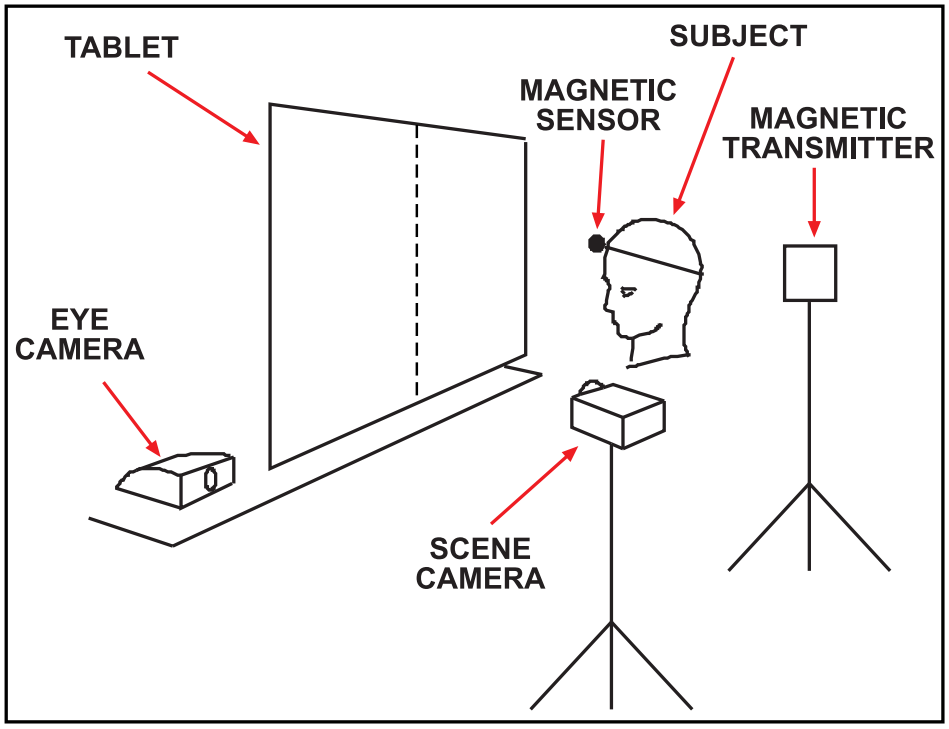


Figure 3.2: The Subject sits in front of a rectangular, vertical Tablet. In the left half of the Tablet hand–drawn images are displayed, while the right half is initially occupied by a white sheet where the Subject is instructed to copy the images. The eye tracker integrates data from the Eye Camera and the Magnetic Sensor and Transmitter; eye position is then superimposed on the Scene Camera video stream, which takes the approximate subjective point of view.

the corneal reflection; these pieces of information are then combined, via simple geometrical calculations, to obtain the 3D angle of gaze direction.
Once this is done, a calibration procedure must be followed for each subject, in order to obtain an estimate of the projection of the gaze vector onto the image plane (the tablet). This amounts to estimating a function that maps 3D gaze angle values to 2D cartesian coordinates.

3.3.4 Preliminary analysis of recordings

The first qualitative analysis was conducted on the video output provided by the instrument: this is the video taken from the point of view of the subject using the Scene Camera (defined in Fig. 3.2), with the eye position displayed as a black cursor superimposed on each frame (see Fig. 3.4).

Then the raw eye tracker data could be displayed in the form of a plot of horizontal and vertical eye positions against time, as shown in figures 3.5(a) and 3.5(b).
The fixations were detected by a Matlab implementation [44] of the standard dispersion algorithm; the dispersion threshold was set to 2.0 deg with a minimum fixation duration of 100 msec. The algorithm detects the fixations, and outputs fixation position ($x-y$ coordinates) and duration. Mainly the spatial coordinates of individual fixations were used in our analysis, while duration was used only to evaluate the total fixation time on relevant areas.

The raw data and the fixations could then be displayed in an X-Y plot superimposed to the digitized version of the stimulus image for the given trial, as shown respectively in figures 3.6(a), and 3.6(b).

Figure 3.3: A picture of the experimental setup at the *Natural Computation Lab* of the University of Salerno. The inset in the upper left corner shows the control monitor for the eye camera.

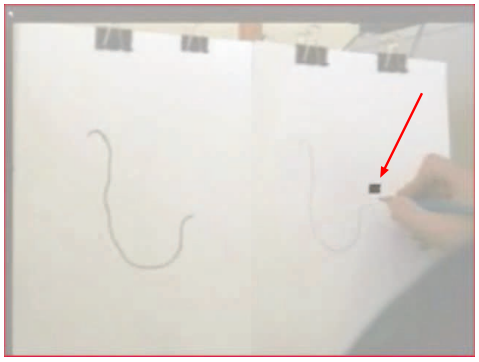
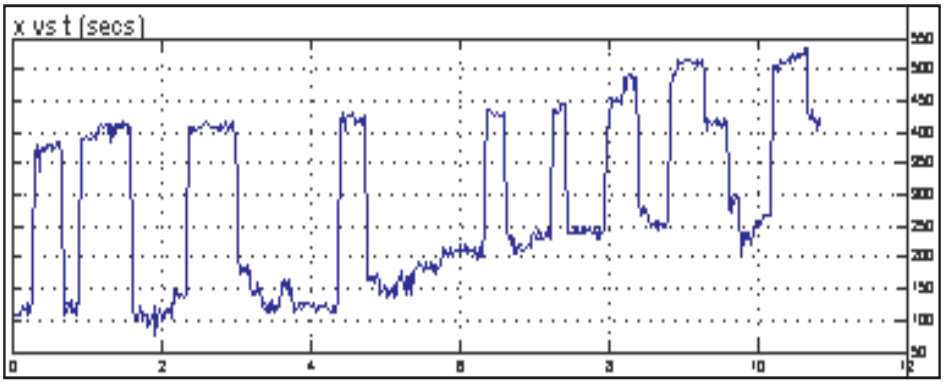
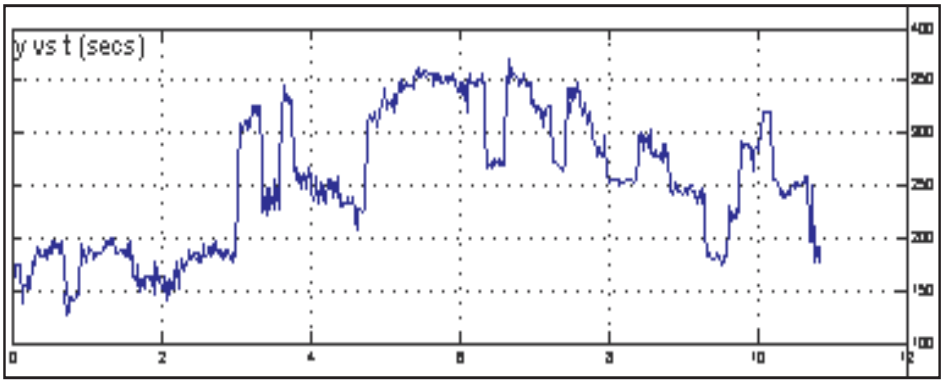


Figure 3.4: Subjective image from the Scene Camera, with the gazepoint shown by the black cursor (the red arrow was added manually).

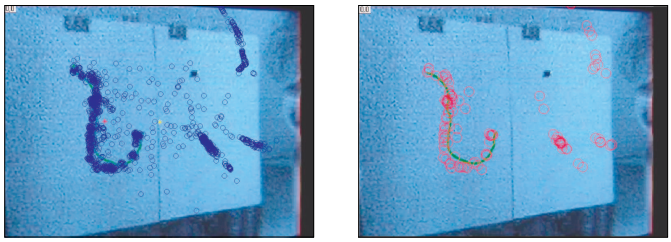


(a)



(b)

Figure 3.5: Raw eye data for one subject in trial 1: 3.5(a) X and 3.5(b) Y *vs* time.



(a) (b)

Figure 3.6: X – Y plot of 3.6(a) raw data and 3.6(b) fixations.

3.4 Analysis of recordings

At present, only very few eye tracking studies on drawing humans have been conducted [123, 46], and no standard measures have been defined for this task (see [52, 25] for a survey of common measures used in eye tracking research). Therefore our analysis of eye data was driven mainly by the hypotheses we made in section 3.2; the aim was to test the correctness of such hypotheses, as well as their implications as to the characterization of the sensorimotor behavior of drawing humans.

3.4.1 Hypothesis 1

In trial 5 the image displayed is composed by two closed contours that are *spatially separated* (see Fig. 3.1(e)). Hypothesis 1 states that in this case the two objects are scanned in two disjoint time intervals,

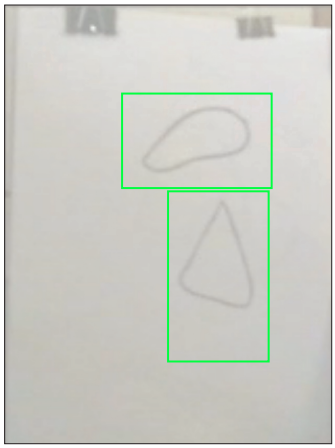
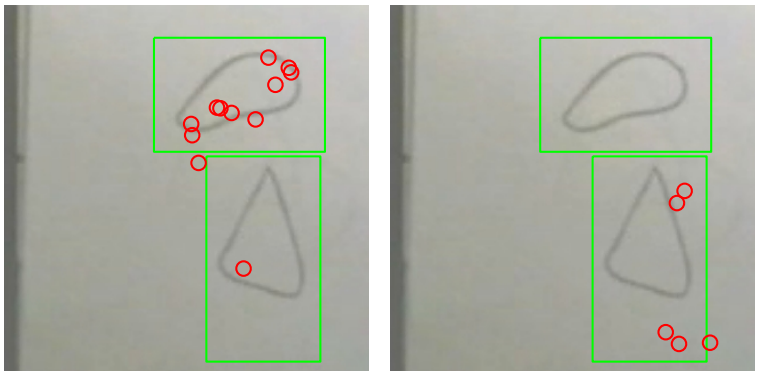


Figure 3.7: Subjective image from the Scene Camera in trial 5, with the ROIs defined by two green rectangles.



(a) (b)

Figure 3.8: The cumulative fixations executed by subject **AP** in trial 5, during each of the two time intervals.

expressing the fact that gaze is directed to an object only when it becomes relevant to the task: in this case, we are assuming that *an object is relevant (almost) only during the time that it is being copied.*

In fact, from qualitative analysis it resulted that all the subjects started drawing the second object only after completion of the first one, independently of which of the two objects was chosen as the first[3]. Thus we first defined, for each subject, two time intervals, T1 and T2, corresponding to the two drawing phases; these were found

[3]Eye data recorded during this trial from 9 subjects were noiseless enough to allow reliable analysis; therefore the results presented in this section refer to those 9 subjects only.

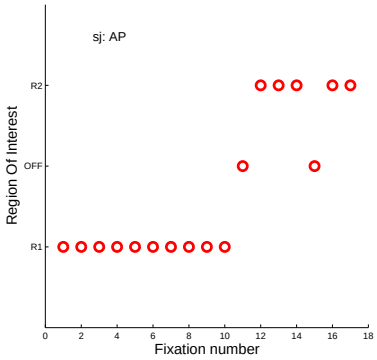
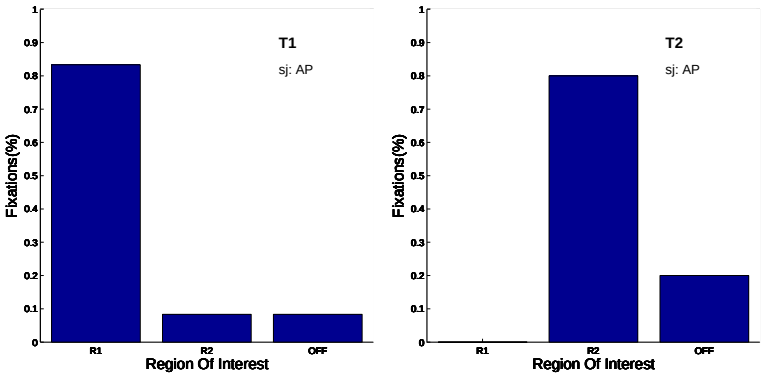


Figure 3.9: Plot of the position (R1, R2, OFF) of each subsequent fixation across the whole execution of trial 5.



(a) (b)

Figure 3.10: The distribution of fixations over the three regions (R1,R2,OFF) executed by subject **AP** in trial 5, during each one of the two time intervals.

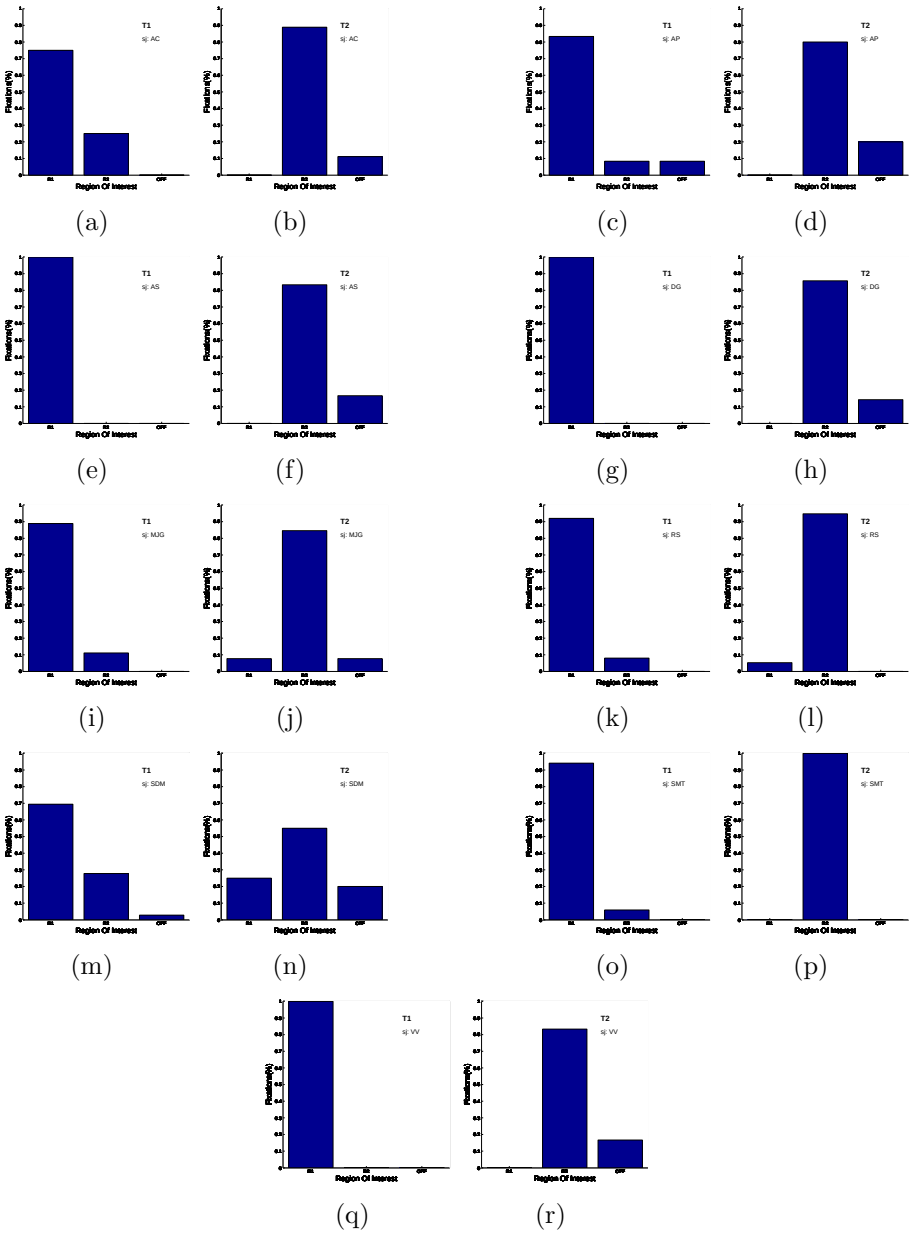
by inspection of the video data. Then we defined two rectangular Regions Of Interest (ROI), R1 and R2, each one containing one of the two objects as shown in Fig. 3.7. Fixations on the left hemifield were then classified in each time interval, as falling in R1, R2 or outside. Fig. 3.8 shows the fixations executed by subject AP in each of the two time intervals, while Fig. 3.9 is a plot of the position (R1, R2, or OFF) of each subsequent fixation. These plots give an idea of the correctness of hypothesis 1 for this single subject, as confirmed by the histogram of the fixations over the three regions in each time interval 3.10.

In Fig. 3.11 we plot, for each subject, the distribution of the number of fixations (F) over the three classes, and in Fig. 3.12 we report the same plot, averaged over the 9 subjects. These plots show good agreement with hypothesis 1, given that

a) the maximum of the distribution is always in the region corresponding to the time interval considered; and, most notably,

b) the percentage of F in the 'wrong' region is always below 27 % for each subject, and below 10 % in average.

It should be noticed that hypothesis 1 does not mention fixations outside the ROI's, while the data show an increase of F in this class when moving from T1 to T2 (see Fig. 3.11 and 3.12 where the average distribution of F is plotted). This fact could be explained by the additional hypothesis that after one object O has been completed then sporadic saccades between O and the next object can be used



(a) (b) (c) (d)

(e) (f) (g) (h)

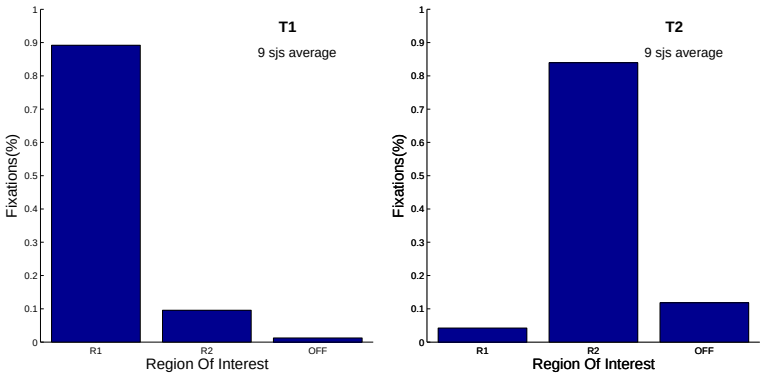
(i) (j) (k) (l)

(m) (n) (o) (p)

(q) (r)

Figure 3.11: Plots of the distribution of the number of fixations over the regions (outside, R1 or R2). Each plot refers to either time interval T1 (left column) or T2 (right column). Each row corresponds to one subject.

to evaluate information, such as the distance and relative size, that are relevant for an accurate drawing. These fixations can be naturally thought of as supporting the evaluation of the emerging result, and were not analyzed further in the present work.



(a) (b)

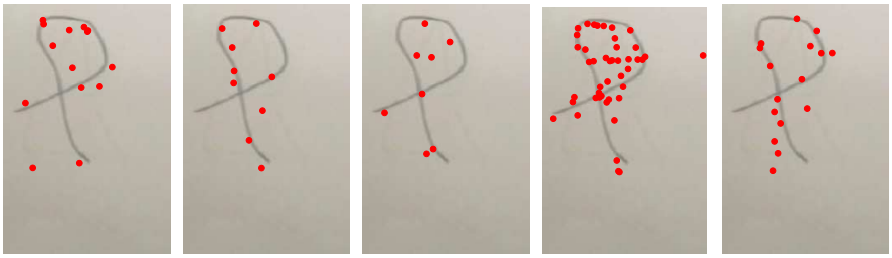
Figure 3.12: The distribution of fixations over the three regions (R1,R2,OFF) during each one of the two time intervals, averaged over 9 subjects.

3.4.2 Hypothesis 2

In trial 4 the image displayed includes an irregular cross and a loop[4] (see figure 3.1(c)). Hypothesis 2 states that at the end of the trial, fixations should be distributed on the image contours according to some *saliency* measure; in other words, we expect to find clusters of fixations around the most salient points.

If we define saliency of a point as proportional to the local intensity and orientation contrast (as proposed e.g. in [51]), then it is evident from qualitative analysis of the cumulative fixation plot for each of 5 subjects (Fig. 3.13) that fixations are distributed preferentially near

[4]The results presented in this section refer to only 5 subjects, as it is intended only to give a proof of principle.

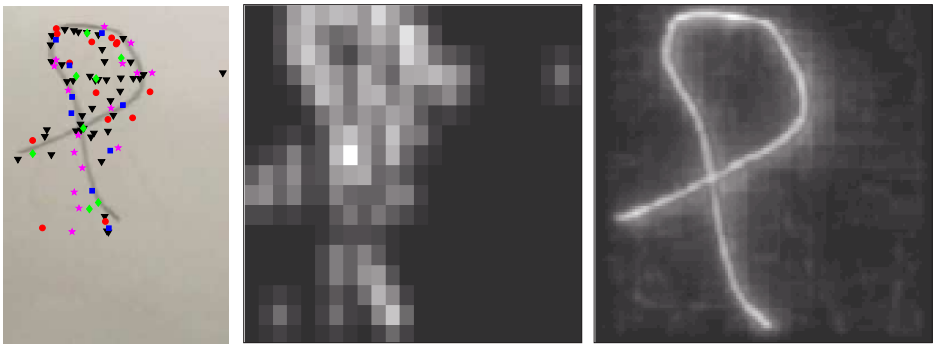


(a) (b) (c) (d) (e)

Figure 3.13: The cumulative plot of fixations in trial 3; from left to right, subjects **AC**, **AP**, **AS**, **RS**, **VV**.

salient regions, as predicted by hypothesis 2.
The general fact that fixations tend to be located near salient points is well known from the literature concerned with purely visual tasks. However, it was not obvious that this should hold also for a visuomanual task such as realistic drawing: in fact in this case *all* the points of the image are *potentially* relevant to the task, since each portion of the image must be copied faithfully.

For a quantitative verification of hypothesis 2, we follow an approach proposed in [135]. We first evaluate the cumulative *fixation map* for the 5 subjects: we divide the image in a fixed number of cells, count the number of fixations per cell by all individuals, assign a gaussian centered on each cell with height proportional to the corresponding value, and eventually normalize the resulting matrix to obtain a 2D probability distribution which can be displayed as a grid of grayscale cells (see figure 3.14(b)).
Then we compare it with the *saliency map* obtained as suggested in [51]. Starting from the original image, early visual features such as color opponents, intensity and orientation are computed in a set of feature maps based on retinal input and represented using pyramids. Then, center-surround operations are implemented as differences between a fine and a coarse scale for a given feature. One feature type encodes for on/off image intensity contrast, two encode for red/green and blue/yellow double-opponent channels and four encode for local orientation contrast. The contrast pyramids for intensity, color, and orientation are summed across scales into three conspicuity maps, which in turn are eventually combined in a saliency map (see Fig. 3.14(c)).
Eventually, the numerical comparison between the two maps requires the definition of a measure of similarity between two matrices. Given that the fixation map represents a 2D probability distribution, also the saliency map can be converted to a probability distribution; then any standard similarity criterion, e.g. the Kullback–Leibler divergence, can be used to evaluate the similarity between the two maps. This result should then be compared with the distance of the experimental fixation map from a map obtained by sampling a uniform distribution over the image contour. We leave this explicit computations for future work.



(a) (b) (c)

Figure 3.14: 3.14(a) The cumulative plot of fixations in trial three for all 5 subjects; 3.14(b) the corresponding *fixation map* [135], and 3.14(c) the *saliency map* [51] of the original image (see text for the explanation of how these maps are obtained).

3.4.3 Hypothesis 3

Our third hypothesis states that *the sequence of fixations on the original scene is constrained to maximize graphical continuity of tracing hand movements.*

In order to explore the correctness and the implications of this hypothesis, we analyze the scanpaths recorded in a trial where the original image is a shape composed by a single line. Fig. 3.15 depicts the cumulative plot of fixations, and the corresponding hand position, at four subsequent stages. The times at which the snapshots have been taken, correspond to the moments during which the following sequence is observed: *hand stops - fixation(s) on the left - saccade - fixation(s) on the right - hand moves*. We interpret the points where the hand stops as keypoints, at which the hand's action needs to be reprogrammed and thus fixations on the original image become necessary.

A qualitative inspection of Fig. 3.15 shows a general tendency of the gaze to move orderly along the image contour, as confirmed by the scanpaths of 11 different subjects[5], plotted in Fig. 3.16; furthermore, all of our subjects used graphically continuous hand strokes. This evidence suggests that the strategy that humans adopt in the drawing task, to facilitate graphical continuity of hand movements, is to

[5]Eye data recorded during this trial from 11 subjects were noiseless enough to allow reliable analysis; therefore the results presented in this section refer to those 11 subjects only.

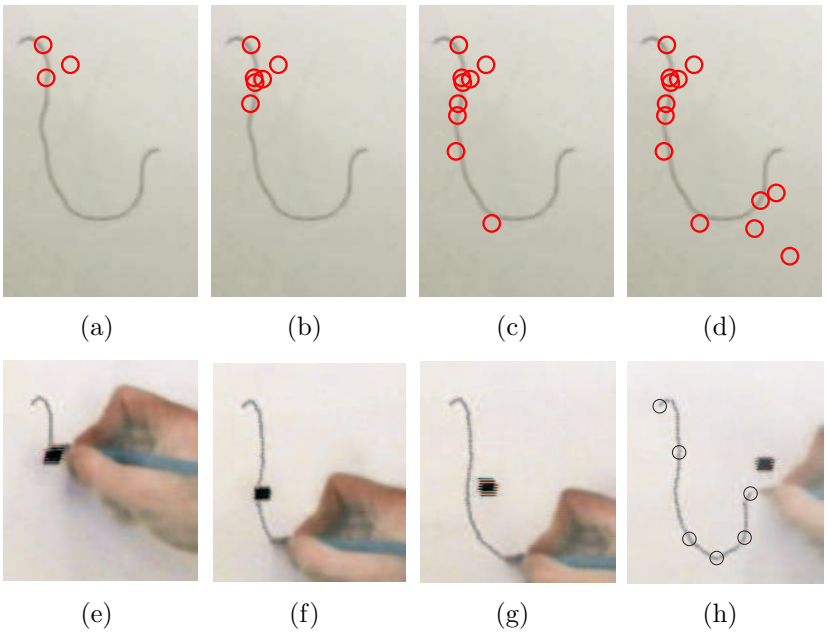


Figure 3.15: The sequence of eye and hand movements by subject **AP** in the drawing task. In the upper row, cumulative fixations on the original image are represented by red circles. In the lower row the solid black square denotes the gazepoint. In 3.15(h) the circles denote the endpoints points of each trajectory segment, found by inspection of the video recording.

move the gaze according to a *coarse grained edge–following* along the contours of the original image.

Thus, we define a procedure (originally proposed in [90]) to evaluate in a quantitative manner the similarity of the recorded scanpaths to the *coarse grained edge–following*; the same procedure can be used then to make a comparison with the scanpaths generated by our as well as other computational models (chapter 4).
As a first step we superimpose an ordered grid on the original image, and then we cluster together all subsequent fixations that fall within a single cell, as one single event. At the end of this procedure, instead of the scanpath we have an ordered sequence of events, each one belonging to a single cell of the grid, as depicted in Fig. 3.17. Then each cell is labeled with a symbol (an ASCII character in the interval 'A' to 'e'), and each sequence of events is converted to a string; this enables a comparison of the strings produced by two algorithms, by two human experimental subjects, or by an algorithm and an experimental subject, using a string matching algorithm. The final value

can be normalized on the basis of the string length.
The string similarity index can be defined through an optimization algorithm, with a cost unit associated to each of three different operations: deleting, inserting and substituting. By sequentially processing the first string to obtain the second string, we get the similarity index as the minimum total cost (this is usually called the Levenshtein distance [66]).

In Fig. 3.18 we report the comparison between the experimental measured scanpath and a) 10000 random strings (i.e. the mean similarity of all the random strings), b) a saliency–based algorithm[51]; c) a perfect edge following and d) the scanpath obtained by the computational model we have proposed, discussed in chapter 4.
For the case a) each random string is formed considering only the pixels where the lines forming input image are present, and their adjacent regions. The probability to extract an empty cell is the half than that of a full one. This fact emphasizes that only occasionally the experimental subjects fixated on white portions of the original image.
The comparison results show that random strings have the lowest string similarity index, meaning that the scanpath in a drawing task can not be considered as a random one. Considering eleven experimental subjects, the average of string similarity index is about 0.098 ± 0.015. Similar results were obtained by the saliency–based scanpath.
Viceversa, better results come from the comparisons with the perfect edge following and the proposed computational model (respectively 0.40 ± 0.15 and 0.39 ± 0.16), thus confirming the validity of our hypothesis 3.

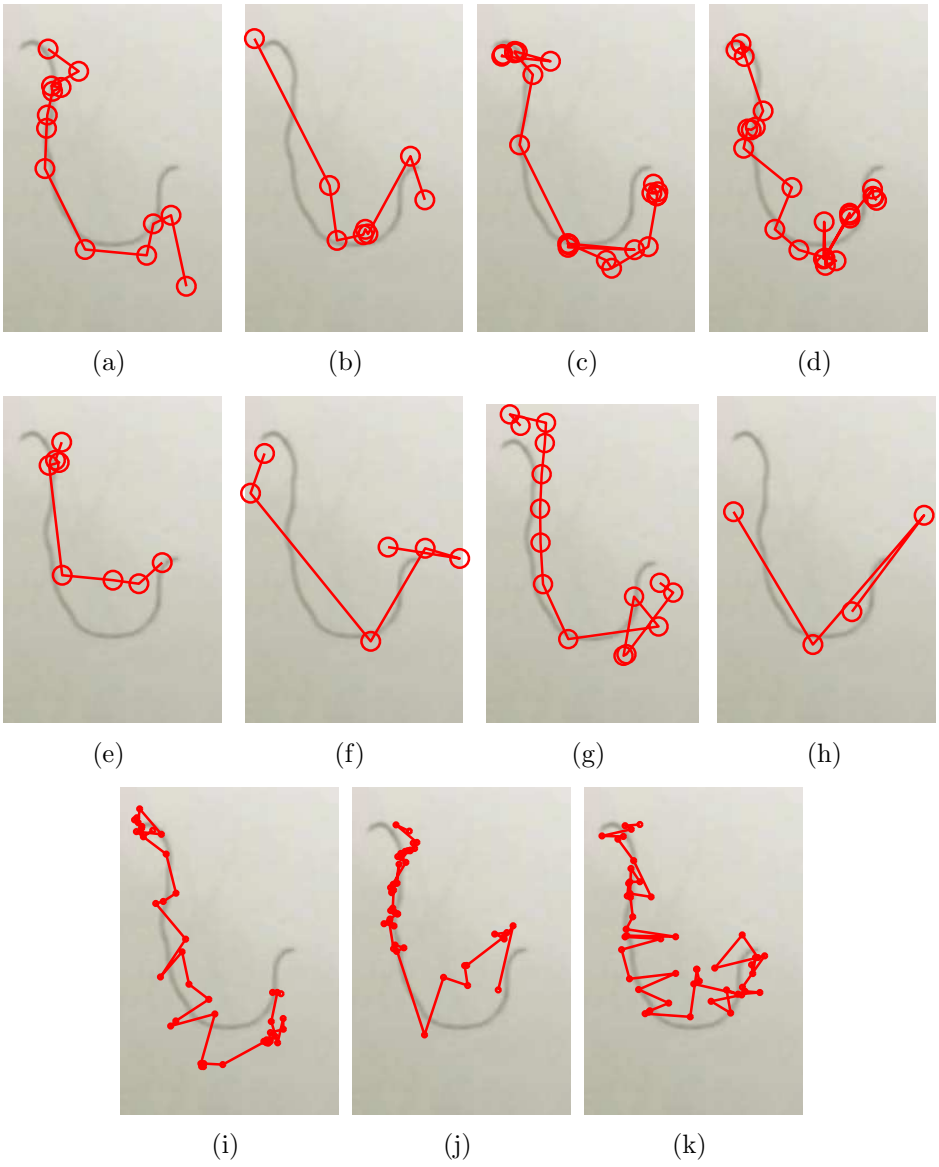
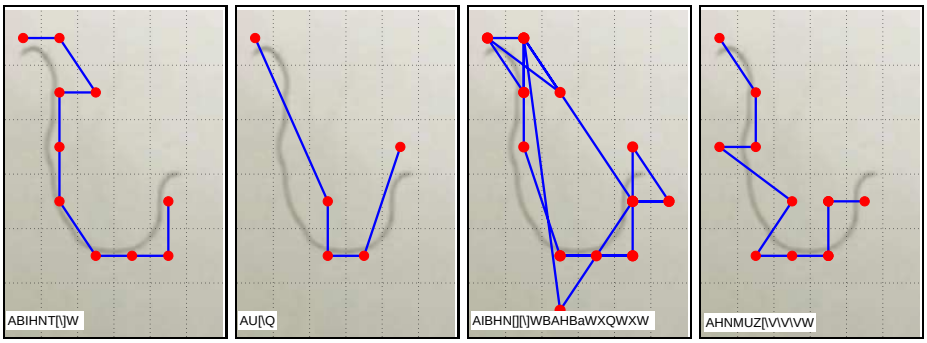


Figure 3.16: From left to right, the scanpath executed by 11 subjects in trial 1.



(a) (b) (c) (d)

Figure 3.17: From left to right, the clustered version of the scanpath, executed by subjects **AP**, **AS**, **AC**, **MJG** in trial 1.

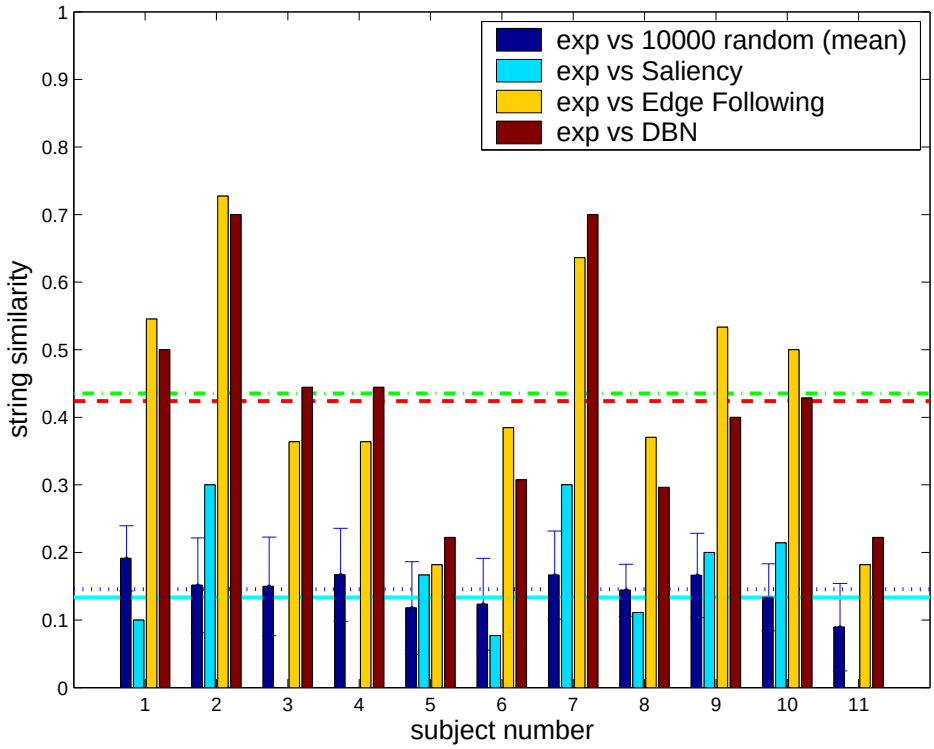


Figure 3.18: The plot shows for each subject (x axis) the mean similarity of the observed scanpath to *i*) 10000 random scanpaths (dark blue with error bar); *ii*) a preattentive scanpath à la Itti (light blue); *iii*) a perfect coarse–grained edge-following (yellow); and *iv*) to the scanpath simulated by our computational model (red, see chapter 4 for a discussion). Horizontal lines denote the respective mean values.

Chapter 4

The Drawing Task. Model Specifications, Simulations and Comparison with Experimental Data

4.1 Introduction

In order to test the performance of the model proposed in chapter 2, and compare the results with human execution, we provided a robotic simulator with the same images presented to human subject in previous eye tracking sessions (chapter 3).
Here we explain the relevant implementation details, mainly those concerning learning, inference and decision in the DBN. Then we outline the implementation of the auxiliary modules, namely those involved in trajectory generation, gazepoint selection, and kinematic inversion. Eventually, we discuss simulation results and give both qualitative and quantitative comparison with experimental data.

4.2 Task description and aims of the simulations

The task we have chosen to test the functional model described in chapter 2 is a realistic drawing task: copying the contour of an irregular shape. The agent is presented with an original image (black on white, 1 pixel stroke width), which is a binarized version of the hand drawn image used in the experiments with human subjects.
Notice that at this stage our model deals only with the fixations on the *original image*, not with fixations on the *drawing hand*; this implies also that feedback on hand movements is purely proprioceptive in this case (we call it the *blind drawing* task).

As discussed in chapter 3, behavioral observations on draughtsmen at work have revealed the existence of a regular execution cycle, where two main phases can be distinguished. During one phase, which corresponds to either the selection of what to draw next or the evaluation of the emerging result, the hand is not drawing, and eye movements distributed over the whole scene can be observed; the other phase is the one during which the hand is tracing and the gaze is moved orderly on small portions of the original image [19].
In order to design an artificial drawing agent, we schematize the above mentioned execution cycle as an oscillation between two main tasks: deciding what to draw next, and actually drawing it. This is described by the following pseudocode:

```
> while(drawing not completed)
>    choose next object;
>    while(object not completed)
>       choose next FOA;
>       draw;
>    end
> end
```

The first task (*choose next object*) must be solved in the cases where the original image is segmented in multiple separated objects: this includes both the case of several disjoint objects, and the case of an object composed by several branches, e.g. a cross. As the experiments presented in chapter 3 show, in this case human subjects almost always direct their fixation to only one object until it has been drawn

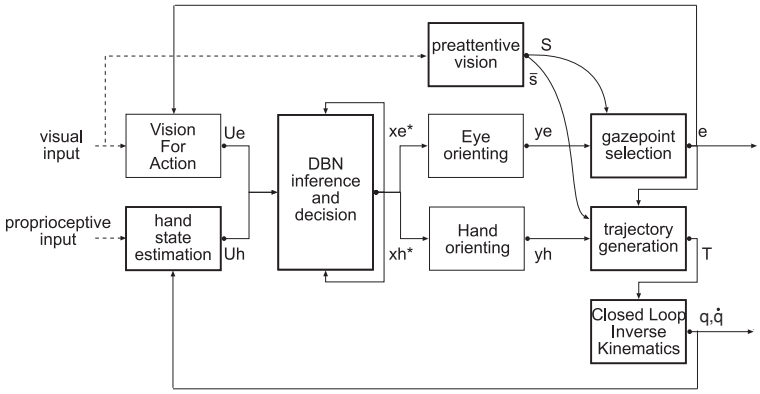


Figure 4.1: A reduced version of the functional model, discussed in chapter 2, for the sensorimotor coordination of a robot involved in a realistic drawing task (see text for explanation)

entirely, and then move to the next one. The solution to this task relies on visual processing mainly related to the Vision for Perception pathway, essentially involving a segmentation of the original scene in separate objects. Many standard techniques exist for this purpose [37].

Once the segmentation is done, the selection process can be expressed in a probabilistic framework. Here we limit ourselves to outlining how this can be done, on the basis of the extended functional model introduced in chapter 2.

Let us assume that a segmentation of the image in separate objets $\{\Theta_1, \ldots, \Theta_K\}$ is available. Let us also assume that each object is further segmented in branches:

$$\Theta_k = \{o_{k,1}, \ldots, o_{k,M}\} \quad . \tag{4.1}$$

If we define n_t^k and $n_t^{k,j}$ as the percentage of the contour that has already been reproduced, at time t, respectively in object k and in branch k, j. These quantities, that are related to the evaluation of task progress, are computed in the *Vision for Perception* modules, and represent the input u^{VP} to the *FEF* module. Then we can denote by

$$p(\Theta_t^k | \Theta_{t-1}^l, n_{1:t-1}^l, n_{1:t-1}^k) \tag{4.2}$$

the a posteriori probability inferred in the *FEF* module. Similarly, for a given object k, the next branch will be distributed according to:

$$p(o_t^{k,j} | o_{t-1}^{k,l}, n_{1:t-1}^{k,l}, n_{1:t-1}^{k,j}) \, . \tag{4.3}$$

The pdf's in eq. 4.2 and eq. 4.3 are then passed on to the *SC* module which in turn will select the next gaze point by maximizing the posterior distribution

$$p(e_t|y_t^e, z_t^e, \Theta_t^k, o_t^{k,l}) \,. \tag{4.4}$$

where y_t^e is the eye–related output of the *Sensorimotor Coupling* module, and z_t^e is the preattentive information encoded by the saliency map.

At the present stage however, our model does not deal with this issue, but rather focuses on controlling the fixations and hand strokes on a single object. Therefore the work presented in the rest of this chapter refers to this task: *reproducing an original image composed by one object*.
For the sake of clarity, we reproduce in Fig. 4.1 a portion of the functional model presented in chapter 2. This corresponds to the modules that we implemented in order to solve the above mentioned task. Recall that the central module, namely *Sensorimotor Coupling*, is reshaped in the form of a DBN, precisely a Input–Output Hidden Markov Model (IOCHMM), depicted in Fig. 4.2. We use the DBN, together with standard Bayesian Decision Theory, as the core module of our functional model. This module generates the sequence of joint eye–hand premotor information that produces the appropriate drawing sensorimotor pattern.

4.3 Joint eye–hand movement planning

4.3.1 Visual and proprioceptive processing

Consider again the system in Fig. 4.1. The visual input is represented by the image of the observed world scene, while the reafferent proprioceptive input is represented by the velocity of the end effector in the drawing plane.

The proprioceptive input is fed into the the *Hand State Estimation* module which, by taking into account internal feedback, computes an estimate of the hand direction (see section 4.3.2 for details); note that here we are not modeling also the speed profile.
The visual input, which more precisely is represented by the scene image together with the point fixated within that image (the fixation point **e** provided as visual feedback, see. Fig. 4.1) follows two routes. In the *Preattentive Vision* module, early visual features (color, intensity, orientation) are extracted through linear filtering across different scales; then, center–surround differences are computed for each feature to yield the feature maps, that are combined in the saliency map $\mathbf{S}$. From such map the agent extracts a list $\bar{\sigma}$ of the n ($n \simeq 6, 7$) most salient points in $\mathbf{S}$, by means of a winner–take–all network; such points represent bottom-up, plausible FOA candidates that can bias higher level gaze planning (see [51] for details).

In the *Vision for Action* module, action–related information is

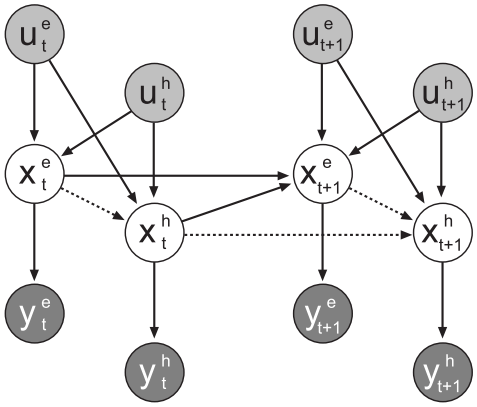


Figure 4.2: The IOCHMM's for combined eye and hand movements. The gray circles denote the input ($\boldsymbol{u}$) and output ($\boldsymbol{y}$) variables. Continuous connections in the hidden layer denote the core process relating hand movements to previous eye movements, while dotted connections highlight the subgraph that represent the complementary process.

computed, within the image region surrounding the fixated point **e** (such region represents the the Focus of Attention, FOA) to provide subsequent modules orientation and curvature information. More precisely, due to the peculiar characteristics of realistic drawing [19], a regular grid is ideally superimposed on the original image and two matrices (N, O) are obtained by assigning to each cell respectively an on/off intensity value (Fig. 4.8(d)) and the average orientation of the contour (Fig. 4.8(e)). Eventually, the visual feature u^e, forwarded to the DBN, is coded as an angular value corresponding to the orientation value of the image contour in the currently fixated cell (see section 4.3.2 for details).

4.3.2 State spaces

In the implementation of the DBN module, discrete state spaces are used for all the variables; we consider a regular grid superimposed on the original image, and eye and hand movements selected by the DBN module are relative to grid cells. In particular, both eye and hand movement models are coded as displacement vectors, originating from the current fixation point or hand position respectively; this choice (encoding the direction rather than the target endpoint) is motivated by literature on neurophysiology, as this appears to be a plausible encoding in the motor and oculomotor areas of primate's brain (see e.g. [106, 55]).

The nodes are distributed on three layers, namely input, hidden and output. The corresponding random variables have the following interpretation and state spaces (also depicted in Fig. 4.3):
$\mathbf{u^e}$: the first input for eye and hand movement planning processes provides the features extracted from the portion of the original image corresponding to the current fixation; in particular u^e is the orientation of the image patch, varying only between 0 and π because lines in the image are not directed:

$$u^e \in \left\{0, \frac{\pi}{8}, \ldots, \frac{7\pi}{8}\right\} \tag{4.5}$$

$\mathbf{u^h}$: the second input for eye and hand movement planning processes concerns information regarding the perceived current position of the hand, as resulting from the elaboration of proprioceptive data. More precisely, u^h is the vector representing the endpoint of the previous movement, with respect to the center of the currently fixated cell.

Since we use only discrete variables, such vector is classified as one out of eight possible directions:

$$u^h \in \left\{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\right\} \tag{4.6}$$

$\mathbf{x^e}$: the state of the eye movement process. Eye movements are coded as displacement vectors, originating from the previous fixation point, as this is the most plausible encoding in biological systems (see e.g. [55]); we consider the length of the displacement as fixed to one grid cell, while the direction can take eight discrete values:

$$x^e \in \left\{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\right\} \tag{4.7}$$

$\mathbf{y^e}$: the eye–movement output, encoding the performed displacement. In principle continuous values should be used for y^e, to include the possibility of errors in the execution; in our implementation however we assume perfect execution, and we have $y^e_t = x^e_t$ at any time step (see section 4.3.4 for a discussion of this assumption).

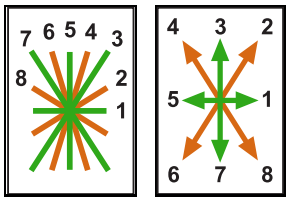
$$y^e \in \left\{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\right\} \quad . \tag{4.8}$$

$\mathbf{x^h}$: the state of the hand–movement process denotes the planned overall direction of the hand movement, within the currently fixated cell. As a slightly different interpretation, x^h should be seen as a class label, where the process of selecting a hand plan is equivalent to the process of selecting an action class; such process then activates a lower–level motor controller that computes the details of the hand kinematics. Formally, x^h is a variable that takes values among eight possible directions:

$$x^h \in \left\{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\right\} \tag{4.9}$$

$\mathbf{y^h}$: the output of the hand–movement process represents the plan that is actually issued and passed forward to a motor controller. The values taken by y^h are the same as for x^h:

$$y^h \in \left\{0, \frac{\pi}{4}, \ldots, \frac{7\pi}{4}\right\} \tag{4.10}$$



(a) (b)

Figure 4.3: A visualization of the state space for the visual input 4.3(a) and eye and hand variables 4.3(b).

4.3.3 Learning the DBN

The problem of learning the parameters associated with the DBN described above is the following: each node has a conditional probability distribution, which describes the probability of taking on each particular value *given* the values of all its parent nodes. In our specific case all the variables are discrete, therefore the associated distribution is in fact a matrix (a Conditional Probability Table — CPT), and each entry in such matrix is treated as a parameter that must be learnt. As discussed extensively in chapter 2, the learning technique we adopt is the Maximum Likelihood Estimate (MLE) of the parameters, using a version of the Baum–Welch algorithm that is adapted to our network; this in turn is a HMM–specific variant of the Expectation Maximization algorithm for exact inference [12].

The system is provided with some *example* sequences that show how the inputs and outputs (observed nodes) are related; from these examples the CPT's of the initial state and the hidden, unobserved layer (*transition* probabilities) are inferred. As a simplifying assumption, here we consider the output mechanism to be perfect, thus the *output* probability distribution is modeled as a delta function, and it needs not be learned:

$$p(y_t^i \mid x_t^i) = \delta_{y_t^i, x_t^i} \tag{4.11}$$

where $i = e, h$.

The examples we use are sequences that reflect the experimental observations on eye–tracked human subjects discussed at length in chapter 2 (see also [18]): hand movements are graphically continuous and correspondingly the scanpath is a coarse–grained edge–following along the contours of the original image.

Since the network is first order Markov, i.e. each node at time t

depends only on nodes at times t and $t-1$, we assume that satisfactory learning results can be obtained using examples that are only two temporal steps long. A possible sequence is the following:

	$t=1$	$t=2$
u_t^e	0	$\frac{\pi}{8}$
u_t^h	π	π
y_t^e	0	$\frac{\pi}{2}$
y_t^h	0	$\frac{\pi}{4}$

(4.12)

In this example sequence the visual input changes across time, it is horizontal in the first time step ($u_1^e = 0$) and then slightly diagonal ($u_2^e = \frac{\pi}{8}$); at the beginning the hand is located at the left of the fixated cell ($u_1^h = \pi$). In the first time step both eye and hand move to the right, but then the input configuration changes, and when $t = 2$ the hand plan is directed upwards along the $45deg$ diagonal, while the eye movement points straight upwards ($y_2^e = \frac{\pi}{2}$, $y_2^h = \frac{\pi}{4}$). Fig. 4.4 depicts the visual input and the eye–hand output corresponding to this specific sequence. The learned joint probability distribution for the eye–hand plans $p(x_{t+1}^e, x_{t+1}^h | u_{t+1}^e, u_{t+1}^h, x_t^e, x_t^h)$ *is* the sensorimotor mapping that characterizes the drawing behavior of the agent; Fig. 4.5 depicts two instances of such map.

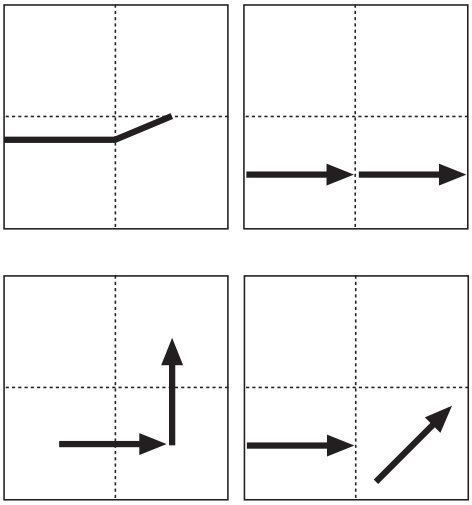
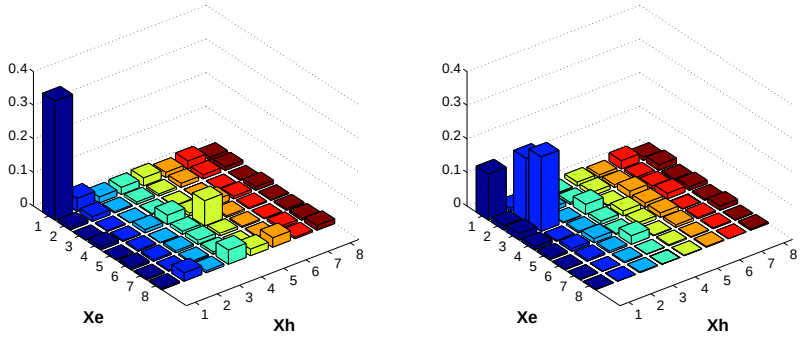


Figure 4.4: The visual input (upper left), proprioceptive input (upper right), and the eye (bottom left) and hand (bottom right) outputs corresponding to the example sequence given in table (4.12).



(a) (b)

Figure 4.5: The joint conditional probability at the hidden nodes (x^e, x^h) at $t = 1$ 4.5(a) and $t = 2$ 4.5(b), in the case that the visual input is the same as in table (4.12).

4.3.4 Decision stage

In order to use the DBN as a control system, we need not just sample it, but rather pass the information it contains to a decision process. As explained in chapter 2, we resort to standard Bayesian Decision Theory, and specifically to the MAP rule; with this choice, the outputs of the network are selected as those that maximize the posterior probability given the inputs and the past history.
In formulae, this amounts to

$$(x^{e\star}_{t+1}, x^{h\star}_{t+1}) = \arg\max \left[p(\bar{X}_{t+1} \mid \bar{u}_{1:t+1}, \bar{y}_{1:t}) \right], \tag{4.13}$$

where $\bar{u} = (u^e, u^h)$ denote the pair of variables representing the visual and hand proprioceptive inputs, respectively, and similarly for $\bar{X}, \bar{y}$. Although in general this requires online inference, as already anticipated in chapter 2 we have run the first simulations under the simplifying assumption that the system has perfect outputs (eq. 4.11). Under this assumption, also the update equation reduces to a product of *delta* functions:

$$\begin{aligned} p(\bar{X}_t \mid \bar{u}_{1:t}, \bar{y}_{1:t}) &= \eta\, p(y^e_t \mid x^e_t)\, p(y^h_t \mid x^h_t)\, p(X^e_t, X^h_t \mid u^e_{1:t}, u^h_{1:t}, y^e_{1:t}, y^h_{1:t}) \\ &= \delta_{x^e_t, y^e_t}\, \delta_{x^h_t, y^h_t} \end{aligned} \tag{4.14}$$

and this in turn implies that the prediction equation for hidden states reduces to

$$p(\bar{X}_{t+1} \mid \bar{u}_{1:t+1}, \bar{y}_{1:t}) = \sum_{x^e_t} \sum_{x^h_t} p(X^e_{t+1}, X^h_{t+1} \mid u^e_{t+1}, u^h_{t+1}, x^e_t, x^h_t)\, \delta_{y^e_t, x^e_t}\, \delta_{y^h_t, x^h_t}$$

$$= \quad p(X^e_{t+1}, X^h_{t+1} \,|\, u^e_{t+1}, u^h_{t+1}, x^e_t = y^e_t, x^h_t = y^h_t)\,. \quad (4.15)$$

This quantity is a matrix that expresses the transition probability distribution, in the case that hidden variables at the previous time step are clamped to the previous observed values.
Thus, in this simplified case, once the network has been trained, it is no more necessary to resort to inference algorithms, since no marginal distribution needs to be computed before applying the decision process. In other words, in this case we just train the network, store the transition distribution (in the discrete case, a matrix called Conditional Probability Table), and then select the next eye–hand movement according to:

$$(x^{e\star}_{t+1}, x^{h\star}_{t+1}) = \arg\max \left[p(X^e_{t+1}, X^h_{t+1} \,|\, u^e_{t+1}, u^h_{t+1}, x^e_t = y^e_t, x^h_t = y^h_t)\right]. \quad (4.16)$$

Notice that, after the DBN has been trained with a sufficient number of examples, it is straightforward to compute at any time step t, given the inputs (u^e_t, u^h_t), and knowing the previous hidden states as $x^i_t = y^i_t$, the posterior hidden state distribution $p(X^e_{t+1}, X^h_{t+1}|u^e_{t+1}, u^h_{t+1}, x^e_t, x^h_t)$. In section 4.3.5 we will introduce an additional mechanism to prevent the eye from moving towards empty cells.

As a result of the learning stage followed by the decision step, we obtain a *sensorimotor map* that encodes the eye and hand directions x^e_t and x^h_t for each given input couple. In Fig. 4.6 we show an instance of such a map in the case of $y^h_{t-1} = x^h_{t-1} = 0$: red and blue arrows denote the direction of the eye and hand plan respectively, for each input couple. In addition, the level of confidence is shown as gray level filling, for the specific eye–hand plan chosen; this is given by the maximum value of the joint conditional probability $p(x^e_{t+1}, x^h_{t+1}|u^e_{t+1}, u^h_{t+1}, x^e_t, x^h_t)$ for any input couple (u^e_{t+1}, u^h_{t+1}). Notice that, due to the choice $x^h_t = 0$, corresponding to a horizontal, rightwards previous hand plan, the confidence is higher in the cases where also the hand input u^h corresponds to a rightwards previous hand plan, be it horizontal or diagonal; this is due to a small number of training examples with $x^h_t = 0$ and different values of u^h . In other words, when the information on the previous hand plan is inconsistent with the actual previous hand plan, the decision taken has a lower level of confidence.

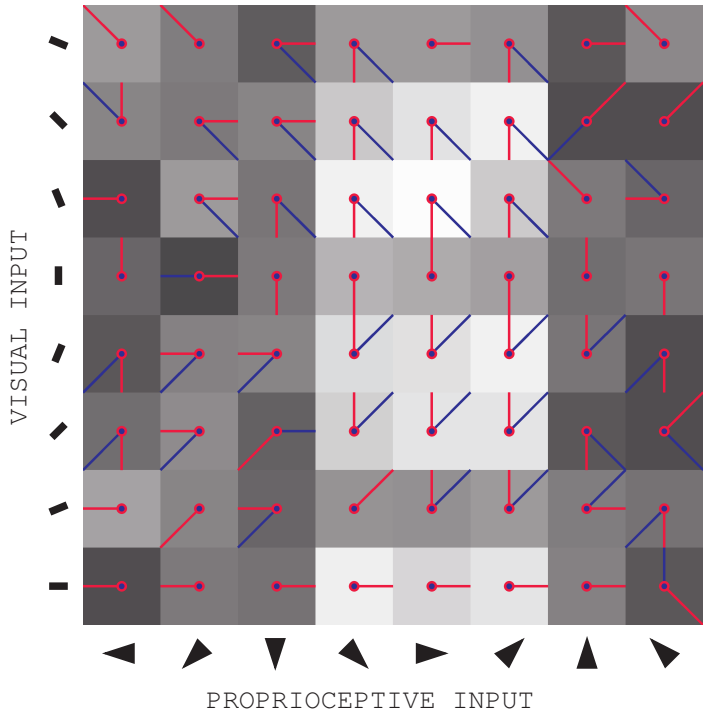


Figure 4.6: The eye–hand policy obtained applying Bayesian Decision Theory to the DBN, in the specific case that $x^e_{t-1} = 0$: red and blue arrows denote the direction of the eye and hand plan respectively, for each input couple. The level of confidence is shown as gray–scale filling for the eye–hand plan chosen, for each input couple. The lighter the pixels, the higher the confidence.

4.3.5 Gazepoint selection and hand trajectory generation

Planning of hand trajectory is achieved by fusing the outputs of different sensorimotor modules of our architecture, in the *Trajectory Generator* module (see Fig. 4.1); here the goal is to reproduce the trajectory planning strategy that can be inferred from the observation of human draughtsmen. Eye tracking experiments have shown that the most common drawing behavior is the following: a) subjects fixate on a location on the original image, b) then move the gazepoint towards the pencil tip, c) draw the corresponding portion of the image and d) stop drawing and go back to point a). Such a cyclic behavior has been discussed in chapter 3. Accordingly, in our model hand trajectory is generated and executed in segments, and the endpoints and intermediate key points of each segment are defined by the fixation points.

Recall that, at any given time step t, the *Gaze orienting* and *Hand orienting* modules provide the next eye and hand movement directions, (y_t^e) and (y_t^h), respectively; meanwhile a set of most salient points $\bar{\sigma}$ is made available by the *Preattentive Vision* module. The latter points are translated to the coordinates of the hand workspace, and are used as the starting and ending points for each trajectory segment.

Gaze points $\mathbf{e}$ are determined by the *Gazepoint Selection* module as follows. Suppose the current gaze location $\mathbf{e}_t \in \bar{\sigma}$; given $\mathbf{e}_t$ and the value of y_t^e, the cell where the gaze point will move next (ϵ_{t+1}) is computed. Then, the next gaze location $\mathbf{e}_{t+1}$ is obtained by finding the most salient point in the image patch $I(\epsilon_{t+1})$ corresponding to the next cell.

The gaze point $\mathbf{e}_t$ and the angular value ϕ_t of the chosen hand direction y_t^h are fed into the *Trajectory Generator* module. This is repeated until it happens again that $\mathbf{e}_{t+\tau} \in \bar{\sigma}$; in this case the sequence of pairs $[(\mathbf{e}_{t+i}, \phi_{t+i})]_{i=0,1,\ldots\tau}$ is interpolated by a spline, setting the slope of the curve at point $\mathbf{e}_{t+i}$ to the value $\tan(\phi_{t+i})$. The resulting curve is the trajectory segment that is fed into the *Inverse Kinematic* module for generating actual motor commands.

4.4 Inverse Kinematics

Movements of a redundant seven degree–of–freedom (DOF) robot manipulator, having a human–like kinematic structure (Fig. 4.7), have been simulated[1].
The drawing task considered here leads to a solution to the inverse kinematics that can be possibly evaluated and compared with arm movements of human experimenters. Previous work in this direction is discussed in [17].

The closed-loop inverse kinematics (CLIK) scheme [109] has been used to obtain the joint variables of the robot manipulator from a differential mapping between task–space and joint–space values, denoted respectively as $\boldsymbol{p}$ and $\boldsymbol{q}$. In solving the kinematic inversion one should keep in mind that in this peculiar case, i.e. the drawing task, only the first two components of the position vector $\boldsymbol{p} = [\, x \quad y \quad z \,]^{\mathrm{T}}$ in the task space are variable, while the z component remains constant during the task execution.
To compute the inverse kinematics we resort to the differential kinematics equation:

$$\dot{\boldsymbol{p}} = \boldsymbol{J}(\boldsymbol{q})\dot{\boldsymbol{q}} \tag{4.17}$$

where $\boldsymbol{J}(\boldsymbol{q})$ is the (3×7) Jacobian matrix. This equation represent the mapping of the (7×1) velocity vector $\dot{\boldsymbol{q}}$ of the joint variables into the task space (3×1) velocity vector $\dot{\boldsymbol{p}}$. It is possible to invert the

[1]This part of the model has been developed in collaboration with Dr. A. De Santis at PRISMA lab, Università di Napoli Federico II.

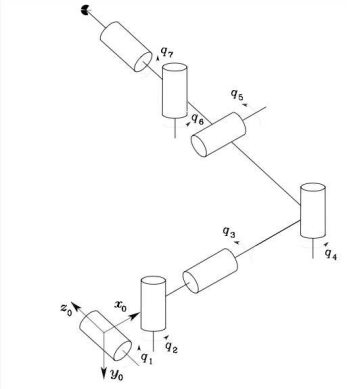


Figure 4.7: The 7 DOF manipulator

equation using the pseudo–inverse of the Jacobian matrix as follows:

$$\dot{\boldsymbol{q}} = \boldsymbol{J}^{\dagger}(\boldsymbol{q})\dot{\boldsymbol{p}} \tag{4.18}$$

where $\boldsymbol{J}^{\dagger} = \boldsymbol{J}^{\mathrm{T}}(\boldsymbol{J}\boldsymbol{J}^{\mathrm{T}})^{-1}$ is a (7×3) matrix; it corresponds to the minimization of the joint velocities in a least-squares sense [109].

In order to contemplate the different characteristics of the available DOF's it could be necessary to modify the velocity distribution with respect to the least-square minimal solution. A possible solution is to consider a weighted pseudo-inverse matrix:

$$\boldsymbol{J}_{W}^{\dagger} = \boldsymbol{W}^{-1}\boldsymbol{J}^{\mathrm{T}}(\boldsymbol{J}\boldsymbol{W}^{-1}\boldsymbol{J}^{\mathrm{T}})^{-1} \tag{4.19}$$

with $\boldsymbol{W}^{-1} = \mathrm{diag}\{\beta_1, \ldots, \beta_7\}$, where β_i is a weighting factor belonging to the interval $[0, 1]$ such that $\beta_i = 1$ corresponds to full motion for the i-th degree of mobility and $\beta_i = 0$ corresponds to freeze the corresponding joint.

Furthermore, redundancy of the robotic arm can be exploited to satisfy secondary tasks, without affecting the primary task, i.e. the motion of the drawing point $\boldsymbol{p}$. To this end, a task priority strategy [78] is used, which leads to the following solution:

$$\dot{\boldsymbol{q}} = \boldsymbol{J}_{W}^{\dagger}(\boldsymbol{q})\dot{\boldsymbol{p}} + \left(\boldsymbol{I}_7 - \boldsymbol{J}_{W}^{\dagger}(\boldsymbol{q})\boldsymbol{J}(\boldsymbol{q})\right)\dot{\boldsymbol{q}}_a \tag{4.20}$$

where $\boldsymbol{I}_7$ is the (7×7) identity matrix, $\dot{\boldsymbol{q}}_a$ is an arbitrary joint velocity vector and the operator $\left(\boldsymbol{I}_7 - \boldsymbol{J}_{W}^{\dagger}\boldsymbol{J}\right)$ projects the joint velocity vector in the null space of the Jacobian matrix.
Discrete–time integration of the joint space velocity can lead to numerical drifts; the CLIK algorithm [109] used here, allows the system to overcome this problem by exploiting the direct kinematics equation to compute an internal feedback signal from the efferent copy of the joint space variables.

The drawing task is performed on a vertical plane. Consequently, the secondary task of minimizing the gravity torques can be transformed to the joint space. This constraint provides an arm posture that is attached to the body. A possible definition of multiple secondary tasks related to the positioning of intermediate parts of the same kinematic structure, including proper trajectory planning, is presented in a more systematic fashion in [102].

4.5 Results

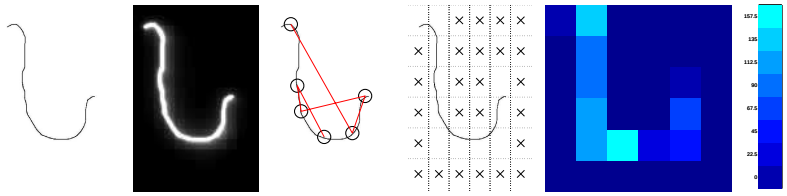
4.5.1 Simulations and qualitative comparison with experimental data

After training the DBN as described above, we have run it on a binarized version of the original image shown to the experimental subjects (Fig. 4.8(a)). The preliminary visual processing, corresponding to the *Vision for Action* and *Preattentive Vision* modules, are shown in Fig. 4.8.

A single run of simulation leads to the time sequences of eye and hand plans $\bar{y}^e, \bar{y}^h$ shown in the two top rows of Fig. 4.9; the bottom row is the sequence of visual inputs, namely the orientation of the image in the region foveated at each time step. The second bottom row shows the confidence level assigned to the eye–hand plan chosen.

The corresponding scanpath is depicted in Fig. 4.10(a), and it can be directly compared to the human eye movement recordings shown in Fig. 4.11. Fig. 4.10(b) shows, in green, the trajectories planned according to the DBN outputs, with the endpoints evidenced by blue circles; these trajectories are computed as splines passing through the points corresponding to the position of each eye fixation, with a slope defined by the associated hand plans.

It is worth remarking that a pure bottom-up, uncoupled scanpath generation would provide a very different result. This can be easily seen, for instance, by feeding the salient points to a winner–takes–all network combined with the inhibition of return[51] in order to ob-



(a) (b) (c) (d) (e) (f)

Figure 4.8: The original image (4.8(a)), the saliency map (4.8(b)), and the most salient locations (4.8(c)) denoted by black circles. The red lines denote the scanpath that would be obtained following the approach proposed in [51]. 4.8(d) shows the imaginary grid superimposed on the image; cells containing an 'X' sign are those evaluated as empty. 4.8(e) depicts the orientation of the image patch contained in each non empty cell; the color code for orientations is explained in 4.8(f).

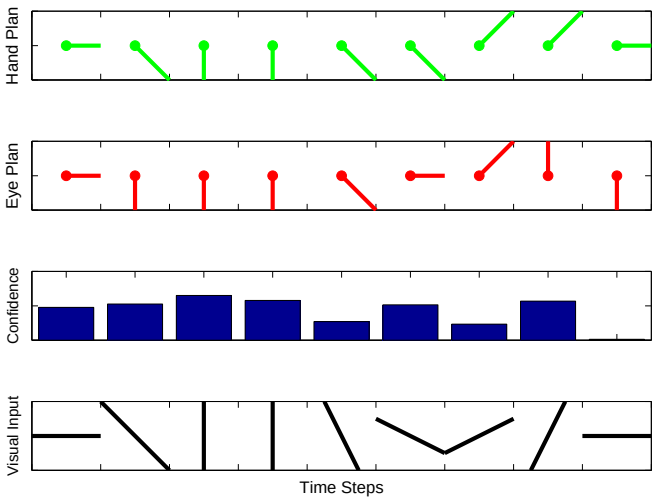
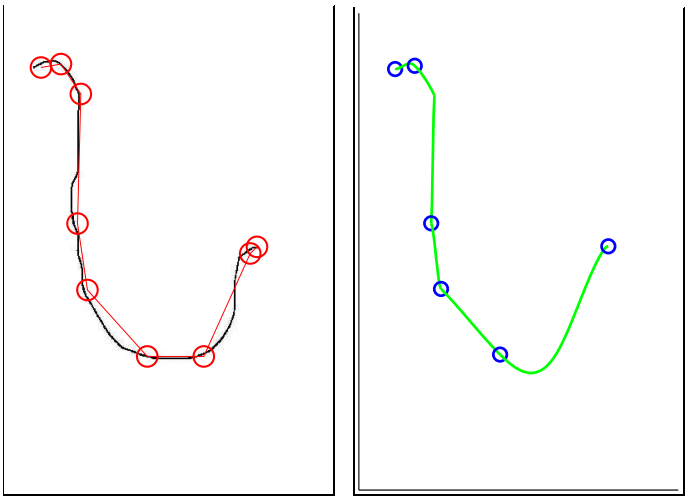


Figure 4.9: The discrete–time evolution obtained as described in section 4.3.5, with time increasing left to right. The bottom row is the sequence of visual inputs, namely the orientation of the image in the region foveated at each time step. The second bottom row shows the confidence level assigned to the eye–hand plan chosen. The two top rows depict respectively the sequences of eye movement plans, in green, and hand movement plans, in red, output by the DBN.

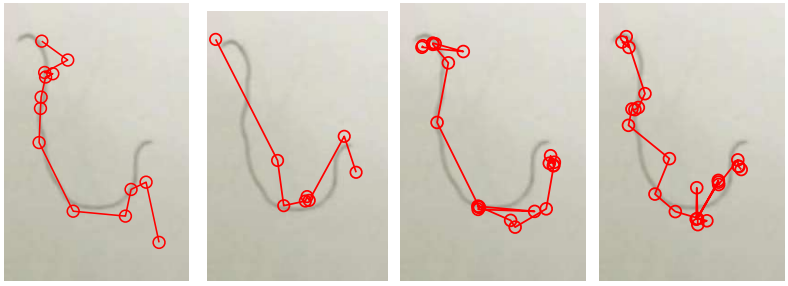
tain the bottom-up fixation sequence; an evaluation of how different this scanpath is from scanpaths either generated by our approach or recorded via eye-tracking, is presented in section 4.5.2.

Fig. 4.10(b) shows, in green, the trajectories planned after the DBN outputs, with the endpoints evidenced by blue circles. For human subjects, such endpoints have been found by inspection of the video recording, as the points where the hand interrupts for a while the drawing movement. In Fig. 4.12 it is possible to observe the temporal sequence of drawing movements by the same subject whose scanpath is in Fig. 4.11(a). The results of the kinematic inversion of such trajectories are shown in Fig. 4.13, where the time histories of the first four joints of the robot are depicted. Finally, Fig. 4.14 shows the pencil trajectory obtained by the simulated robotic arm, with blue circles denoting the endpoints of each trajectory segment. It can be recognized that the simulated trajectory and segmentation points are qualitatively following those recorded experimentally.



(a) (b)

Figure 4.10: The final scanpath (4.10(a)) and planned hand trajectory (4.10(b)); the blue circles in (4.10(b)) denote the starting and ending points of each trajectory portion. Both eye and hand movements start from the upper left corner.



(a) (b) (c) (d)

Figure 4.11: The scanpath executed by 4 human subjects in the drawing task, trial 1, are reproduced here for a visual comparison with the simulated scanpath in Fig. 4.10(a).

4.5.2 Quantitative comparison with experimental data

For a quantitative comparison of both the mobility distribution and the final trajectory, direct measurements of the pencil tip, wrist and elbow would be required. However, the results obtained allow us to present a numerical comparison of the simulated and recorded eye movements.

As explained in chapter 3, in order to make such a comparison, we

first convert the scanpaths to strings of ASCII symbols, and then we can evaluate string similarity as the Levenstein distance between any two strings.

In Fig. 4.15 is reported the similarity between the experimental measured scanpath and a) 10000 random strings (i.e. the mean similarity of all the random strings), b) a saliency–based algorithm [51]; c) a perfect edge following and d) the proposed DBN algorithm.
For the case a) the random string is formed considering only the cells where the lines forming input image are present, and their adjacent cells. The probability to extract an empty cell is the half of that of a full one. This fact emphasizes that only occasionally the experimental subjects fixated on white portions of the original image.

The comparison results show that random strings have the lowest string similarity index, meaning that the scanpath in a drawing task can not be considered as a random one. Considering eleven experimental subjects, the average of string similarity index is about 0.098 ± 0.015. Similar results were obtained by the saliency–based

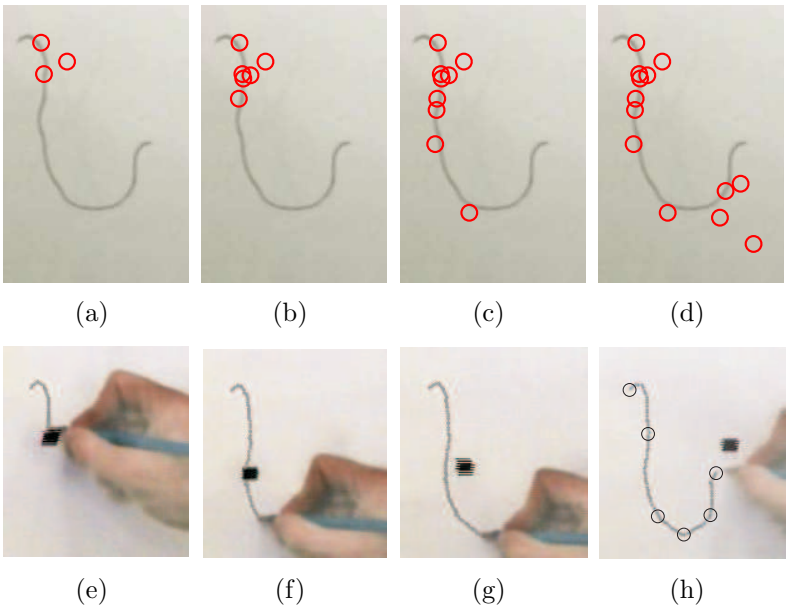


Figure 4.12: The sequence of eye and hand movements by subject **AP** in the drawing task. In the upper row, cumulative fixations on the original image are represented by red circles. In the lower row the solid black square denotes the gazepoint. In 4.12(h) the circles denote the endpoints points of each trajectory segment, found by inspection of the video recording.

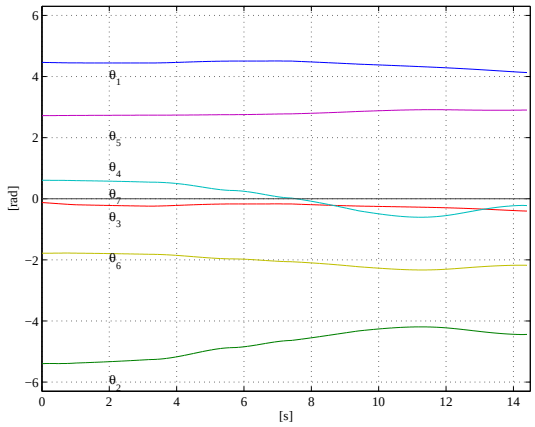


Figure 4.13: The time history of joint motions for the considered trajectory; angular variables correspond to the rotational degrees of freedom depicted in Fig. 4.7.

scanpath. Better results come from the comparisons with the perfect edge following and the proposed DBN algorithm. (respectively 0.40 ± 0.15 and 0.39 ± 0.16).

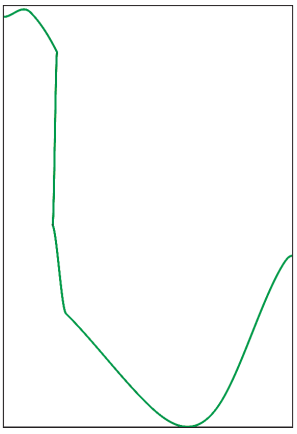


Figure 4.14: The actual position of the end effector computed via direct kinematics from the joint variables. The trajectory has been translated in world coordinates considering square pixels (1 mm).

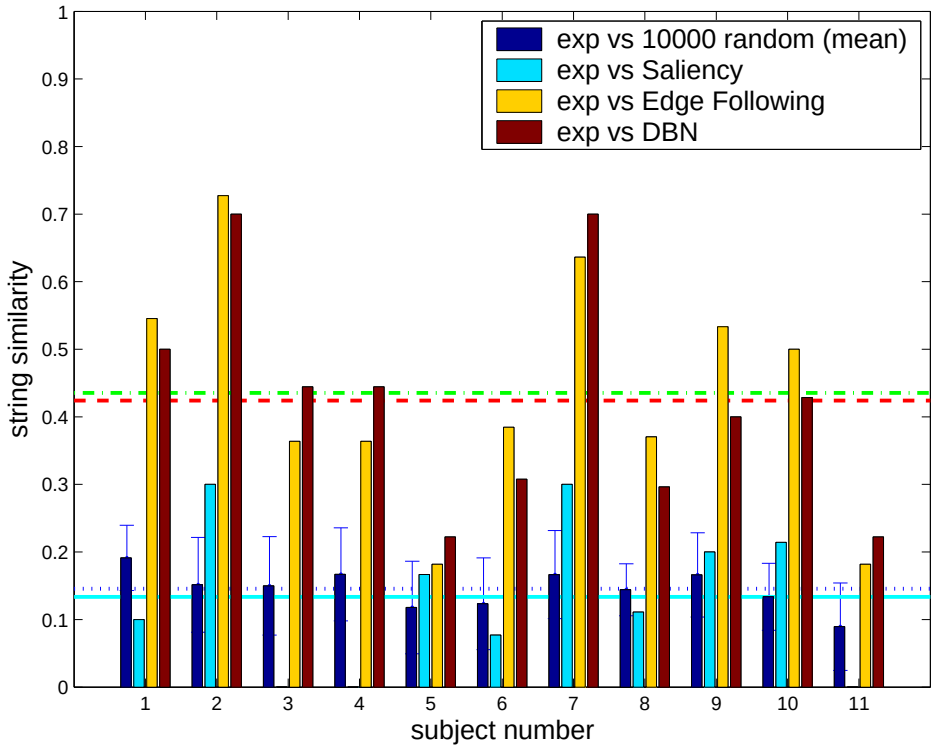


Figure 4.15: The plot shows for each subject (x axis) the mean similarity of the observed scanpath to 10000 random scanpaths (dark blue with error bar), the similarity to a preattentive scanpath à la Itti (light blue), the similarity to a perfect coarse–grained edge-following (yellow), and the similarity to the scanpath simulated by the DBN (red). Horizontal lines denote the respective mean values.

Chapter 5

Discussion

The research project presented in this thesis tackled the general problem of understanding the mechanisms underlying sensorimotor coupling in humans. The complementary issue of designing a sensorimotor control architecture for a situated artificial agent, was considered as well.

The first result of this work (see **chapter 2**) is that we provided the outline of a modular functional model of eye–hand coordination, inspired by the functional organization of the primate brain areas involved in sensory and motor processing; then, with the aim of providing in a principled way a computational theory of the underlying processes, we analyzed further the core modules, and formalized them by means of probabilistic techniques, namely a novel kind of Dynamic Bayesian Network, that we called *Input–Output Hidden Markov Model* (IOCHMM).
For this specific network — that to the best of our knowledge was never adopted before in the literature on sensorimotor coordination — we derived explicitly the analytic solution to the problems of inference (evaluating the conditional distribution of hidden states, given the observations), learning (estimating the most suitable values for the parameters that describe the generative model) and decision (selecting, at each time step, the most appropriate values for the variables that describe eye–hand actions, in order to solve the given task), in a particular case of interest, namely in the case of discrete variables.

The second achievement of the present work (see **chapter 3**) was to design and realize eye–hand tracking experiments in the case study we chose, namely realistic drawing. The first thing to notice is that

in the current literature on eye–tracking only few papers [123, 46] from a single research team reported experiments on this specific task; furthermore, while the above mentioned experiments were mainly *exploratory*, our analysis was instead driven by some preliminary hypotheses stemming out of a general theoretical description of the drawing process and its neural underpinnings.
As a matter of facts, the experimental results are in quantitative agreement with our three basic hypotheses (discussed in the opening of chapter 3). Probably the most interesting finding is that, due to the tight coupling of eye and hand movement generation in this specific task (in particular, the constraint that eye movements should support graphically continuous hand movements) we observed a scanpath that was never reported before, which resembles a coarse–grained edge–following on the contours of the original image to be portrayed.

Further results of this project are presented in **chapter 4**, where we discuss the implementation details for the computational model as applied to the drawing task. Such details include mainly the explanation of a number of simplifying assumptions, of the choices made for the state spaces of the stochastic variables, and the generation of a suitable training set.
Two points deserve further discussion here: first, at the end of the training procedure the DBN parameters are set to a value, to which corresponds a generative model (the joint probability distribution of all variables) that, together with the decision stage, represents a *sensorimotor map*; namely, a couple of eye–hand responses for each given collection of sensory stimuli (visual and proprioceptive). As shown in chapter 4, such map can be used as the core module of a bio–inspired situated artificial agent that should solve the given task.
The second point worth mentioning, is that the simulation results include a sequence of eye–hand movements on any given input image. Therefore, although the computational model was *not* designed to fit the experimental data (in other words, the DBN was *not* trained with experimental data), nevertheless it produces a kind of observable behavior that — much in the vein of *algorithmic explanations* [122] — can be directly compared to experimental observations. Thus, we first provided a qualitative (visual) comparison of eye–hand sequences produced by the model with those recorded on human subjects, and the similarity of the two behaviors was readily apparent. Eventually, we defined a mathematical procedure to obtain a numerical evaluation of how the simulated scanpath captures the regularities of the

observed ones, and the result was successful (particularly if compared with other existing influential models), notwithstanding the simplifying assumptions that are at the basis of our first implementation.

At this stage of the project, it can be said that the initial work plan has been fulfilled in terms of the development of a computational model, the realization of eye–tracking experiments, and the comparison of the respective results.
As a final remark, it is important however to acknowledge that a number of directions could be explored from this point on, and we like to conclude by evidencing some potentially interesting extensions.

- The simplifying assumptions we made in the implementation of the computational model could be weakened, in order to include continuous variables and noisy outputs for the DBN.

- The DBN for sensorimotor coupling could be based on a more general graph structure, taking advantage of existing techniques for structure learning that allow to define the most suitable conditional dependencies among variables on the basis of the training data.

- The learning technique adopted in our implementation provides a way to get point estimates of the relevant parameters; however, a fully bayesian approach to learning could be followed, in order to learn probability distributions on the parameters, that express uncertainty on the parameters due e.g. to a restricted training set.

- At a more general level, the formulation by means of probabilistic graphical models could be extended to the whole functional model, not only for the interest of formal consistency, but even more interestingly for including in a unified mathematical framework also the processes related to inverse kinematic and forward models (that are usually treated in the framework of optimal motor control, separately from the control of active sensors).

- On the experimental side, novel recording sessions should be planned to get a better recording of eye movements on the drawing hand (namely fixations providing visual feedback on hand

movements). Furthermore, numerical records of arm joints dynamics should be obtained as well, to enable a direct comparison with the hand motor outputs of the model.

- Many interesting extensions of the eye–hand tracking experiments could be tested, including a comparison of the behaviors of different groups of subjects (e.g. experts vs beginners), and a comparison of the behavior observed in different drawing modalities (e.g. realistic drawing, stylized drawing, drawing from memory or from imagination, ...)

- In perspective, it would be interesting to apply the functional model in the solution of sensorimotor tasks other than drawing (e.g. motion planning for mobile robots), in order to test the generality of the proposed architecture.

Bibliography

[1] Y.E. Aloimonos, A. Bandopadhay, and I. Weiss. Active vision. In *Proceedings First Int. Conf. Computer Vision, ICCV*, London, UK, 1987.

[2] F.C. Anderson and M.G. Pandy. Dynamic optimization of human walking. *Journal of Biomechanical Engineering*, 123:381–390, 2001.

[3] J. Atkinson, F.W. Campbell, and M.R. Francis. The magic number 4 +/- 0: A new look at visual numerosity judgements. *Perception*, 5(3):327–34, 1976.

[4] H. Attias. Planning by probabilistic inference. In *Proceedingd of the 9th International Conference on Artificial Intelligence and Statistics*, 2003.

[5] D.H. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[6] D.H. Ballard, M.M. Hayhoe, F. Li, and S.D. Whitehead. Hand–eye coordination during sequential tasks. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 337:331–338, 1992.

[7] D.H. Ballard, M.M. Hayhoe, P.K. Pook, and R.P.N. Rao. Deictic codes for the embodiment of cognition. *Behavioral and Brain Science*, 20:66–80, 1998.

[8] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annuals Mathematical Statistics*, 41:164–171, 1970.

[9] S. Becker and G.E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, 1992.

[10] R.J. van Beers, A.C. Sittig, and J.J. van der Gon Denier. Integration of proprioceptive and visual position-information: an experimentally supported model. *Journal of Neurophysiology*, 81:1355–1364, 1999.

[11] Y. Bengio and P. Frasconi. Input-output hmm's for sequence processing. *IEEE Transactions on Neural Networks*, 7:1231–1249, 1996.

[12] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Berlin, 2007.

[13] C. Blakemore and F.W. Campbell. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology*, 203:237–260, 1969.

[14] G. Boccignone, V. Caggiano, G. Di Fiore, A. Marcelli, and P. Napoletano. A bayesian approach to situated vision. In M. de Gregorio, V. Di Maio, M. Frucci, and C. Musio, editors, *Brain, Vision and Artificial Intelligence*, volume 3704, pages 367–376. Lecture Notes in Computer Science, 2005.

[15] J.S. Bruner. On perceptual readiness. *Psychological Review*, 64:123–152, 1957.

[16] G.T. Buswell. *How people look at pictures*. University of Chicago press, Chicago, 1935.

[17] V. Caggiano, A. De Santis, B. Siciliano, and A. Chianese. A biomimetic approach to mobility distribution for a human-like redundant arm. In *First IEEE-RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, Pisa*, 2006.

[18] R. Coen Cagli, P. Coraggio, G. Boccignone, and P. Napoletano. The bayesian draughtsman: A model for visuomotor coordination in drawing. In *Advances in Brain Vision and Artificial Intelligence*, volume 4729 of *LNCS*. Springer-Verlag, 2007.

[19] R. Coen Cagli, P. Coraggio, and P. Napoletano. Drawbot — a bio–inspired robotic portraitist. *Digital Creativity Journal*, 18(1):24, 2007.

[20] E. Datteri, G. Teti, C. Laschi, G. Tamburrini, P. Dario, and E. Guglielmelli. Expected perception: an anticipation-based perception-action scheme in robots. In *Proceedings of IROS 2003, IEEE/RSJ International Conference on Intelligent Robots and System*, pages 934–939, 2003.

[21] E. Datteri, G. Teti, C. Laschi, G. Tamburrini, P. Dario, and E. Guglielmelli. Expected perception in robots: a biologically driven perception-action scheme. In *Proceedings of ICAR 2003, 11th International Conference on Advanced Robotics*, volume 3, pages 1405–1410, 2003.

[22] G. Deco, O. Pollatos, and J. Zihl. The time course of selective visual attention: theory and experiments. *Vision Research*, 42:2925–2945, 2002.

[23] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[24] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Rev Neuroscience*, 18:193–222, 1995.

[25] A.T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, London, 2007.

[26] J. Duncan and G.W. Humphreys. Visual search and stimulus similarity. *Psychology Review*, 96:433–458, 1989.

[27] S.R. Ellis and J.D. Smith. Patterns of statistical dependency in visual scanning. In R. Groner, G. W. McConkie, and C. Menz, editors, *Eye movements and human information processing*, pages 221–238. Amsterdam: Elsevier Science Publishers, 1985.

[28] R. Engbert, A. Longtin, and R. Kliegl. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42:621–636, 2002.

[29] M.O. Ernst and M.S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415:429–433, 2002.

[30] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[31] G. Feng. From eye movement to cognition: Toward a general framework of inference. *Psychometrika*, 68:551–556, 2003.

[32] G. Feng. Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research*, 7:70–95, 2006.

[33] J.M. Findlay and I.D. Gilchrist. Visual attention: the active vision perspective. In M. Jenkins and L. Harris, editors, *Vision and Attention*. Springer Verlag, 2001.

[34] J.M. Findlay and I.D. Gilchrist. *Active Vision - The Psychology of Looking and Seeing*. Oxford University Press, 2003.

[35] J.M. Findlay and R. Walker. A model of saccadic eye movement generation based on parallel processing and competitive inhibition. *Behavioral and Brain Science*, 22:661–674, 1999.

[36] T. Flash and N. Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience*, 5:1688–1703, 1985.

[37] D.A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall International, 2002.

[38] V. Gallese. Embodied simulation: From neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences*, 4:23–48, 2005.

[39] V. Gallese and D. Freedberg. Mirror and canonical neurons are crucial elements in esthetic response. *Trends in Cognitive Sciences*, 11(10):411, 2007.

[40] W.S. Geisler and R.L. Diehl. Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London B*, 357:419–448, 2002.

[41] A.P. Georgopoulos, A.B. Schwartz, and R.E. Kettner. Neuronal population coding of movement direction. *Science*, 233:1416–1419, 1986.

[42] Z. Ghahramani. *Computation and Psychophysics of Sensorimotor Integration*. Ph.d. thesis, Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 1995.

[43] Z. Ghahramani, D.M. Wolpert, and M.I. Jordan. Computational models of sensorimotor integration. In P. G. Morasso and V. Sanguineti, editors, *Self-Organization, Computational Maps and Motor Control*, pages 117–147. Elsevier Press, 1997.

[44] D.R. Gitelman. Ilab: a program for postesperimental eye movement analysis. *Behavioral Research Methods, Instruments and Computers*, 34:605–612, 2002.

[45] M. A. Goodle and G. K. Humphrey. The objects of action and perception. *Cognition*, 67:181–207, 1998.

[46] E. Gowen and R.C. Miall. Eye-hand interactions in tracing and drawing tasks. *Human Movement Science*, 25:568–585, 2006.

[47] S.S. Hacisalihzade, L.W. Stark, and J.S. Allen. Visual perception and sequences of eye movement fixations: A stochastic modeling approach. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):474–481, 1992.

[48] C.M. Harris and D.M. Wolpert. Signal–dependent noise determines motor planning. *Nature*, 394:20, 1998.

[49] M.M. Hayhoe and D.H. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4), 2005.

[50] N. Hogan. An organizing principle for a class of voluntary movements. *Journal of Neuroscience*, 4(11):2745–2754, 1984.

[51] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews - Neuroscience*, 2:1–11, 2001.

[52] R.J. Jacob. What you look at is what you get: Eye movement-based interaction techniques. In *Human Factors in Computing Systems: CHI 90 Conference Proceedings*, pages 11–18, (1990).

[53] R.S. Johansson, G. Westling, A. Backstrom, and J. Randall Flanagan. Eye–hand coordination in object manipulation. *Journal of Neuroscience*, 21:6917–693, 2001.

[54] M.I. Jordan and D.M. Wolpert. Computational motor control. In M. Gazzaniga, editor, *The Cognitive Neurosciences*. Cambridge: MIT Press, 1999.

[55] E.M. Klier, H. Wang, and J.D. Crawford. The superior colliculus encodes gaze commands in retinal coordinates. *Nature Neuroscience*, 4(6), 2001.

[56] C. Koch and S. Ullman. Shift in selective visual attention: Towards th underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[57] K.P. Kording and D.M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427:244–247, 2004.

[58] K.P. Kording and D.M. Wolpert. Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10, 2006.

[59] G. Kreiman, C. Koch, and I. Fried. Category-specific visual responses of single neurons in the human medial temporal lobe. *Natural neuroscience*, 3:946–953, 2000.

[60] F. Lacquaniti, C. Terzuolo, and P. Viviani. The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica*, 54:115–130, 1983.

[61] M.F. Land. Predictable eye-head coordination during driving. *Nature*, 359:318–320, 1992.

[62] M.F. Land and S. Furneaux. The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 352:1231–1239, 1997.

[63] M.F. Land and P. McLeod. From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, 3:1340–1345, 2000.

[64] M.F. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311–1328, 1999.

[65] T.S. Lee and D. Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

[66] V.I. Levenshtein. inary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii. Nauk SSSR*, 163:845–848, 1965.

[67] J. Liechty and R. Pieters. Global and local covert visual attention: Evidence from a bayesian hidden markov model. *Psychometrika*, 68:519–541, 2003.

[68] D. Marr. *Vision*. Freeman, S. Francisco,CA, 1982.

[69] G.W. McConkie, P.W. Kerr, and B.P. Dyre. What are normal eye movements during reading: Toward a mathematical description. In J. Ygge and G. Lennerstrand, editors, *Eye Movements in Reading*, pages 315–327. Tarrytown, NY: Pergamon, 1994.

[70] G.W. McConkie, P.W. Kerr, M.D. Reddix, and D. Zola. Eye movement control during reading: I. the location of initial eye fixations. *Vision Research*, 28:1107–1118, 1988.

[71] G.W. McConkie and K. Rayner. The span of the effective stimulus during a fixation in reading. *Perception and Psychophysics*, 17:578–586, 1975.

[72] G. Metta, G Sandini, L Natale, L. Craighero, and L. Fadiga. Understanding mirror neurons: A bio-robotic approach. *Phenomenology and the Cognitive Sciences*, 2:197–232, 2006.

[73] R.C. Miall, D.J. Weir, D.M. Wolpert, and J.F. Stein. Is the cerebellum a smith predictor? *Journal of Motor Behavior*, 25:203–216, 1993.

[74] A.D. Milner and M.A. Goodale. *The Visual Brain in Action*. Oxford University Press, 1995.

[75] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[76] D. Mumford. On the computational architecture of the neocortex ii. *Biological Cybernetics*, 66:241–251, 1992.

[77] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Phd dissertation, Berkeley, University of California, Computer Science Division, 2002.

[78] Y. Nakamura. *Advanced Robotics: Redundancy and Optimization*. Addison-Wesley, Reading, Mass., 1991.

[79] P. Napoletano. *A Bayesian Approach to Situated Vision*. Phd dissertation, University of Salerno, Department of Electrical and Information Engineering Natural Computation Laboratory, 2007.

[80] T.A. Nazir and J.K. O'Regan. Some results on translation invariance in the human visual system. *Spatial Vision*, 5(2):81–100, 1990.

[81] U. Neisser. *Cognition and Reality. Principles and Implications of Cognitive Psychology*. W.H. Freeman, S. Francisco, CA, 1976.

[82] D. Noton and L. Stark. Scanpaths in saccadic eye movements during pattern perception. *Science*, 171:308–11, 1971.

[83] J.K. O'Regan. Optimal viewing position in words and the strategy–tactics theory of eye movements in reading. In K. Rayner, editor, *Eye movements and visual cognition: scene perception and reading*, pages 333–354. Springer–Verlag New York, 1992.

[84] J.K. O'Regan. Solving the 'real' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, 46(3):461–688, 1992.

[85] J.K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–1011, 2001.

[86] J. Pelz and R. Canosa. Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, 41:3587–3596, 2001.

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

[88] A. Pignocchi. Motor perception. perceiving pictures as artifacts. Manuscript under review, 2007.

[89] D. Popovic, R.B. Stein, N. Oguztoreli, M. Lebiedowska, and S. Jonic. Optimal control of walking with functional electrical stimulation: a computer simulation study. *IEEE Transactions on Rehabilitation Engineering*, 7:69–79, 1999.

[90] C.M. Privitera and L.W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000.

[91] Z. Pylyshyn. Is vision continuous with cognition? the case for cognitive impenetrability of visual perception. *Brain and Behavioral Science*, 2001.

[92] Z. W. Pylyshyn. Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80:127–158, 2001.

[93] Z.W. Pylyshyn. Situating vision in the world. *Trends in Cognitive Sciences*, 4(5):197–207, 2000.

[94] L.R. Rabiner. A tutorial on hmm and selected applications in speech recognition. In *Proceedings of IEEE*, pages 257–286, 1989.

[95] N. Ramnani. The primate cortico–cerebellar system: anatomy and function. *Nature Reviews Neuroscience*, 2006(7).

[96] E.D. Reichle, A. Pollatsek, D.F. Fisher, and K. Rayner. Toward a model of eye movement control in reading. *Psychology Review*, 105:125–147, 1998.

[97] I. Rezek and S.J. Roberts. Estimation of coupled hidden markov models with application to biosignal interaction modelling. In *NNSP International Workshop on Neural Networks for Signal Processing*, 2000.

[98] R.D. Rimey and C.M. Brown. Controlling eye movements with hidden markov models. *International Journal of Computer Vision*, 7(1):47, 1991.

[99] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umiltà. Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25:31–40, 1987.

[100] G. Rizzolatti, L. Riggio, and B.M. Sheliga. Space and selective attention. In C. Umiltà and M. Moscovitch, editors, *Attention and Performance*, volume XV, pages 231–265. Cambridge, MA: MIT Press, 1994.

[101] S. Saida and M. Ikeda. Useful visual field size for pattern perception. *Perception and Psychophysics*, 171:119–125, 1979.

[102] A. De Santis, P. Pierro, and B. Siciliano. The virtual end–effectors approach for human–robot interaction. In Lenarcic Roth, editor, *Advances in Robot Kinematics*. Springer, 2006.

[103] J. Saunders and D.C. Knill. Visual feedback control of hand movements. *Journal of Neuroscience*, 24:3223–3234, 2004.

[104] J.D. Schall. Neural basis of saccade target selection. *Reviews in the Neurosciences*, 6:63–85, 1995.

[105] K.H. Schlingensiepen, F.W. Campbell, G.E. Legge, and T.D. Walker. The importance of eye movements in the analysis of simple patterns. *Vision Research*, 26(7):1111–17, 1986.

[106] A.B. Schwartz. Direct cortical representation of drawing movements. *Science*, 265:540–543, 1994.

[107] E.L. Schwartz. Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, 20:645–69, 1980.

[108] O. Schwartz, A. Hsu, and P. Dayan. Space and time in visual context. *Nature Reviews Neuroscience*, 8(11), 2006.

[109] L. Sciavicco and B. Siciliano. *Modelling and Control of Robot Manipulators*. Springer-Verlag, London, UK, 2000.

[110] B.M. Sheliga, L. Craighero, L. Riggio, and G. Rizzolatti. Effects of spatial attention on directional manual and ocular responses. *Exp Brain Res*, 114:339–351, 1997.

[111] M. Shepherd, J.M. Findlay, and G.R.J. Hockey. The relationship between eye movements and spatial attention. *Quantitative Journal of Experimental Psychology*, 38:475–491, 1986.

[112] S. Shiori and M. Ikeda. Useful resolution for picture perception as a function of eccentricity. *Perception and Psychophysics*, 18:347–36, 1989.

[113] S. Shipp. The brain circuitry of attention. *Trends in Cognitive Sciences*, 8:223–230, 2004.

[114] S. Shipp. The importance of being agranular: a comparative account of visual and motor cortex. *Philosophical Transactions of the Royal Society B*, 360:797–814, 2005.

[115] A. Simpkins and E. Todorov. Optimal trade–off between exploration and exploitation. Manuscript under review, 2007.

[116] S.J. Sober and P.N. Sabes. Multisensory integration during motor planning. *Journal of Neuroscience*, 23:6982–6992, 2003.

[117] S.J. Sober and P.N. Sabes. Flexible strategies for sensory integration during motor planning. *Nature Neuroscience*, 8:4, 2005.

[118] N. Sprague, D. Ballard, and A. Robinson. Modeling embodied visual behaviors. *ACM Transactions on Applied Perception*, 4:2, 2007.

[119] P. Suppes. Eye-movement models for arithmetic and reading performance. In E. Kowler, editor, *Eye movements and their role in visual and cognitive processes*, pages 455–477. Amsterdam: Elsevier Science Publishers, 1990.

[120] G.G. Sutton and K. Sykes. The variation of hand tremor with force in healthy subjects. *Journal of Physiology*, 191:699–711, 1998.

[121] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[122] G. Tamburrini. *I Matematici E Le Macchine Intelligenti*. Bruno Mondadori, 2002.

[123] J. Tchalenko, R. Dempere-Marco, X.P. Hu, and G.Z. Yang. Eye movement and voluntary control in portrait drawing. In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*. Elsevier, Amsterdam, 2003.

[124] E. Todorov. Optimality principles in sensorimotor control. *Nature Neuroscience*, 7:907–915, 2004.

[125] E. Todorov and M.I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5:1226–1235, 2002.

[126] A. Treisman. The binding problem. *Current Opinion in Neurobiology*, 6:171–178, 1996.

[127] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[128] J. Trommershauser, L.T. Maloney, and M.S. Landy. Statistical decision theory and the selection of rapid, goal–directed movements. *Journal of the Optical Society of America: A*, 20, 2003.

[129] S. Ullman. Visual routines. *Cognition*, 18:97–159, 1984.

[130] L.G. Ungerleider and M. Mishkin. Two cortical visual systems. In D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield, editors, *Analysis of Visual Behavior*. MIT Press, Cambridge,, 1982.

[131] Y. Uno, M. Kawato, and R. Suzuki. Formation and control of optimal trajectories in human multijoint arm movement: Minimum torque–chqnge model. *Biological Cybernetics*, 61:89–101, 1989.

[132] P. Viviani. Eye movements in visual search: Cognitive, perceptual and motor control aspects. In E. Kowler, editor, *Eye movements and their role in visual and cognitive processes*, pages 353–393. Amsterdam: Elsevier Science Publishers, 1990.

[133] P. Viviani and T. Flash. Minimum-jerk model, two-thirds power law, and isochrony: converging approaches to the movement planning. *Journal of Experimental Psychology*, 21:32–53, 1995.

[134] D. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11:1317–1329, 1998.

[135] D.S. Wooding. Fixation maps: Quantifying eye-movement traces. In *Eye Tracking Research and Applications (ETRA) Symposium*, 2002.

[136] A.L. Yarbus. *Eye movements and vision.* Plenum Press, New York, 1967. English trans. by L.A. Riggs.

[137] J. Yedidia, W. Freeman, and Y. Weiss. Bethe free energy, kikuchi approximations, and belief propagation algorithms. Technical report, MERL, 2000.

[138] S. Zeki. *A Vision of the Brain.* Backwell Science, Oxford,UK, 1993.

[139] S. Zeki and M. Lamb. The neurology of kinetic art. *Brain*, 117:607–636, 1994.

[140] S. Zhong and J. Ghosh. Hmms and coupled hmms for multi-channel eeg classification. In *Proceedings of IEEE International Joint Conference on Neural Networks*, pages 1154–1159, 2002.