Università degli Studi di Napoli "Federico II"

Facoltà di Scienze Matematiche Fisiche e Naturali

ROMOACRE, a RObotic MOdel for ACtion REcognition

Tesi di Dottorato in Fisica Fondamentale e Applicata

XIX Ciclo

Tutore:
Prof. Giuseppe Trautteur

Autore:
Massimiliano Sorrentino

Coordinatore:
Prof. Gennaro Miele

Contents

Introduction

Action recognition is a task of everyday life which is performed by the central nervous system. The aim of this research is to build a neurophysiological model of action recognition in humans. The model proposed is called ROMOACRE (a **RO**botic **MO**del for **AC**tion **RE**cognition).

The results of the research are presented in this thesis, which is made up of five chapters, the first three of which present the background knowledge and previous studies, while the fuorth and fifth chapters illustrate a first attempt which we call the precursor of ROMOACRE and finally ROMOACRE.

In the first chapter we describe very briefly the central nervous system, which represents the largest part of the nervous system; together with the peripheral nervous system, it has a fundamental role in the control of behavior. The basic component of the CNS is the neuron, whose main role is to process and transmit information. Neurons have excitable membranes, which allow them to generate and propagate electrical impulses. The neuron is made up of a soma, an axon, the synapses, and dendrites. The dynamics of a neuron is based on the difference of potential between the interior and the exterior of the membrane. The axon is in a state of potential ON/OFF. In the ON state neurotransmitters are emitted, then neurotransmitters reach the receptors of the postsynaptic neuron. A flux of ions produces an exitatory or inhibitory potential. Then the potentials are summed in the soma and if the sum reaches a threshold the neuron enters the state ON (it fires or emits a spike). The significative variable is the frequency of the trains of spikes. Then we propose the formalized neuron by McCulloch and Pitts. After this we describe the lobules, the sulcus, the cortical areas and the circuits of the brain with special attention to visual and motor areas. We can distinguish two visual pathways: the form patway (ventral) and the motion pathway (dorsal). The first recognises forms in the visual field while the second recognises motion in the visual field.

In the second chapter we describe the mirror mechanism which is mainly involved in action recognition. Neurons in motor area F5 (see later) are divided in two classes: canonical neurons and mirror neurons. Canonical neu-

rons discharge only when a subject performs an action, while mirror neurons discharge also when a subject sees an action. Prior to the the discovery of mirror neurons by Rizzolatti it was known that the monkey (Macaca nemestrina) frontal cortex is subdivided into several different areas, among which area F5 is particularly interesting for its possible omology with Broca's area of human brain. With the term mirror neurons are indicated those neurons that become active when the monkey observes meaningful hand actions performed by the experimenter; the simple presentation of objects, even when held by hand, does not evoke the neuron discharge. The majority of mirror neurons (about 60%) are selective for one type of action (for instance grasping). Some are highly specific and fire selectively during the observation of a particular type of hand configuration used to grasp or manipulate an object (for instance precision grip, but not whole hand prehension). The remaining neurons are activated by observation of two or more hand actions. The actions most represented are: grasp, put object on a surface in front of a monkey, manipulate.

The activity of mirror neurons represents the action. Generally there are mechanisms which inhibit the observer to imitate the movements of the actor. When the observed action is of particular interest it can happen that a short initial part of the movement is executed by the observer. Then the actor recognizes an intention of the observer and the observer will notice that his involuntary answer will modify the behaviour of the actor. This will establish a primitive dialog which might be the basis of linguistic comunication. So, human language is probably evolved from a mechanism which originally was not related to speech comunication but to action recognition; similarly, a language grammar has evolved starting from the prelinguistic grammar.

The rostral part of monkey ventral premotor cortex is called area F5. Electro-physiological studies have shown that in this area there is a motor representation of mouth and hand actions. Neurons related to hand actions discharge when the monkey executes specific goal-directed hand actions such as grasping, holding, tearing, and manipulating objects. It has been proposed that these neurons constitute a sort of vocabulary of hand actions. As we already said, part of these neurons discharge both when the monkey performs specific goal-directed hand actions and when it observes another monkey or an experimenter performing the same or a similar action. These neurons are called mirror neurons because the observed action seems to be reflected as in a mirror, in the motor representation for the same action of the observer. Some MEP (Motor Evoked Potentials) studies have also been conducted on humans. During hand action observation there was an increase of amplitude of motor evoked potentials recorded from those hand muscles, normally recruited when the observed action is actually performed

by the observer. An fMRI (functional Magnetic Resonance Imaging) study was conducted on humans, which showed that the mirror neuron system is complex and related to different body actions performed not only with the hand, but also with the foot and the mouth. The actions showed could be either transitive (the mouth/hand/foot was acted upon an object) or intransitive (the mouth/hand/foot action was performed without an object). The observation of both transitive and intransitive actions, compared to the observation of a static image of the same action, led to the activation of different regions in the premotor cortex and Broca's area.

The third chapter deals with the meaning of "action" and "action recognition". An action is a generic movement of the human body, like moving a hand in the space; a meaningful action is an action which has a meaning, like walking, running, imitating grasping (without an object); obviously the meaningful actions constitute a subset of the set of the actions; a goal oriented action is a meaningful action in which the variables of the procedure (meaningful action) are instantiated. There are two possible meanings for action recognition: pragmatic recognition (the action is recognized when the observer is able to imitate it) and semantic recognition (the action is recognized when the observer is able to classify it). Following this distinction two models of action recognition are presented: the model by Giese and Poggio, which is a model of semantic recognition and the model by Demiris and Johnsson, which is a model of pragmatic recognition.

The model by Giese and Poggio is based on four assumptions: the model is divided into two parallel processing streams, analogous to the ventral and dorsal streams, that are specialized for the analysis of form and optic flow information, respectively; both pathways comprise hierarchies of neural features detectors that extract form or optic flow features with increasing complexity along the hierarchy. The position and size invariance of the feature detectors also increases along the hierarchy; the model assumes that the hierarchy in both pathways is predominantly feedforward (apart from local feedback loops), without the need of top-down signals. Although such signals might be important, in particular for longer stimulus presentations, good recognition performance can be achieved in most cases also without them. Recordings in the STS (Superior Temporal Sulcus) have found short latencies for the recognition of biological movements. None of these facts rules out the use of feedback processing. However, it indicates a hierarchical feedforward architecture as the core circuitry, underlying immediate recognition, that might be modulated by recurrent loops and higher-level interactions over longer time intervals; the representation of motion is based on a set of learned patterns. These patterns are encoded as sequences of snapshots of body shapes by neurons in the form pathway, and by sequences of complex

optic flow patterns in the motion pathway. This assumption is a central postulate of the model.

The model by Demiris and Johnsson is a combination of two kinds of models: inverse models and forward models. Inverse models are also known as controllers, or behaviours. Given a goal and the current state of the controlled system, they output the necessary motor commands that are needed in order to achieve or maintain that goal. These models are used frequently in control engineering and have been used for modeling motor planning and control. Inverse models have been hypothesized to exist in the premotor cortex (F5 mirror). Forward models are also known as predictors. Given motor commands and the current state of the controlled system, they output the predicted next state of that system. Like inverse models, they have been used in motor control modeling, and they have been hypothesized to exist in the cerebellum.

On the basis of some experiments on the motor cortex in humans and on brain disorders we state that action recognition is a two steps process: pragmatic recognition (the action is mentally reharsed, i.e. motor commands needed to perform it are extracted from visual information) and semantic recognition (the action is classified analyzing the motor commands). The model of action recognition proposed here is based on the assumptions indicated above.

Finally we report a work by Jacoboni about the role of mirror neurons in grasping the intention behind the goal.

In the fourth chapter we describe the precursor of ROMOACRE. ROMOACRE is composed of the following three computational stages: human pose estimation from images of body silhouette; evaluation of the motor commands of the action from human poses sequence; recognition of the action from motor commands. First of all we briefly present previous works about pose estimation. Then we present the k-NNW method which we will use in this model for pose estimation. Human pose estimation in ROMOACRE is performed in two stages: the first stage takes as input the raw data from the image and produces as output a vector of snapshot units selective for body silhouette shapes (form pathway); the second stage performs human pose estimation with the k-NNW method using as weights the output of the first stage. The form pathway is composed of three computational levels. The first level of the form pathway consists of local orientation detectors that model simple cells in the primary visual cortex (V1). Consistent with other models of simple cells, these detectors are modelled as Gabor filters. The second level of the form pathway contains position and scale tolerant bar detectors, which extract local orientation information. Within a limited range, thir responses are independent of the spatial position and scale of con-

tours within their receptive fields. They might correspond to complex-like cells in area V1, or to neurons that are increasingly invariant to position changes in areas V2 and V4. Many neurons in areas V2 and V4 are selective for more complex form features that are similar to corners or junctions. Such features were not necessary to achieve sufficent selectivity of the form pathway for body silhouette shapes. The third level of the form pathway contains snapshot neurons that are selective for body silhouette shapes. These model neurons are similar to view-tuned neurons in monkey inferotemporal cortex (IT) which are selective for complex shapes and can become tuned to complex shapes through learning. Neurons with a similar property in the cortex might be located in the STS of monkey and humans. Activity that is selective specifically for human body shapes has been found in the human lateral occipital complex, occipital and fusiform face areas and monkey STS. Then we assumed that the human pose is described by eight angles and we measured these angles on the subject with an alidade. We implemented the 2-NNW method with a neural network. Using another neural network we performed motor command evaluation. Finally we performed action recognition by confronting the evaluated motor commands with the known motor commands. We tested the model with three actions.

In chapter five we describe ROMOACRE. Roughly speaking it is composed of the same computational stages of its precursor: human pose estimation, evaluation of motor commands, action recognition. The first difference between ROMOACRE and its precursor is that in the precursor the actions are real, videos are captured and the poses are measured; in ROMOACRE, instead, actions are generated, videos are produced and poses are exported. In order to generate actions we used POSER, a third-party software which allows the creation of 3D motion of the human body. POSER produces BVH files which contain information about the poses (the body model we used is constituted by 19 3D joints). Three sets of 15 128x128 images of the silhouette of the human body performing each action was produced. In order to perform body pose estimation we extracted a set of shape descriptors from the image. This operation is performed in the form pathway, consisting of a hierarchy of five levels: S1 layer models simple cells; C1 layer models complex cells; S2 layer models composite feature cells; C2 layer models complex composite cells; VTU layer models view tuned cells. VTU layer is only introduced to test selectivity and invariance (in scale and position) of the model, but it is not a part of the model. Instead we stop the computation of the form pathway to the C2 layer, then in this model we make action recognition without performing image recognition according to experimental data. Anyway the model shows a good invariance for stimulus scale and position. We use the C2 units as shape descriptors and we perform a regression from

the space C2 (256 parameters) to the space of body pose (57 parameters). The regression method used is that of radial basis function neural network. As training set we used the 45 frames of the three actions generated. So, for each frame we extracted the C2 parameters and we know the body pose parameters from the BVH files. We evaluated the known motor commands for each action and the motor command of the percieved action. We cannot subtract the angles of two consecutive poses as they are Cardan angles. We can evaluate the rotation matrix for each joint, i.e. the matrix that rotate the parent joint to the child joint. Then we define a motor command as the matrix that transforms the previous joint into the next joint. At this point the model is able to perform pragmatic recognition. In order to perform semantic recognition we found a method to compare the perceived motor commands with the known motor commands. The model has been tested with three actions.

Chapter 1

The central nervous system

In this chapter part of the biological nervous system will be summarily described with special attention to the biological neuron, the cortical areas, and their connections.

The central nervous system (CNS) represents the largest part of the nervous system. Together with the peripheral nervous system, it has a fundamental role in the control of behavior.

The CNS originates from the neural plate, a specialised region of the ectoderm, the most external of the three embryonic layers. During embryonic development, the neural plate folds and forms the neural tube. The internal cavity of the neural tube will give rise to the ventricular system. The regions of the neural tube will differentiate progressively into transversal systems. First, the whole neural tube will differentiate into its two major subdivisions: spinal cord (caudal) and brain (rostral). Consecutively, the brain will differentiate into brainstem and prosencephalon. Later, the brainstem will subdivide into rhombencephalon and mesencephalon, and the prosencephalon into diencephalon and telencephalon.

The CNS is covered by the meninges, the brain is protected by the skull and the spinal cord by the vertebrae. The rhombencephalon gives rise to the pons, the cerebellum and the medulla oblongata, its cavity becomes the fourth ventricle. The mesencephalon gives rise to the tectum, pretectum, cerebral peduncle and its cavity develops into the mesencephalic duct or cerebral aqueduct. The diencephalon gives rise to the subthalamus, hypothalamus, thalamus and epithalamus, its cavity to the third ventricle. Finally, the telencephalon gives rise to the striatum (caudate nucleus and putamen), the hippocampus and the neocortex, its cavity becomes the lateral (first and second) ventricles.

The basic pattern of the CNS is highly conserved throughout the different species of vertebrates and during evolution. The major trend that can be

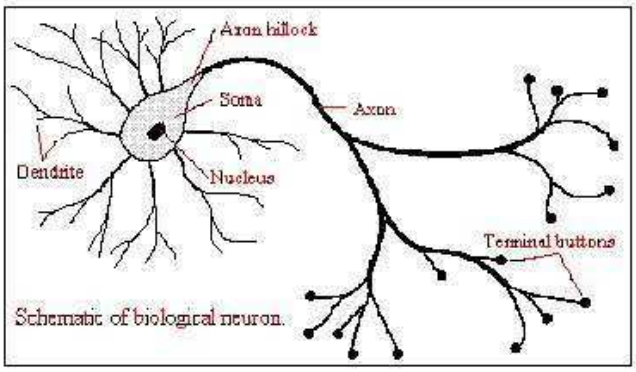

Figure 1.1: The biological neuron.

observed is towards a progressive telencephalisation: while in the reptilian brain that region is only an appendix to the large olfactory bulb, it represents most of the volume of the mammalian CNS. In the human brain, the telencephalon covers most of the diencephalon and the mesencephalon. Indeed, the allometric study of brain size among different species shows a striking continuity from rats to whales.

1.1 The biological neuron

1.1.1 The components of a neuron

Neurons are a major class of cells in the nervous system; they are sometimes called nerve cells, though this term is technically imprecise, as many neurons do not form nerves. In vertebrates, neurons are found in the brain, in the spinal cord and in the nerves and ganglia of the peripheral and autonomic nervous systems. Their main role is to process and transmit information. Neurons have excitable membranes, which allow them to generate and propagate electrical impulses.

They have cellular extensions which send and receive information. Their

size is typically 4 to 100 micrometres in diameter, depending on the type of neuron and the species it is from. Most neurons are highly specialized and differ widely in appearance.

They can be considered as made by the following different parts (fig. 1.1):

- the soma: the central part of a neuron;
- the axon: the (long) thread coming out from the soma;
- the synapses: the small branches of an axon. Each synapsis ends with a terminal button. The terminal buttons are connected to the dendrites of other neurons. The space between a terminal button and a dendrite is called synaptical gap.
- the dendrites: the short threads which come out from the soma.

1.1.2 Dynamics of a neuron

The dynamics of a neuron is based on the difference of potential between the interior and the exterior of the membrane. The potential of the interior of the membrane is negative compared the the potential of the exterior of the membrane, the difference being about -70 mV. Such difference of potential, called resting potential, is due to a different concentration of ions. The species involved are Na^+, K^+ and Cl^-; the different concentration of such ions causes the difference of potential indicated above.

The membrane is traversed by ionic channels which, in certain conditions, allow the passing of ions from the interior to the exterior of the membrane; moreover the ionic channels are sensitive to the difference of potential between the interior and the exterior of the membrane. These ionic channels can be considered as doors which stop or allow the passing of ions. According to chemical elements involved there are two types of ionic channels:

- ionic channels reacting to Na^+, which stop or allow the passing of Na^+ through the membrane;
- ionic channels reacting to K^+, which stop or allow the passing of K^+ throgh the membrane.

It is possible to stimulate a neuron in such a way that a difference of potential is produced which depolarizes the membrane. This happens because of the variation of permeability of the membrane following the opening of ionic channels which allow the passing of positive ions from the exterior to the interior of the membrane. If the stimulus is larger than a threshold

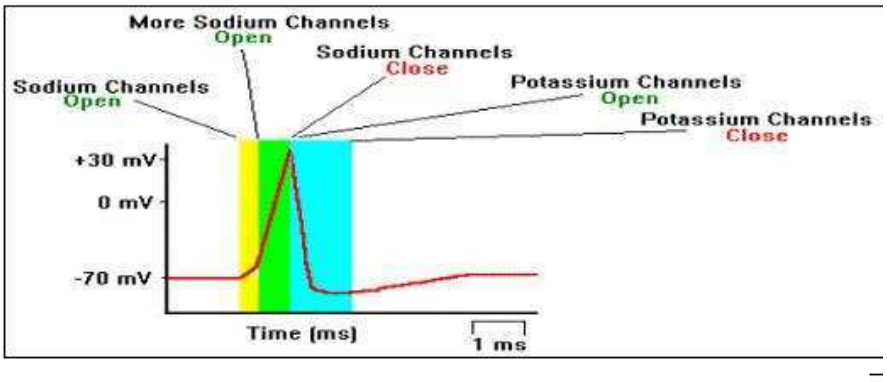

Figure 1.2: The action potential of a neuron.

of approximately 50 mV the opening of the Na^+ channels increases as it is sensitive to the difference of potential. In this way a large number of ions enter the cell. Such stream generates the action potential; it is said that the neuron fires or that there is a spike. The action potential has a maximum value of about 35 mV (fig. 1.2).

After the action potential follows the closing of the ionic channels sensitive to Na^+ ions, while those sensitive to K^+ ions get open. These produce the depolarization of the membrane thus bringing positive charges to the exterior of the cell until the resting potential is reestablished. The action potential is propagated along the axon of the neuron up to the terminal buttons. When the action potential reaches the terminal buttons the neurotransmitters are released in the synaptical gap. These neurotransmitters link to the receptors of the postsynaptical membrane (i.e. the membrane of the dendrites of another neuron) causing (generally) the opening of the ionic channels sensitive to Na^+ and, consequentely, a depolarization of the membrane. When a sufficient number of dendritic receptors are activated and the depolarization of the membrane reaches a typical threshold value, then also in this neuron there will be an action potential.

Moreover, the so called sodium pump is required to expel sodium ions

from the interior of the nerve axon so that the interior sodium ion concentration is held to about 10% that of the exterior fluid. At the same time the pump drives potassium ions from a low external concentration to a 30 times higher internal concentration. The pumping rate must keep up with the leakage of the two kinds of ions and the influx of ions at the occurrence of the spike.

To summarize, the dynamic of the communicative process among neurons is the following:

- the axon is in a state ON/OFF. In the state ON it propagates the signal: action potential or spike. The shape and the amplitude of the propagated signal is very stable: the signal has always the same shape and amplitude. In the state OFF there is no signal propagated along the axon;

- when the signal reaches the end of an axon (the terminal buttons) it causes the secretion of neurotransmitters (molecules) towards the postsynaptic membrane (we recall that there is no contact between the terminal buttons and membrane);

- the neurotransmitters reach the postsynaptical membrane. On the postsynaptic side these neurotransmitters link to receptors thus causing the opening of ionic channels which will produce an ionic stream in the postsynaptical neuron. The amount of the ionic stream entering the neuron - caused by the presynaptic spike - is a parameter that specifies the efficiency of the synapsis;

- the postsynaptical potentials (PSP) caused by the entering ionic streams are summed in the soma. Each single PSP is valued about 1 mV. These potentials can be activators or inhibitors. The activators depolarize the postsynaptic membrane and increase the possibility of a spike in the postsynaptic membrane. The inhibitors iperpolarize the postsynaptic membrane and decrease the possibility of a spike in the postsynaptic neuron;

- if the total sum of the PSP which reaches a given threshold the probability of a spike in the postsynaptic neuron becomes very high. The threshold value is about tens of mV (50 mV);

- the period from the spike of the presynaptical neuron to the spike of the postsynaptical neuron is around 1-2 ms;

- there is a period of about 1-2 ms during which the neuron cannot produce a second spike. This period is called absolute refractary period. The maximum frequency of a spike is then 500-1000 periods per second;
- the duration of a spike is around 5-7 ms.

1.1.3 The formal neuron

The model of neuron which we consider here is the original schematic model by McCulloch and Pitts which is made of:

- an elementary computational unit n_i representing the cell body (soma);
- a certain number of input lines (channels, connections) logically connected to N_i;
- each input line is a combination of a dendrite with a terminal button therefore the number of input lines is equal to the number of connections between the terminal buttons and the dendrites of a neuron;
- the input channels are activated by signals received from the logical input to which they are connected. Such logical inputs represent the presynaptic action. This can either activate the input channel (the axon carries an action potential) or not activate the input channel (there is no action potential on the presinaptic axon);
- a single output line (channel, connection) represents the fact that a single neuron produces one single relevant output (produces a spike or not);
- to each single output line is associated a parameter w_{ij}: the index i refers to neuron (postsynaptic) which we are considering in the hipothesis of many neurons, the index j refers to the various input channels of such neuron. The value of w_{ij} expresses the efficiency of the connection between a terminal button and the neural dendrite when there is an action potential on such button;
- each logical input is represented by a variable x_i which takes values 0, 1. $x_i = 1$ when it activates the input channel, and $x_i = 0$ when it does not activate the channel;
- the output channel is represented by a variable o_i which takes values 0, 1. $o_i = 1$ if on the output channel there is a spike, otherwise $o_i = 0$.

A number of further models have been introduced which do not concern us here.

1.2 The brain cortex

This section describes the lobules and the sulcus, the cortical areas, and the circuits of the brain (fig. 1.3). This description is generally acceptable both for the brain of the humans and for the brain of the macaque (macaca nemestrina) and it is based, as concernes the cytoarchitecture, on Broadmann's classification of the cortical areas.

1.2.1 The brain lobules and sulcus

The brain cortex is divided in four main lobules:

- frontal lobule;
- parietal lobule;
- temporal lobule;
- occipital lobule.

It contains the following sulci:

- occipital temporal sulcus;
- calcarine sulcus;
- principal sulcus;
- superior temporal sulcus;
- lateral sulcus;
- lunate sulcus;
- superior temporal sulcus;
- inferior occipital sulcus.

1.2.2 The cortical areas

Here follows a short description of the cortical areas which we are interested in:

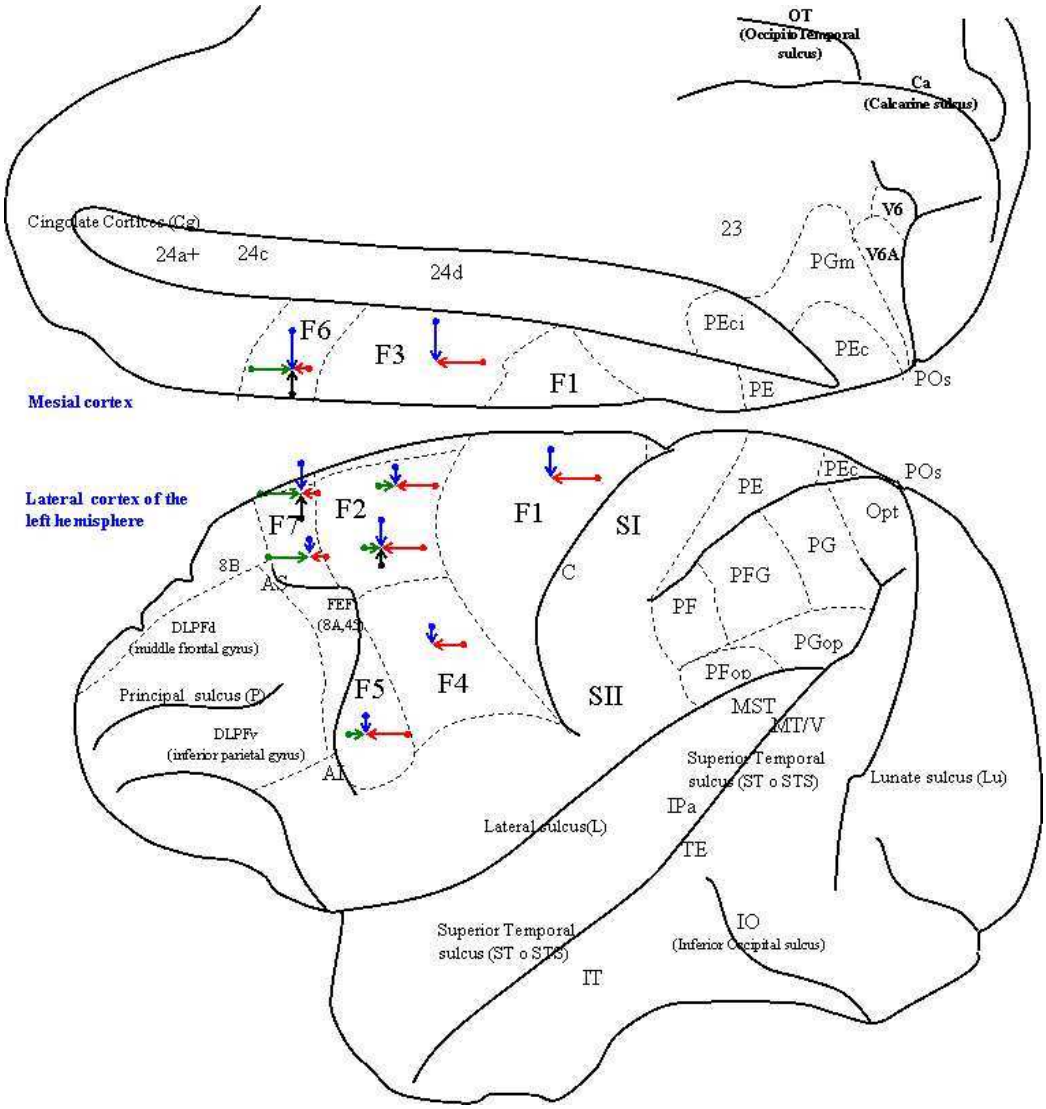

Figure 1.3: Cortical areas. The arrows indicate the connections among the areas.

- MT/V5: the MT area, also called V5, is positioned in the interior of the superior temporal sulcus, close to MST area (middle superior temporal area). MT area receives signals from layer 4b in V1, from the thick stripe in V2 and from V3. MT area is the center of motion perception. In MT area neurons are selective to the direction and speed of motion, but are non-selective to orientation and wave lenght. Their main task is to analyze the global motion of objects and surfaces, whereas neurons in V1 and V2 extract local motion of edges. The behaviour of MT neurons is very similar to our motion perception. MT area has a retinotopic map, though the receptive field of its neurons is generally very large;

- LIP: the LIP area is reciprocally connected with V3A, FEF, V6, V6A, V3, V4, MST, PG, VIP, PE. Furthermore it receives connections from PF, 7m and sends connections to MT, TE, TEO. LIP receives and integrates visual information concerning the space and the object from the areas of the dorsal and ventral stream, with the information from somatosensorial areas and sends this information to the visuomotor AIP area. LIP is related to the preparation of saccadic movements, to visual answers for 3D vision, to aspects like attention and anticipation. LIP is active also during reaching and grasping movements. The responses of neurons in LIP are similar to those of AIP;

- AIP: this area is involved in the circuit of the AIP-F5 canonicals (see chapter 2). It displays three different classes of neurons:

  - motor dominant neurons: they are active during the manipulation of objects in light or in dark conditions but they are not active during only the observation of objects. Many of them are more or less selective in respect to various objects;
  - visual and motor neurons: they are active both during observation and manipulation of objects. They are less active during manipulation in the dark and during the observation of the object;
  - visual dominant neurons: they are active only during manipulation in light condition and during the observation of the object.

  Motor dominant neurons are not influenced by the object position. Visual and motor neurons are selective for the same type of object in case of observation and manipulation;

- VIP: in this area there are two main classes of neurons activated by sensorial stimuli: visual neuron and bimodal neurons. This area is involved in the VIP-F4 circuit.

Visual neurons are strongly selective in respect to direction and speed of the stimulus. Their receptive field is generally wide. A considerable percentage of these visual neurons codifies the space in respect to the body of the agent. They react in a selective way to a visual stimulus that moves in space within a reachable distance, close to a specifical part of the body.

Bimodal neurons react to both visual and tactile stimuli. Tactile receptive fields are mainly located on the face. For these neurons the visual receptive field includes a 3D region around a tactile receptive field. The visual and tactile receptive fields correspond to:

- location: a neuron which responds, for example, to a visual stimulus produced in the right high corner of the visual field reacts also to a light touch of the skin in the right high corner of the face;
- dimension: if a neuron has a wide receptive field it has also a wide somatosensorial receptive field;
- direction: a neuron which reacts to a visual stimulus moving towards a certain direction reacts also to a somatosensorial stimulus which is moving in the same direction.

This correlation is valid also for eye movements; that is to say that the visual space to which a neuron reacts stays anchored to the somatosensorial receptive field to which it also reacts.

- Cingulate: it has some connections from 24d towards F1, few connections from 24d towards F2d, various connections from 24a+b, 24d towards F2vr, strong connections from 24d and 24c towards F3, few connections from 24c towards F4, various connections from 24c and to a less extent, from 24a+b towards F7 SEF, few connections from 24a+b and 24c towards F7, strong connections from 24c and, to a less extent, from 24a+b to F6;

- Prefrontal: there are few connections from 46d towards F2d, F2vr, and F5; strong connections from 8a, 8b, 45, 12, and to a less extent from 46d, 12arb towards F7. There are strong connections from 46d and 8B towards F7 ventral, strong connections from 46d, 46v and, to a less extent, 8B towards F6;

- DLPF (dorso lateral prefrontal cortex): it is made of a dorsal part (DLPFd or middle frontal gyrus) and a ventral part (DLPFv or inferior frontal gyrus) separated by a main sulcus (P). The caudal part of

DLPFv and DLPFd form the Frontal Eye Field (FEF, area 8A and 45 of Walker, this one situated under 8A). Close to 8A area there is 8B area. Rostrally to FEF we found area 46. FEF area receives connections from V4 and MT regions, from LIP area and from area 46. It sends strong connections towards area F7 SEF;

- Arcuate sulcus: it is situated in the frontal lobule and is made of two zones: superior arcuate sulcus (AS) and inferior arcuate sulcus (AI);

- F1 (primary motor area): it is situated in area 4. From a functional point of view it controls the arms, face, and legs. It is situated in the frontal part of the central sulcus on both lateral and mesial surface. From F1 about 20-30% neurons sends direct connections towards the motor neurons of the spinal cord. On F1 arrive connections from F2,F3,...,F5. The area is somatotopically organized, i.e. the output from a certain subarea of F1 controls a certain body part;

- F2: it is a premotor area situated rostrally to F1 and posteriorly to F7. It lies completely in the lateral cortex. From a functional point of view it controls arms and legs. Area F2 and F3 seem to belong to the same circuit. The neurons of F2 area send connections to F1 area, but there are also connections from F2 directly to the spinal cord. F2 can be devided in F2d (F2 dimple) and F2vr (F2 ventrorostral). F2d receives connections from PEc, PEip, PFG and less from 24d (cingulate) 46d (prefrontal). F2vr receives many connections especially from MIP, V6A (intraparietal sulcus) and from the cingulate; it receives less from 46d (prefrontal) and MST (temporal). For this reason F2 belongs to the parieto-dependent motor areas. It is possible to identify two circuits: PEip/PEc-F2d (it appears to be apparently involved in the planning and control of the movements on the basis of somatosensorial information) and MIP/V6A-F2vr (it appears to be apparently involved in the planning and control on the basis of somatosensorial and visual information);

- F3: it is a premotor area situated rostrally to F1 and posteriorly to F6. It lies basically in the mesial cortex (it is in area 6 and corresponds to the so called supplementary motor area, SMA proper). From a functional point of view it controls arms, legs and face. Area F2 and F3 seem to belong to the same circuit. Area F3 is organized in a somatotopical way and sends connections to area F1, though there are connections that go directly to the spinal cord. F3 receives mainly from F2 and F4 areas (25%), and from F5, F6, F7 areas (20%). It

receives many connections from PEc, PE, PEip (20%), from SII, PFG (posterior parietal zone), from the cingulate (24d, 24c) (20%) and from F1 (15%). For this reason it belongs to the parieto-dependent motor areas;

- F4: it is a premotor area situated in area 6. Functionally it controls the arms and the face. It is situated in the caudal zone of the frontal cortex between F5 and F1. Neurons of area F4 send connections to area F1, though there are connections that go from F4 directly to the spinal cord. F4 receives many connections from VIP, PEip, SII (intraparietal sulcus, anterior and posterior parietal zone) and few connections from the cingulate (24d, 24c). For this reason it belongs to the parieto-dependent motor areas. This area is involved in the VIP-F4 circuit. Many neurons of this area react to sensorial stimuli, especially to somatosensorial stimului. They can be subdivided in two classes: somatosensorial neurons and bimodal neurons (visual and somatosensorial). Somatosensorial neurons have tactile receptive fields, typically wide, localized on face, chest, arms and hands. Bimodal neurons have visual receptive fields and, as in the case of the VIP area, there is a correspondence with tactile fields and generally confined to peripersonal space. It is noticed that an increase of the stimulus speed causes an increase in the depthness of the receptive field. Furthermore, movement of the eyes does not imply a change in these fields. Many F4 neurons react also to reaching movements. Often a correlation is found between somatosensorial and visual receptive fields and the direction of movement (as an example, a neuron with a visual and tactile receptive field in the region of face responds also to movements towards the space over the shoulder);

- F5: it is a motor area situated in area 6. Functionally it controls the arms. It is located in the inferior part of the arcuate sulcus, in the frontal cortex of the caudal zone. In this area we distinguish two zones: F5c (dorsal convexity of the inferior arcuate sulcus) and F5ab (posterior bank of inferior arcuate sulcus). Both sectors receive many connections from the second somatosensorial area (SII, located in the inferior zone of the post central sulcus) and from PF/PFG area. F5ab receives input selectively also from AIP area (interior of the intraparietal sulcus). F5c is the main location of mirror neurons, while F5ab is the main location of canonical neurons. F5 area is involved in two circuits: PE-F5 mirror and AIP-F5 mirror. We notice that neurons of F5 area send strong connections to F1 area. Neverthless there are connections that go from

F5 directly towards the spinal cord where there are groups of neurons that control proximal movements. One hypothesis is that through F1 neurons more precise hands movements are controlled, while through the connections that go directly from F5 towards the spinal cord higher level hands movements are controlled;

- F6: it is an anterior motor area situated in mesial zone of area 6. It is located in the anterior part of the caudal zone of the frontal cortex rostrally to F3. From a functional point of view it controls the arms. It is supposed that F6 area represents a control unit which establishes when (because of internal or external conditions) a movement has to be performed. F6 does not send connections to F1, it receives connections from F5 and F7 (40%), from the prefrontal lobule (46d, 46v) (20%), from the cingulate lobule (24a+b) (20%), from F2, F3, F4 (15%), from the parietal lobule (PFG, PG) (15%) and from STP (5%). For this reason it belongs to the prefronto-dependent motor areas;

- F7: it is the anterior motor area situated, in area 6 of the parietal zone. From a functional point of view it controls the eyes. It is located in the anterior part of the caudal zone of the frontal cortex rostrally to F2. F7 can be divided in F7-SEF (supplementary eye field) and F7-nonSEF. F7-nonSEF and F6 belong to the same circuit. F7-SEF is involved in the oculomotor circuit. F7 does not send connections to F1. It has many connections towards the other motor areas. F7 receives many connections from the prefrontal lobule (8a, 8b, 45, 12, 46d, 12arb towards SEF and 46d, 8B towards nonSEF) and not many connections from the cingulate lobule (24c, 24a+b towards SEF and 24a+b, 24c towards nonSEF). It has enough connections from the temporal lobule (STP) towards SEF. It has enough connections from the parietal lobule (PGM, V6A, PG/PP) towards nonSEF. It has few connections from the parietal lobule (LIP) towards F7-nonSEF. For this reason it belongs to the prefronto-dependent motor areas. It is possible to distinguish two circuits: LIP-F7SEF (it can be important for controllong saccadic movements) and PGm-F7nonSEF (it can be important for the selection of conditional movements and for visualization of the stimulus in the space for reaching movements).

1.3 The cortical circuits

Here follows a short description of the cortical circuits which we are interested in:

- AIP-F5 canonical: we suppose that there is a AIP-F5 canonical circuit which allows the codification (AIP) of the geometrical properties of an object, which are basics in order to interact with the object. AIP-F5 canonical circuit transforms these geometrical properties into appropriate movements of the hand in order to interact with the object;
- PF-F5 mirror circuit: the role of this circuit is to detect the activity of mirror neurons as a system for the classification and recognition of actions;
- VIP-F4 circuit: the previous data support the hypothesis that VIP and F4 belong to a circuit for the codification of peripersonal space and for transforming the position of an object into an appropriate movement towards the object itself.

1.3.1 Parieto-frontal cortical circuits for actions control and space perception [41] [42] [43]

Motor areas (F1,F2,...,F7) can be divided into two main classes:

- parieto-dependent areas: F1,F2,...,F5. These areas have strong connections mainly with parietal areas belonging to the superior parietal lobule (SPL) and to the inferior parietal lobule (IPL) located in the posterior zone of the parietal lobule;
- fronto-dependent areas: F6 and F7. These areas have strong connections with prefrontal areas and with the cingulate.

The two classes are differentiated has follows:

- the fronto-dependent areas (F6 and F7) do not have direct connections with F1 and with the spinal cord. They send instead connections to the encephalic trunk;
- the parieto-dependent areas (F1,...,F5) send direct connections to the spinal cord (about 20-30% of area F1 sends directly to the motor neurons of the spinal cord). The areas F2,...,F5, moreover, have direct connections towards F1.

It is interesting to point that each single motor area receives generally strong connections from one single parietal area. In the same way each parietal area sends strong connections towards one single area. The pairs of parietal and motor areas which have strong connections have very similar functions. From what noted above it is possible to suppose that parietal and motor areas form a series of circuits independent from each other:

1. MIP/V6A-F2vr circuit (dorsal premotor cortex, PMd): it transforms visual and somatosensorial stimuli in order to control the movement of the arm towards a target;

2. FEF-F7SEF circuit (dorsal premotor cortex, PMd): it codifies the position of an object in space in order to orient and coordinate the arm-body movements;

3. V6A-PGm-F7 (dorsal premotor cortex, PMd): it codifies the position of object in space in order to orient and coordinate the arm-body movements;

4. IP-F4 circuit (ventral premotor cortex, PMv): it codifies the peripersonal space and transforms the position of objects into appropriate movements towards the same object;

5. AIP-F5 canonicals circuit (ventral premotor cortex, PMv): it extracts the specific characteristics of an object which are useful for manipulation;

6. PF-F5 mirror circuit (ventral premotor cortex, PMv): it represents the action, independently of its being executed or observed;

7. LIP-FEF circuit (ventral premotor cortex, PMv): it codifies spatial variables and it controls ocular movements.

We can try to include these circuits in a more general framewok. Historically the following event was observed: a ventral circuit (occipital-temporal-prefrontal), called WHAT pathway, which transformed visual information into a cognitive codification of the object; a dorsal circuit (occipital-parietal), called WHERE pathway, which transformed visual information into a codification of the position of the object. The information coming from these circuits was then integrated into the premotor areas in order to perform the action itself through the primary motor area (figg. 1.4, 1.5).

Such opinion is now considered obsolete as the dorsal channel (WHERE pathway) is in fact a channel directly dedicated to the transformation of visual and somatosensorial information into actions. Moreover the dorsal channel appears to be divided into two channels: dorso-dorsal and ventro-dorsal (fig. 1.6). The dorso-dorsal channel supplies the necessary information for action control. The ventro-dorsal channel has a fundamental role both in the organization of the action and in the perception of the space and of the action.

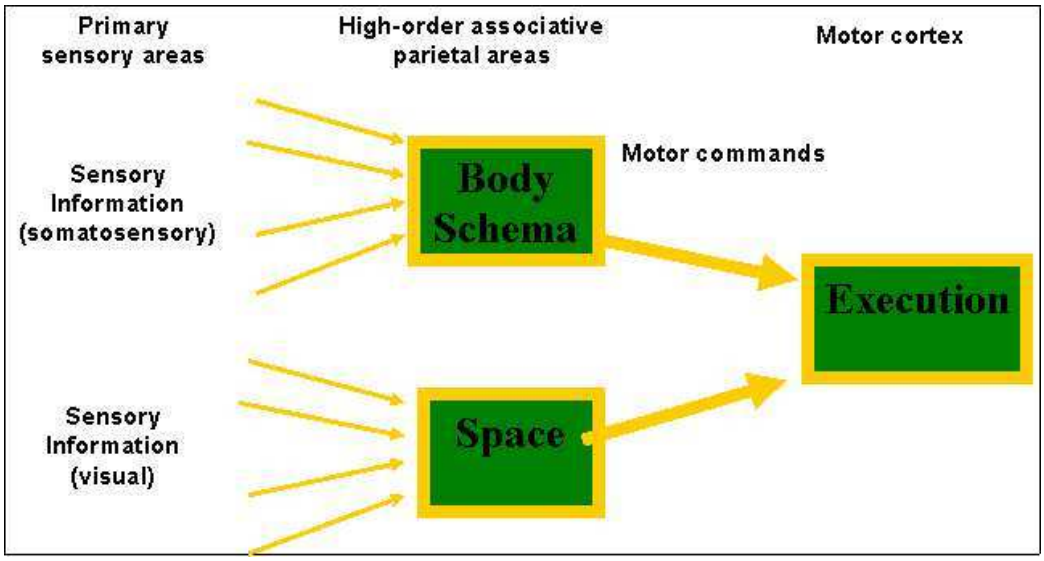

Figure 1.4: Action execution schema.

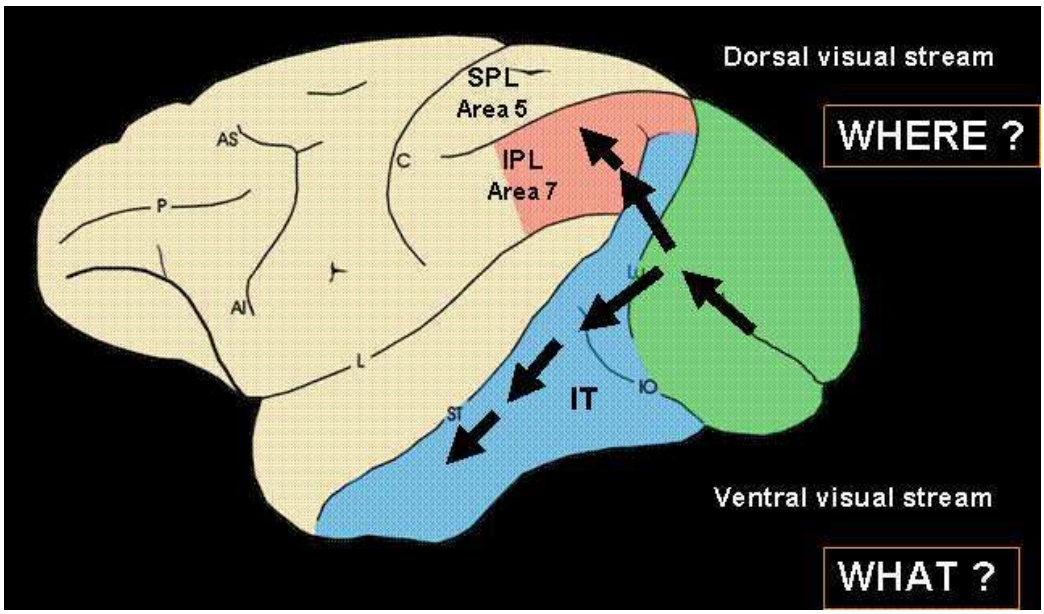

Figure 1.5: What and where pathways.

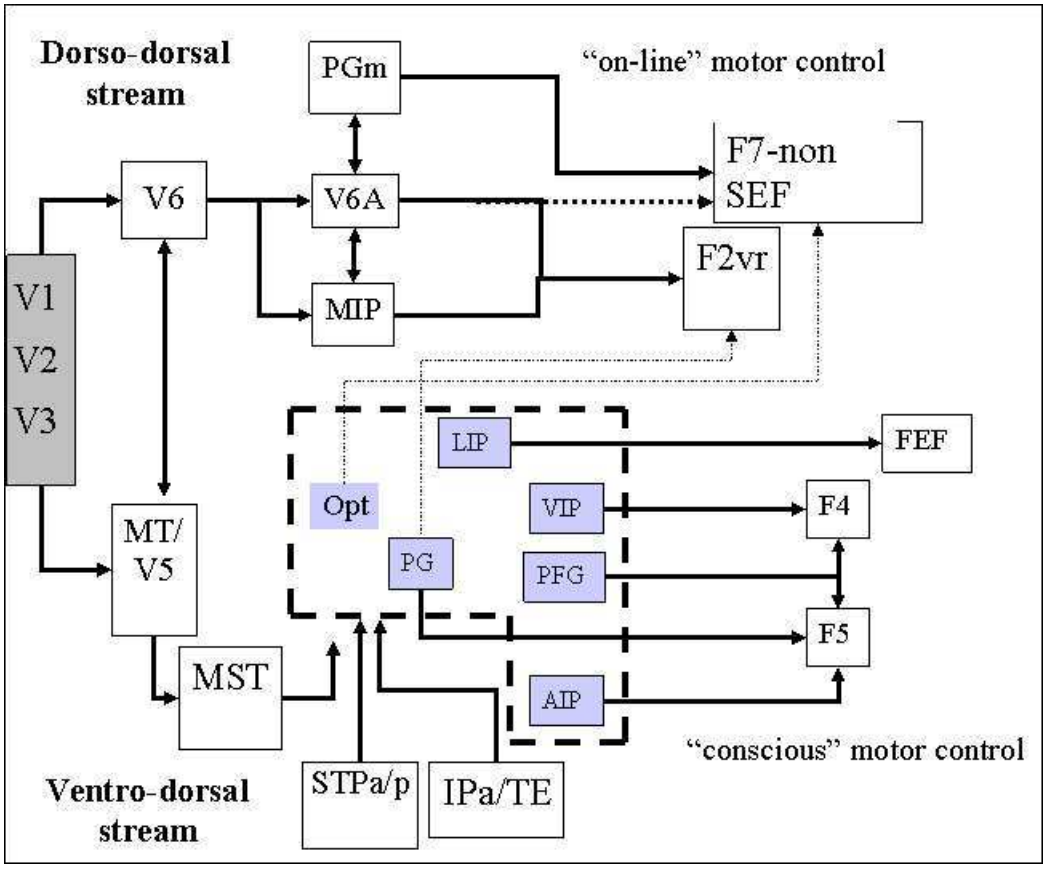

Figure 1.6: Dorso-dorsal and ventro-dorsal stream.

The dorso-dorsal channel is organized as follows: V6 receives strong connections from the visual areas V1(4b), V2, V3, V3A. V6 seems to be a purely visual area and it belongs cytoarchitectonically to the occipital cortex; neverthless it sends connections only towards SPL (superior parietal lobule). Specifically V6 sends mainly connections towards V6A (cytoarchitectonically closer to the parietal areas) and MIP. V6A and MIP, besided being reciprocally connected, form the two circuits $MIP - V6A \leftrightarrow F2Vr$ and $V6A/PGm \leftrightarrow F7 - nonSEF$.

The ventro-dorsal channel is organized as follows: MT/V5 shares with V6 the input coming from V1(4B). MT/V5 receives connections from V1 and sends connections to MST (besides being connected with V6). The areas of IPL are also connected with STPa/p (superior temporal polisensory areaa) and IPa/TE. IPL is involved in four basic cirsuits: $VIP \leftrightarrow F4$, $AIP \leftrightarrow F5canonical$, $PF \leftrightarrow F5mirror$, and $LIP \leftrightarrow FEF$.

Therefore at a more abstract level we can suppose to have (fig. 1.7):

1. a ventral channel which goes from the visual areas to the superior temporal sulcus up to the prefrontal areas 12/45 and from here either towards 46b, connected to the ventral motor areas (F2,F3,...) either towards 46d, connected to the dorsal motor areas (F6,F7);

2. a ventro-dorsal channel which goes from the visual areas, to the IPL, up to ventral motor areas (conscious motor control);

3. a dorso-dorsal channel which goes from the visual areas, to the SPL, up to the dorsal motor areas (online motor control).

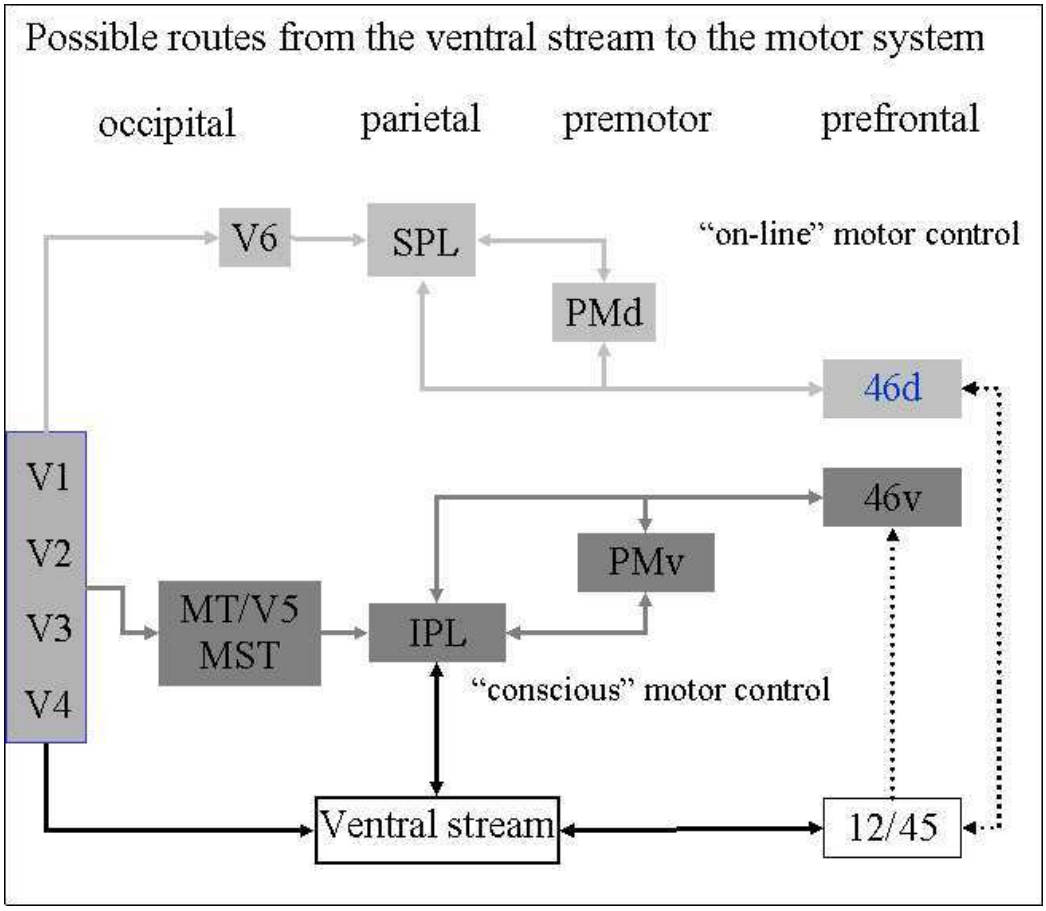

Figure 1.7: Possible routes from the ventral stream to the motor system.

Chapter 2

The mirror mechanism

This chapter deals with the mirror mechanism, which is mainly involved in action recognition.

In the next section we describe the cortical circuit in which the mirror neurons are involved.

2.1 AIP-F5 canonical circuit and PF-F5 mirror circuit

During a goal directed action(grasping, holding, putting down an object) performed by a subject, the area mainly involved is the area F5 situated in the inferior part of the arcuate sulcus (caudal zone of the frontal lobule).

In this area we distinguish two classes of neurons: purely motor neurons and visual neurons (which react to visual information). These last ones can be further subdivided into two groups:

- mirror neurons, situated mainly in F5c, dorsal convexity of the inferior arcuate sulcus;
- canonical neurons, situated mainly in F5ab area, posterior bank of the inferior arcuate sulcus.

Both sectors receive information from the second somatosensorial area (SII, inferior zone of the post central sulcus) and from PF area; F5ab receives selectively input also from AIP area (interior of the intraparietal sulcus). Generally the neurons of F5 area (canonical and mirror neurons) behave as follows: they are selective in respect to the type of action (to the goal of the action: catch, put down, hold an object) and to the modality of the action (grasp with fingers, grasp with full hand).

We can formalize what we said above in the following way:

- according to the type of action performed a special group of neurons is active, which are selective to the type of the action (also active during the whole action);
- according to the modality of the action another special group of neurons is active which are selective to the modality with which the type of action (i.e. grasp an object) is performed (grasp an object with fingers);
- some neurons which are selective to different specific phase of the action are active and not active during the action.

Mirror neurons have characteristics similar to other neurons in F5 area during the performance of an action. Then mirror neurons are also selective in respect to the type of action (goal) and to the modality of the action (grasp with fingers, grasp with full hand). Mirror neurons seem to become active mainly during the final phase of the action (when the hand reaches the object).

During the performing of an action, mirror neurons cannot be distinguished from the canonical neurons; this means that canonical neurons are also selective to the type and modality of the action.

The activity of mirror neurons only, or of canonical neurons only (caused by stimulation) should not produce movements and/or part of an action.

We can state that in order to have a goal directed action it is necessary to activate both mirror neurons and canonical neurons and, obviously, purely motor neurons. We notice that neurons of F5 area send strong connections to F1 area, where about 20-30% of the neurons send connections to the motor neurons of the spinal cord. Neverthless there are connections that go from F5 directly to the spinal cord towards groups of neurons which control proximal movements. One hypothesis is that through the neurons of F1 more precise hand movements are controlled, while through connections which go directly from F5 to the spinal cord higher level movements of the arm are controlled.

Mirror neurons are active also when the subject observes another subject to perform goal directed actions. There is a congruence between mirror neurons which are active when the subject performs an action and when the subject observes the performing of an action. This congruency can be more or less strong. We can summarize as follows:

- mirror neurons which are selective for the type of action performed, are active for the same type of action observed indipendently from the modality of the action observed;

- mirror neurons which are selective for the type and modality of action performed are active for the same type and modality of the action observed;

- also during the observation of the action there are possibly mirror neurons which become alternatively active and not active during the observation of the action.

Canonical neurons, instead, become active, beside the case when the action is performed by a subject, also when the observer just sees an object which can be grasped or manipulated, without any performance. These neurons are selective to the modality with which the object can be grasped. We can summarize as follows:

- if in the scene there is an object which can be manipulated a group of canonical neurons starts to be active;

- if the subject receives the command to grasp or manipulate the object, other canonical neurons, mirror neurons, and purely motor neurons of the area F5 become active;

- if the subject does not interact manually with the object but another subject starts to interact with the object it is not known whether the canonical neurons become active or not. It is known that prefrontal areas stop the activity of the F5 area if another subject grasps the object. Furthermore F6 area can distinguish wether a graspable object is within a reachable distance or not.

In case of an action performed by the subject the activity of canonical neurons and mirror neurons should be dependent from visual information (the object) and from the goal. They have the following biological counterpart:

- activity of the canonical neurons: it is based on input coming from AIP area and from F6 area which receives signals from prefrontal areas (area 24, cingulate medial zone, DLPF area);

- activity of mirror neurons: it is based on input coming from PF area and, possibly, visual areas.

In case of an observed action the activity of mirror neurons only should be dependent from visual information (object and external agent), which have the following biological correspondence; activity of mirror neurons is based on input coming from PF area and, possibly, from visual areas.

2.2 The discovery of mirror neurons [44]

This section describes the discovery of mirror neurons by Rizzolatti et al..

2.2.1 Discovery background

Prior to the the discovery of mirror neurons by Rizzolatti et al. it was known that the monkey (Macaca nemestrina) frontal cortex is subdivided into several different areas, among which area F5 is particularly interesting for its possible omology with Broca's area of human brain.

F5 is located in the ventro-rostral part of area 6, just caudal to the lower arm of the arcuate sulcus. F5 has enough somatotopical organization. Hands movements are represented mostly in its dorsal part, while mouth movements tend to be represented ventrally.

The properties of F5 neurons controlling hands movements were extensively studied and found to have both motor and sensory properties. As far as the motor properties are concerned they have two main characteristics:

- most neurons discharge selectively during goal directed hand movements such as grasping, holding and manipulating;
- many neurons are specific for particular types of hand prehension such as precision grip and whole hand prehension.

For the sensory properties, the most interesting aspect is that a considerable part of F5 neurons fire at a presentation of a 3D object, in the absence of a movement. In many cases the discharge occurs only if there is a match between the object size and the type of grip coded by the neuron.

F5 receives a strong input from the inferior parietal lobule and from AIP area, located in the lateral bank of the inferior parietal sulcus rostral to the oculomotor LIP area. As in the case of F5, a large number of neurons in AIP are related to hand movements. About 40% of AIP neurons discharge during the appropriate hand movements both in darkness or in the light (motor dominant neurons). The remaing neurons discharge stronger (visual and motor neurons) or exclusively (visual dominant neurons) in the light. Some of these neurons became active when the monkey fixates a still object and is not required to make a movent toward the object.

These data indicate that AIP and F5 form a cortical circuit which transforms visual information of the intrinsic properties of the object into hand movements that allow the monkey to interact with the objects. Motor information is then transferred to F1, to which F5 is directily connected, as well as to various subcortical centers for movement execution.

Rizzolatti et al. discovered that a particular subset of F5 neurons discharge when the monkey observes meaningful hand movements made by the experimenter. Following this discovery Rizzolatti called this particular subset "mirror neurons". Movements included placing or taking away objects from a table, grasping food, manipulating object. There was always a link between the effective observed movement and the effective executed movement.

These data suggest that area F5 has an observation-execution matching system. When the monkey observes a motor action that belongs to its movement repertoire, this action is automatically retrieved. The retrieved action is not necessarely executed; it is only represented in the motor system. The assuption was that this observation-execution mechanism plays a role in understanding the meaning of motor events.

Previous data were considered which showed that an observation-execution matching system does exist in man and that the cortical region involved in this matching is part of the region usually referred to as Broca's area.

2.2.2 Testing the mirror properties

Mirror properties were tested by performing, in front of the monkey, a series of motor actions related to food grasping (i.e. presenting the food, putting it on a surface, grasping it, giving it to a second experimenter or taking it away from him), to manipulation of food or other objects (i.e. breaking, tearing, folding), or intransitive gestures (non object related) with or without emotional content (threatening gestures, lifting the arms, waving the hand, etc.).

In order to verify whether the recorded neuron coded specifically hand-object interaction, the following actions were also performed:

- movements of the hand mimiking grasping in the absence of the object;
- prehension movements of food or other objects performed with tools (i.e. forceps, pincers);
- simultaneous combined movements of food and hands, spatially separated one from the other.

All experimenter's actions were repeated on the right and on the left of the monkey at various distances (50 cm, 1 m, 2 m).

2.2.3 The mirror neurons

Fig. 2.1 shows a lateral view of the monkey brain. Mirror neurons were recorded from the dorsal convexity of the cortex (shadowed area) and the

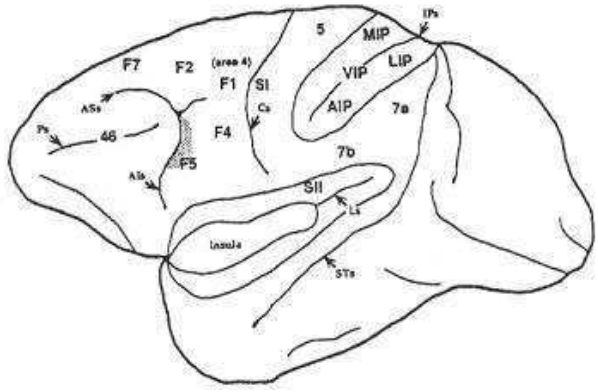

Figure 2.1: Lateral view of the monkey brain. The shadowed area shows the anatomical localization of the recorded neurons.

adjacent posterior bank of the arcuate sulcus. Both these cortices are part of area F5. Mirror neurons represented, approximately, 20% of the recorded neurons.

With the term mirror neurons were indicated those neurons that become active when the monkey observed meaningful hand actions permormed by the experimenter. The simple presentation of objects, even when held by hand, did not evoke the neuron discharge. The majority of mirror neurons (about 60%) were selective for one type of action (i.e. grasping). Some were highly specific selectively firing during the observation of a particular type of hand configuration used to grasp or manipulate an object (i.e. precision grip, but not whole hand prehension). The remaining neurons were activated by observation of two or more hand actions. The actions most represented were: grasp, put object on a surface in front of a monkey, manipulate.

2.2.4 The mirror mechanism

Neurons that are selectively activated by complex biologically meaningful visual stimuli have been observed in many high-order cortical areas and in the amygdala. These neurons respond to the sight of hands, faces and particular types of body movements. Among them are neurons located in the depth of the superior temporal sulcus, that are specifically responsive to hand-object interaction.

Mirror neurons of area F5 share with these complex neurons the property of being responsive to meaningful stimuly. Neurons with mirror properties have been described only in F5. It is likely, however, that they are not unique to this area, but do exist in other frontal and parietal cortical areas that control the organization of goal directed movements.

An explanation of mirror neurons that comes naturally to mind is that they are related to motor preparation. When the monkey observes an action, he starts automatically to prepare the same action. The preparation explanation is unsatisfactory for two reasons:

- the discharge of mirror neurons caused by the observation of a movement is not followed by the movement that, supposedly, was prepared;

- mirror neurons cease firing when the food is moved toward the animal and becomes available to him. If the firing of mirror neurons were related to motor preparation the neuron activity should have increased and not decreased in the phase that precedes movement execution.

A more sophysticated interpretation of mirror neurons was given by Jeannerod [45] who made the example of a pupil learning how to play a musical instrument. The pupil is completely still, watching the teacher who demonstrates an action that he must imitate and reproduce later. Although the pupil is immobile he must form, in his brain, an image of the teacher action. Jeannerod's view is that the neurons responsible for the motor schema formation are the same that the pupil will later activate during planning and preparation of the action. According to him mirror neurons are neurons that internally represent an action.

The explanation favoured by Rizzolatti is similar to that proposed by Jeannerod, in the sense that mirror neurons are neurons that represent internally actions. However whereas the emphasis given by Jeannerod is on learning, it is possible that mirror neurons play a role in the understanding of motor events.

The expression "understanding motor events" indicates the capacity of an individual to recognize the presence of another individual performing an action, to differenciate the observed action from other actions, and to use this information in order to act appropriately. Some of the mechanisms mediating operations of this type are linked to emotion and depend on the integrity of limbic structure. In contrast to the "understanding" based on the affective valence of the stimuli, the "understanding" mediated by the mirror neurons appears to be disjoint from emotional and vegetative responses. The meaning of the observed action does not result from the emotion it evokes, but from a

matching of the observed action with the motor activity which occurs when the individual performs the same action.

What is important to stress here is that the proposed mechanism is based on a purely observation-execution matching system. The affective valence of the stimuli, even if possibly present, does not play a role in this "understanding" system. Later we will present the importance of this point for understanding the development of the observation-execution matching system in man.

The presence of an observation-execution matching mechanism in monkey's premotor cortex suggests that a similar mechanism may exist also in man. To test this prediction the excitability of the motor cortex in a group of normal human subjects was studied. The subjects were stimulated in four conditions:

- while they observed an experimenter grasping 3D objects;
- while they looked at the same 3D objects;
- while they observed an experimenter tracing geometrical figures in the air with his arm;
- while detecting the dimming of a light.

Motor evoked potentials were recorded from arm muscles. The results showed a significant increase of the motor evoked potentials in the two conditions in which the subject observed movements. Furthermore, the increase was found only in those muscles that were active when the subjects executed the previously observed actions. Although it is premature to draw any firm conclusion on this last point, because only two type of movements were tested, neverthless the obtained data strongly suggest that in man there is an observation-execution matching system similar to that found in the monkey premotor cortex.

Admitted that an observation-execution matching system exist in man, the next problem is to assess where it is located. This problem was addressed using positron emission tomography (PET). The most striking result of the experiment was the presence, in grasping observation condition, of a highly significant activation in the posterior part of the left inferior frontal gyrus. This region corresponds to the rostral part of Broca's area as defined by Penfield and Roberts. Other active regions were presented in the occipital lobule and in the middle temporal gyrus.

Homologies between cortical areas of different species are always difficult to draw especially when one deals with the speech areas wich are unique

to humans. In man, the frontal region related to speech (Broca's area) is formed by areas 44 and 45 of Broadman. Area 44 has basically an agranular structure while in the second a granular layer is present.

Mesulam suggested that the ventral part of inferior area 6 (F5) and area 45 are the areas which might be homologs of the human frontal speech areas. F5 area might be the anatomical homologs of human Broca's area. Two major differences can be observed:

- in F5 there is a large hand representation, while Broca's area is classically thought of as an area related to the control of musculature responsible for spoken word production;
- F5 is an area receiving visual and somatosensory inputs, while Broca's area is mostly related to auditory input

In F5, in addition to the hand field, there is also a large mouth-face field located laterally to the hand field. It is very likely that in man this field has grown in relation to speech development and the great motor difficulies that speech poses. However a mouth field exist in F5 and, conversely, a hand field appears to exist in the Broca's region.

Furthermore, the mirror mechanism is supposed to be at the basis of the evolution of language. The activity of mirror neurons represents the action. Generally there are mechanisms which inhibits the observer to imitate the movements of the actor. When observed action is of particular interest it can happen that a short initial part of the movement is executed by the observer. Then the actor recognizes an intention of the observer and the observer will notice that his involuntary answer will modify the behaviour of the actor. This will establish a primitive dialog which is at the base of language comunication.

So, human language is probably evolved from a mechanism which originally was not related to speech comunication but to action recognition; similarly, a language grammar has evolved starting from the prelinguistic grammar.

Chapter 3

Pragmatic and semantic action recognition

3.1 Introduction

This chapter deals with the meaning of "action" and of "action recognition". Firstly we present what is meant by action (fig. 3.1):

- an action is a generic movement of the human body, like moving a hand in the space;
- a meaningful action is an action which has a meaning [1], like walking, running, imitating grasping (without an object); obviously the meaningful actions constitute a subset of the set of the actions;
- a goal directed action[2] is a meaningful action in which the variables of the procedure which implements the meaningful action are instantiated by a specific object. This is a very restricted interpretation of "goal directed" because, for instance, the action "grasping" is not goal directed unless it is instantiated in "grasping that cup".

We introduce now what is meant by action recognition. There are two possible meanings for action recognition:

[1] "Meaning" is not intended here as a concept of philosophy of the language. We use "meaningful action" to denote an action which belongs to an experimentally known repertoire of actions performed by the organism.

[2] "Goal directed action " is not intended in its common sense. It denotes an action in which an agent performes definite movements with respect to some object. So "Walking" should not be a goal directed action

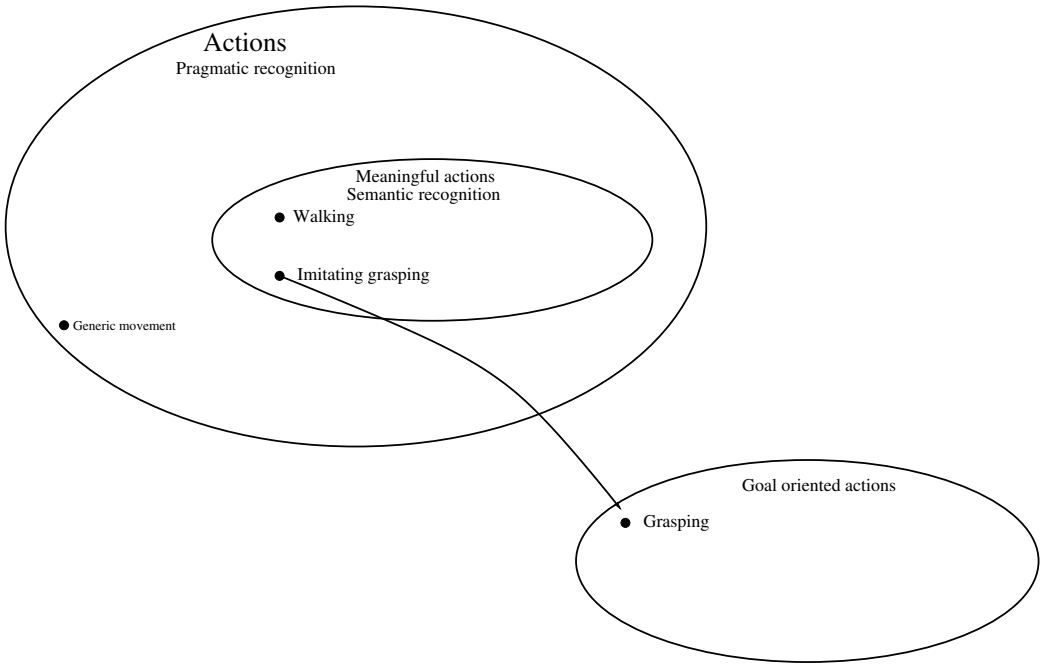

Figure 3.1: Type of actions.

- pragmatic recognition: the action is recognized when the observer is able to imitate it;
- semantic recognition: the action is recognized when the observer is able to classify it.

An action that does not belong to the set of meaningful actions cannot be recognized in the semantic way, while a meaningful action can be recognized in both pragmatic and semantic ways.

Here after we present two models of action recognition:

- the model by Giese and Poggio [15], which implements semantic recognition;
- the model by Demiris and Johnson [47], which implements pragmatic recognition.

3.2 The model by Giese and Poggio

Giese and Poggio developed a neural mechanism for the recognition of biological movements. The amount and complexity of data available in several

areas in cognitive neurosciences are continuosly increasing. Consequently, pure intuition, and the qualitative mental models associated with it, are becoming less appropriate for interpreting experimental results and for planning new experiments. Therefore quantitative computational theories can be an effective tool for summarizing existing data, and for testing the consistency of possible explanations.

These models may help to organize knowledge and to use it to provide explanations and to propose and plan new experiments.

Moreover they are also used as a framework to organize the results of the experiments.

In the model under discussion two main sets of questions are addressed:

- is it possible to semantically recognize biological movements in a way that is consistent with experimental data, and also uses plausible neural mechanisms?
- what are the roles of the ventral and dorsal pathways for the recognition of visual stimuli induced by biological movements?

This computational model gives an interpretation of the data and provides answers to these questions. It also points to issues that cannot be answered by the model and by the available experimental results. For instance:

- how is the information from the two pathways combined?
- what is the role of time in the ventral pathway?
- how does the attention influence the recognition process?

Notice that the actions considered are meaningful, but not goal directed.

3.2.1 Recognition of biological movements

Recognition of complex biological movements - gestures, facial expressions and motor actions - is biologically important for activities such as detecting predators, selecting prey and courtship behaviour. Gestures and facial expressions are also central to social communications in primates and humans. In fact humans recognize biological movements accurately and robustly, as shown in classical psychophysical work by Johansson [48] (fig. 3.2).

The information carried by biological movements is illustrated in the experiment by Gunnar Johansson. He attached ten light bulbs to the joints of actors, who were videotaped while they performed complex movements, such as walking, running, or dancing in the dark. From the videos, which

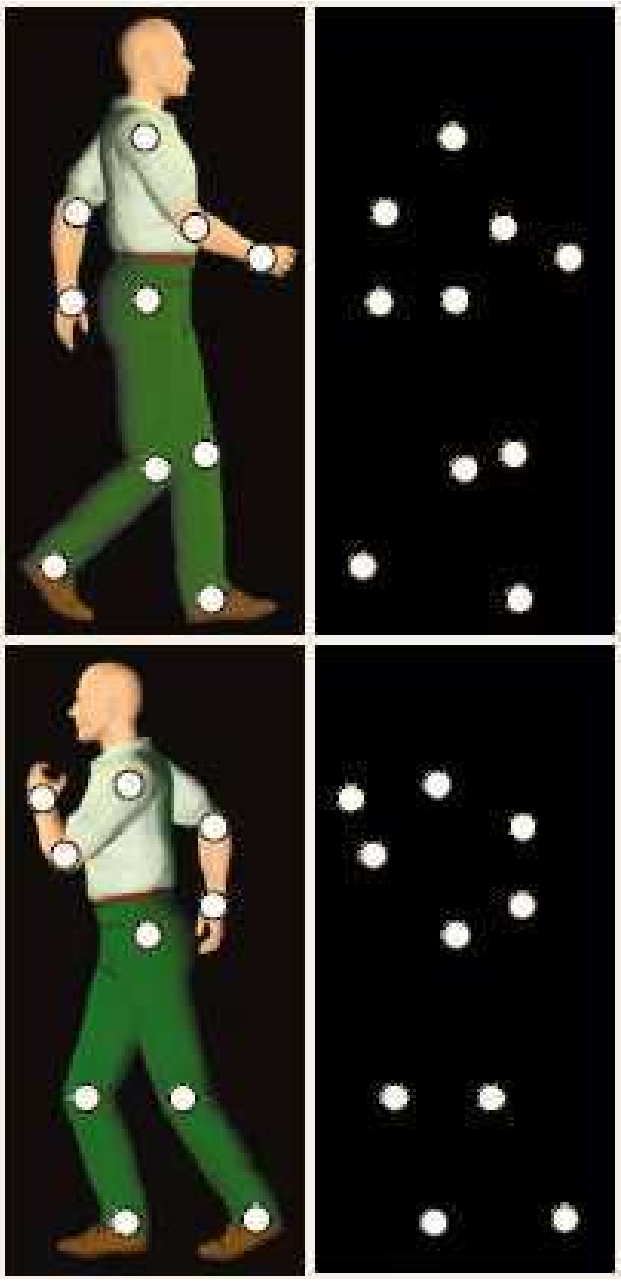

Figure 3.2: The experiment by Johansson.

showed only ten light dots moving against a dark background, subjects could immediately recognize the action. In addition, the dots were spontaneously interpreted as a human being. Furthermore if the subjects saw individual dot-frames from the videos presented as static pictures, they neither perceived the dots as human nor where able to identify the actions.

Subsequent studies showed that many complex actions can be recognized on the basis of such "point-light displays", including facial expressions, American Sign Language, arm movements, and various full-body actions.

3.2.2 The role of ventral and dorsal pathway

Visual information arriving at the cortex is first processed in the occipital lobule. From there, two main pathways project information to two different lobules:

- the dorsal pathway projects visual information into the parietal lobule;
- ventral pathway projects visual information into the temporal lobule.

The ventral pathway is hypothesized to play the major role in object identification. The temporal lobule receives visual input from the ventral pathway, and the object(s) in the visual scene are compared to stored representations in the object-memory system, also located in the temporal lobule. The ventral pathway is also hypothesized to be a "perception" pathway, in that it computes spatial relations among components of objects, allowing for their identification.

The dorsal pathway is hypothesized to play the major role in spatial localization of stimuli. The parietal lobule receives input not only form the dorsal pathway, but also from the auditory and somatosensory centers in the brain. This information is integrated into a coherent spatial representation of the world. The dorsal pathway is also hypothesized to be an "action" pathway, in that it computes spatial relations between the organism and the environment, thus allowing for the organism's effective interaction with the environment.

The roles of the ventral and dorsal visual processing streams in the recognition of biological movements are unclear. It seems likely that the dorsal pathway, which is specialized for the processing of motion information, contributes substantially to the perception of biological movements, in particular as the perception of actions is possible without well-defined form information. At the same time, when confronted with full pictures, instead of dot-pictures, subjects can recognize gait patterns from individual stationary key frames,

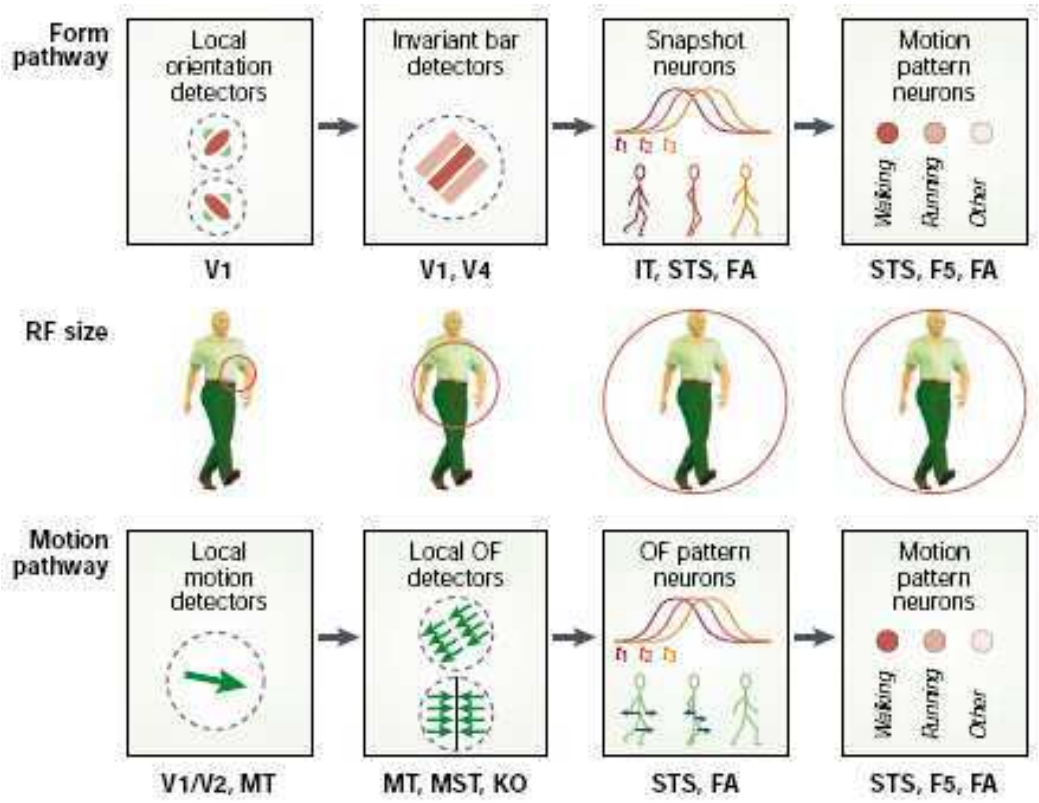

Figure 3.3: The neural model by Giese and Poggio.

and from stimuli with strongly degraded motion information, indicating that the ventral pathway is involved.

Neurophysiological and imaging experiments support the existence of neurons that respond selectively to human body configurations, so biological movements might be recognized by analyzing sequences of body shapes that correspond to "snapshots" from movies of complex movements. The motion is essential in order to identify the dot-pictures. The results of functional Magnetic Resonance Imaging (fMRI) indicate that normal movement stimuly activate areas in both pathways, whereas point-light stimuli tend not to activate form-selective areas.

3.2.3 The neural model

The neural model by Giese and Poggio is based on four assumptions which are consistent with established anatomical and physiological facts (fig. 3.3):

- the model is divided into two parallel processing streams, analogous to

the ventral and dorsal streams, that are specialized for the analysis of form and optic flow information, respectively;

- both pathways comprise hierarchies of neural features detectors that extract form or optic flow features with increasing complexity along the hierarchy. The position and size invariance of the feature detectors also increases along the hierarchy;

- the model assumes that the hierarchy in both pathways is predominantly feedforward (apart from local feedback loops), without the need of top-down signals. Although such signals might be important, in particular for longer stimulus presentations, good recognition performance can be achieved in most cases also without them. Recordings in the STS have found short latencies for the recognition of biological movements. None of these facts rules out the use of feedback processing. However, it indicates a hierarchical feedforward architecture as the core circuitry, underlying immediate recognition, that might be modulated by recurrent loops and higher-level interactions over longer time intervals;

- the representation of motion is based on a set of learned patterns. These patterns are encoded as sequences of snapshots of body shapes by neurons in the form pathway, and by sequences of complex optic flow patterns in the motion pathway. This assumption is a central postulate of the model.

The model by Giese and Poggio extends a previous model meant for the recognition of stationary objects [49] by integrating form information over time in the ventral pathway, and by adding the dorsal pathway.

The two pathways are kept separate. This is a semplification: in monkey and human brain the two processing streams interact at several levels.

Both pathways consist of a hierarchy of neural feature detectors. In addition they contain neural circuits that make recognition sequence-selective.

The form pathway (ventral)

The form pathway analyze biological movements by recognizing sequences of snapshots of body shapes. Several neurophysiologically plausible models for the recognition of stationary form have been proposed. The form pathway of the model extends a model for object recognition that consists of a hierarchy of neural detectors that process form features of increasing complexity.

These detectors correspond to different classes of neurons in the ventral visual pathway. Consistent with neurophysiological data, the receptive field sizes and the position and scale invariance of the neural detectors increase along the hierarchy.

The first level of the form pathway consists of local orientation detectors that model simple cells in the primary visual cortex (V1). Consistent with other models of simple cells, these detectors are modeled as Gabor-like filters. The model contains orientation detectors for eight preferred orientations, and two spatial scales that differ by a factor 2. The sizes of the receptive fields are in the range of those of neurons in monkey V1.

The next level of the form pathway contains position-invariant and scale-invariant bar detectors, which extract local orientation information. Within a limited range, their responses are independent of the spatial position and scale of contours within their receptive fields. They might correspond to complex-like cells in area V1, or to neurons that are increasingly invariant to position changes in areas V2 and V4. The receptive field size of the invariant bar detectors is typical of neurons in area V4. Many neurons in areas V2 and V4 are selective for more complex form features that are similar to corners or junctions. Such features were not necessary to achieve sufficient selectivity of the model for motion recognition. A neurophysiologically plausible mechanism for achieving position and scale invariance is the pooling of the responses of neurons with similar preferred orientations, but with different receptive field positions and spatial scales. It is assumed that this pooling is accomplished by a nonlinear maximum-like operation rather than by linear summation. Subpopulations of complex cells in the visual cortex of cats and neurons in area V4 of macaques show behaviour that is compatible with a maximum computation.

The next level of the form pathway contains snapshot neurons that are selective, for instance, for body shapes. These model neurons are similar to view-tuned neurons in monkey inferotemporal cortex (area IT), which are selective for complex shapes and can become tuned to complex shapes through learning. Like view-tuned neurons in area IT, the snapshot neurons have large receptive fields ($> {}^\circ 8$) and show substantial position-invariance and scale-invariance. As in previous models of view-tuned neurons, the snapshot neurons are modeled by gaussian radial basis function. These neurons receive inputs from the invariant bar detectors on the previous hierarchy level. The centers of the basis functions are adjusted during training so that each snapshot neuron encodes one key frame from a training sequence. In the simulations, each movement pattern is encoded by 21 snapshot neurons representing regularly sampled key frames (this number is not crucial for the performance of the model). The model does not address how an optimum

set of key frames can be learned automatically. Neurons with similar properties in the cortex might be located in area IT in monkeys, and in the STS of monkeys and humans. Activity that is selective specifically for human body shapes has been found in the human lateral occipital complex, occipital and fusiform face areas and monkey STS.

The highest hierarchy level of the form pathway consists of motion pattern neurons. These model neurons temporally smooth and summate the activity of all snapshot neurons that contribute to the encoding of the same movement pattern. Each motion pattern neuron encodes a single action, such as walking or fighting or other actions. Each snapshot neuron has asymmetric lateral connections that pre-activate snapshot neurons that encode subsequent body configurations. The other snapshot neurons are inhibited. The output signals of all snapshot neurons that are involved in the encoding of the same motion pattern are summed in a motion pattern neuron. In this way sequence selectivity is obtained. According to physiological data, motion pattern neurons in monkey and human cortex are probably located in the STS, the premotor cortex (area F5) and possibly the fusiform and occipital face areas.

The motion pathway (dorsal)

The motion pathway recognizes biological movements by analysing optic flow patterns. Consistent with neurophysiological data, it consists of a hierarchy of neural detectors for optic flow features of increasing complexity. As in the form pathway, the receptive field sizes, invariance of the detectors and complexity of the extracted features increase along the hierarchy.

The first level of the motion pathway consists of local motion detectors that correspond to direction-selective neurons in V1 and component motion-selective neurons in area MT. Many neurophysiologically plausible models for local motion estimation have been proposed. For the simulations reported in the Giese and Poggio paper, optic flow patterns have been computed and the responses of motion-sensitive neurons with physiologically realistic parameters have been calculated. Their equivalent receptive field sizes are in the range of direction-selective neurons in V1, and of foveal neurons in area MT.

The second level of the motion pathway consists of neurons with larger receptive fields that analyze the local structure of the optic flow fields induced by movement stimuli. There are two types of local optic flow detector. The first is selective for translation flow and corresponds to motion pattern neurons in area MT. The model includes neuron populations with four preferred directions and with a receptive field size similar to monkey MT neurons.

The second class of local optic flow detectors is selective for expansion and contraction flow. Neurons with such opponent motion selectivity have been found in several areas in the dorsal processing stream, including areas MT, MST, and MSTl. They are probably also present in the kinetic occipital area (KO) in humans. Neurons in area MST have substantial position and scale invariance. In the model, such position invariance is obtained by pooling the signals from position-specific motion edge detectors through a maximum operation. The receptive field size of the motion edge-selective detectors is in the range of neurons in areas MT and MSTl in the macaque monkey.

The optic flow pattern neurons on the third level of the motion pathway are equivalent to the snapshot neurons in the form pathway. Their existence is a prediction of the model. These detectors are selective for complex optic flow patterns that arise for individual moments of biological movement patterns. Like the snapshot neurons, the motion pattern neurons are modelled by gaussian radial basis functions that receive their inputs from the previous hierarchy level. After training, the centres of the basis functions correspond to the optic flow patterns that are characteristic for individual moments of the learned movement. It is assumed that such optic flow pattern neurons might be found at different locations in the visual cortex, in particular in the STS, fusiform and occipital face areas, and, perhaps area MST.

The output signals of the optic flow pattern neurons are summed and temporally smoothed by the motion pattern neurons of the motion pathway. They are modelled in the same way as the motion pattern neurons of the form pathway. Alternatively, a single set of motion pattern neurons might integrate the information from both pathways. Motion pattern neurons are probably located in the STS, fusiform and occipital face areas, and, perhaps area F5 in the premotor cortex.

Sequence selectivity

Movement recognition is selective for temporal order. In the model, sequence selectivity results from asymmetric lateral connections between the snapshot neurons in the form pathway (and between the optic flow pattern neurons in the motion pathway). With this circuitry, active snapshot neurons pre-exite neurons that encode temporally subsequent configurations, and inhibit neurons that encode other configurations. Significant activity can arise only when the individual snapshot neurons are activated in the correct temporal order.

Asymmetric lateral connections are one physiologically plausible implementation of sequence selectivity. The neural activity is strongly reduced if the order of the input frames is reversed or randomized with respect to the

training sequence.

3.3 The model by Demiris and Johnson

In this section a cognitive architecture will be presented that allows the imitation and learning of actions, and describe its computational implementation on two robots. Following earlier work, the architecture employs a generative, or motor-based simulation approach to imitate actions: we understand other people's actions by mentally rehearsing them. These approaches have been advocated at a more theoretical level for several years. The algorithmic and representational requirements for such approaches through a robotic implementation using distributed, hierarchical structures will be examined. Moreover it will be demonstrated how new composite action representations can be learned by imitation.

3.3.1 Different approaches to understanding actions

Fig. 3.4 shows the classical approach to understanding actions: the imitator observes the demonstrated action, which is represented in some symbolic manner, classifies it and subsequently sends it to the motor system for execution. This approach, although highly successful in the robotics domain, is not compatible with a variety of experimental data from the life sciences. The most important difference is that the decomposition above enforces a strong decoupling of the perceptual and motor areas, and does not hypothesize any involvement of the motor systems during observation. On the contrary, as we said previously, Positron Emission Tomography (PET) and fMRI data have shown strong activation of the motor areas of the human brain during the observation of actions, as well as during their execution. The architecture that is used in these experiments attempts to overcome these discrepancies by involving the motor systems of the observer during the demonstration process. The observer uses its motor systems to generate hypotheses during the demonstration (what could I do if I was in that situation/configuration?), and ranks the hypotheses according to the accuracy of the predictions they generate as the demonstrated action unfolds. Action planning, imagination, execution and perception share a generative computational substrate.

3.3.2 Architecture of the model

In the described model the autors adapt a hierachical, distributed action representation approach, that emphasizes concurrent execution of several

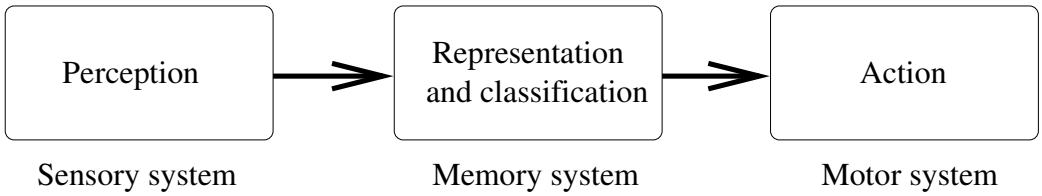

Figure 3.4: Classical approach to action imitation.

modules, rather than a monolithic classifier of the type depicted in fig. 3.4. Here follows a description of the components of the architecture, and how they are combined to perform the functions this architecture supports.

The essence of the biological data described earlier, i.e. that the motor systems are involved during the demonstration of an action, has been captured through the use of inverse and forward models:

- inverse models are also known as controllers, or behaviours. Given a goal and the current state of the controlled system, they output the necessary motor commands that are needed in order to achieve or maintain that goal. These models are used frequently in control engineering and have been used for modeling motor planning and control. Inverse models have been hypothesized to exist in the premotor cortex (F5 mirror);

- forward models are also known as predictors. Given motor commands and the current state of the controlled system, they output the predicted next state of that system. Like inverse models, they have been used in motor control modeling, and they have been hypothesized to exist in the cerebellum.

Fig. 3.5 shows a typical arrangement of an inverse and a forward model, as used in motor control. With this arrangement, an inverse model, encoding a particular behaviour, is given the current state and a target goal. The inverse model outputs the motor commands that would achieve that target goal, and these are sent to the musculoskeletal system, which in turn, feeds back proprioceptive information. However, this feedback takes a significant amount of time before it is available, so, in parallel to this, the motor commands are also sent to a forward model which will output a prediction of what the next state will be. This predicted state is compared with the target goal, and errors are used to adjust its confidence on how well it is performing. Subsequently, parallelism is introduced to this arrangement. Multiple pairs of inverse and forward models, encoding different behaviours, are sent

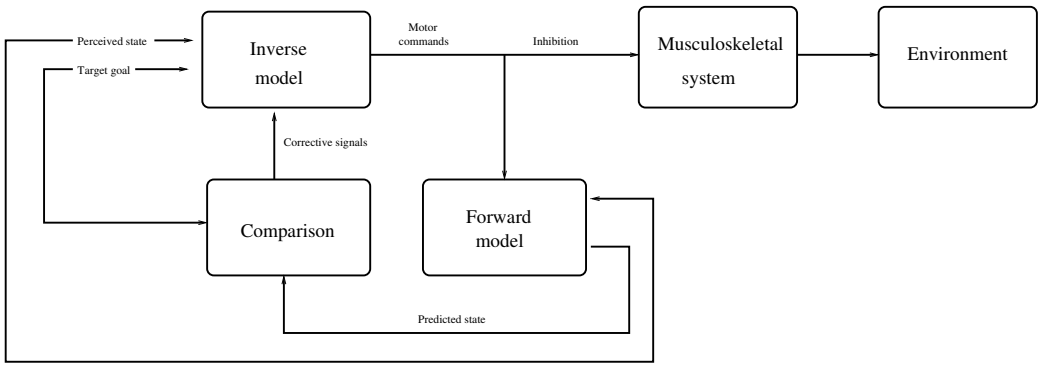

Figure 3.5: Basic structure of the model: a pair of inverse and forward models.

to the target goal. They all propose motor commands that they hypothesize will achieve the target goal. These motor commands are sent to the forward models (each inverse model sends its motor commands to its corresponding forward model) which output a set of predicted next states. The pair that has generated the next state prediction most compatible with the target goal receives reinforcement (increases its confidence), while the other pairs decrease their confidences. The errors can also be used so that the behaviours can adapt their internal parameters (if any) so they can suggest better hypotheses. For example, an error in the prediction of the gripper's next position will decrease the confidence of the corresponding behaviour, but it will also alter the parameters of the behaviour; for example it will alter the gains in a Proportional Integrative Derivative (PID) controller that moves the gripper, so that the predictions can potentially be improved.

3.3.3 Implementation of the model

The implementation of the model uses two robots in an experimental scenario involving imitation of gripper movements. In order to simplify the visual processing, color markers have been added to the grippers and robots are placed facing each other. They are equipped with a camera and a two degrees of freedom gripper (fig. 3.6). The imitator robot "mentally reharses" and ranks hypotheses during the demonstration, and subsequentally replicates the one with the highest confidence. If it does not recognize the behaviour but recognizes its components, it assembles a composite inverse model and adds a new pair of inverse-forward models to its repertoire.

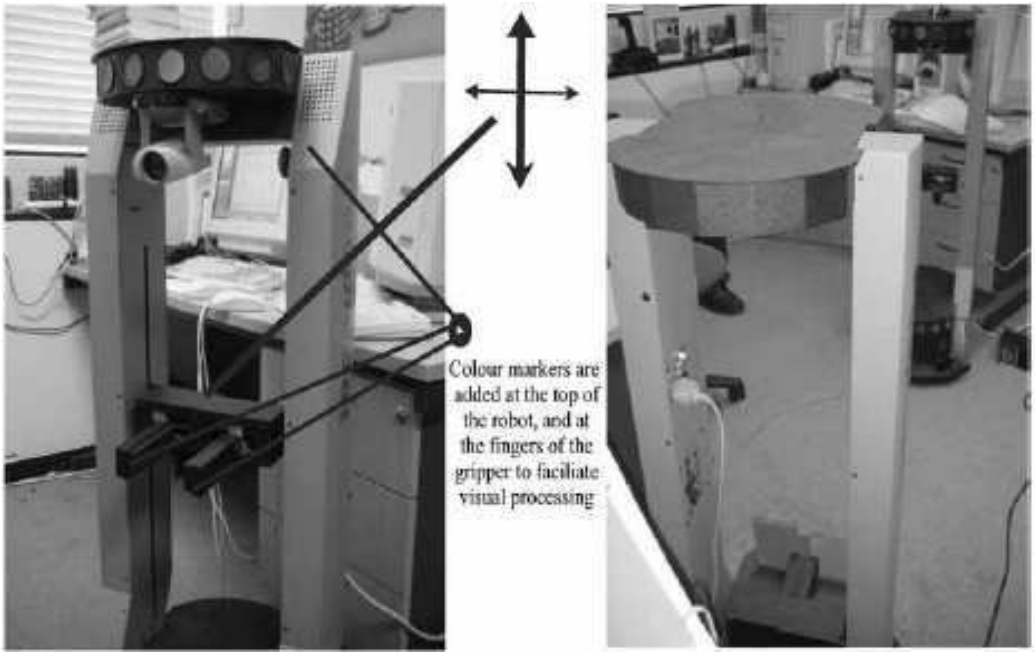

Figure 3.6: Left: imitator robots; right: robotic setup

Implementation of inverse models

The most basic inverse model, i.e. a primitive, is a single node implemented as a simple motor program, which sends out motor commands to the robots and to the forward model. Composite inverse models can be created, by having more basic ones arranged serially and/or concurrently in a graph (fig. 3.7) and indicating the start time of each of them.

Implementation of forward models

For the construction of forward models the speed of movement for each degree of freedom was calculated by having the corresponding robot part (i.e. the robot gripper) executing a full movement between the two limits (top-bottom for the gripper platform, and open-close for the gripper fingers), and recording proprioceptive values and timer values at the beginning and at the end of the movement (the gripper is only capable of moving with a single constant velocity for each degree of freedom). Proprioceptive values are scaled between the values 0.0 and 1.0. The forward models can then output predicted proprioceptive values at the next time step (100 ms intervals are used) by using the known velocity, the direction of movement and the current

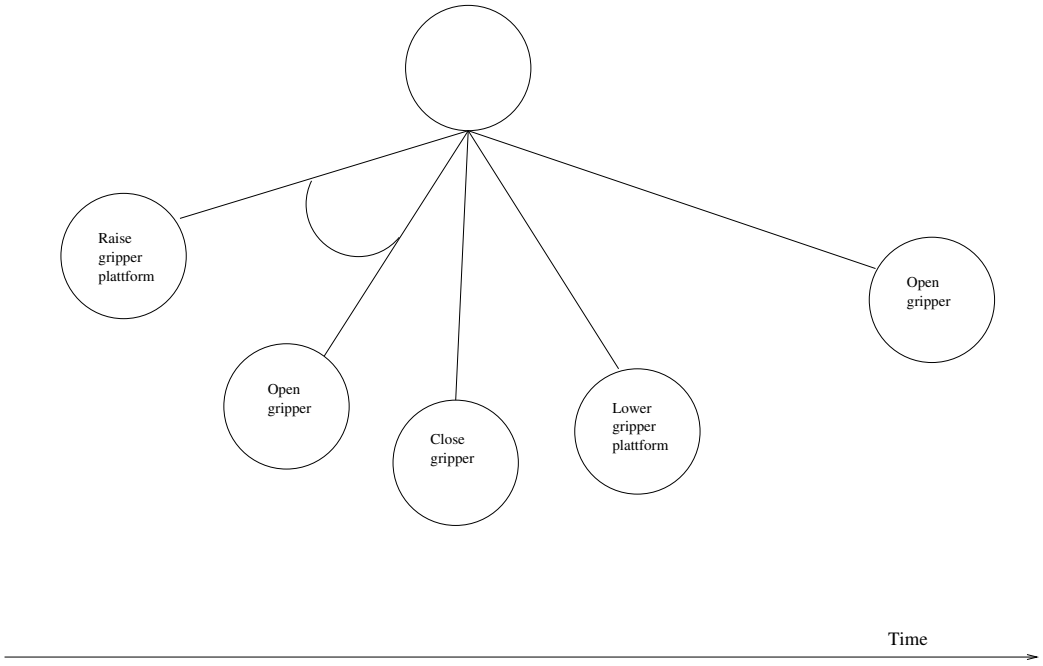

Figure 3.7: An example inverse model.

recorded proprioceptive value.

Implementation of the framework

For the architecture to work on an imitator robot it also requires the extraction of the visual information of the state of the demonstrator, and of the imitator's proprioceptive state. The two quantities extracted are the height of robot gripper from base and the gripper finger positions. A visual calibration preceeds the experiment; the demonstrator robot executes a sequence involving moving the gripper from the lowest possible position to the highest possible position, and subsequently opening and closing the gripper fingers completely. The visual coordinates during the calibration are collected and normalised to give positional information in the range of 0.0 to 1.0.

The inverse models are executed as parallel processes in the robots. During observation, the motor commands they generate are inhibited from being sent to the motor system, and they are only sent to the forward models. Each of the forward models output predictions for the state of each of the degrees of freedom that are involved in this movement, which are compared to the actual values that come at the next time step. The prediction error is

given by:

$$E(t) = \sum_{i=1}^{N} |x_i(t) - px_i(t)| \tag{3.1}$$

where $x_i(t)$ is the actual value of the demonstrators state at time t, and $px_i(t)$ is the predicted value that was given by the forward model for time t. N is the number of degrees of freedom involved in this behaviour. The confidence value of the inverse model is then accumulated according to the following update rule:

$$C(t) = \begin{cases} C(t-1) + Kr\frac{1}{E(t)} & if E(t) < A \\ C(t-1) - KpE(t) & otherwise \end{cases}$$

where $C(t)$ is the confidence of the inverse model at time t, A is a constant threshold value which is set experimentally, and Kr and Kp are gain constants. The value A is essential as a threshold describing the boundary between the reward and punishment regions.

Having presented two models of action recognition let's now introduce the relation between pragmatic and semantic recognition.

3.4 The relation between pragmatic and semantic recognition

In humans, several experiments have investigated the interplay between action generation and action perception [50].

Fadiga et al. stimulated the motor cortex of human observers and recorded the MEP (Motor evoked potentials) from hand muscles, utilising the assumption that if action observation activates the premotor cortex (as it does in monkeys), this activation should induce an increase of the motor evoked potentials elicited by the magnetic stimulation of the motor cortex. They found a significant increase of the MEP when subjects observed movements, and additionally the pattern of muscle activation was very similar to the pattern of muscle contraction present during the execution of the same action, i.e. the increase was present only in those muscles that are active when the human subjects executed these actions. During the rehersing of the motor commands when the observer sees an action the muscle contraction is inhibited but this mechanism is not well known. A different set of experiments with human subjects used PET brain scanning as a way of mapping the brain regions whose activations are associated with the observation of hand actions, as well as mental rehersal. Concerning these studies there is a very interesting brain disorder: the "imitation behaviour". The patients affected

by this disorder imitate the demonstrator's gesture although they were not instructed to do so, and some times even when told not to do so. An explicit direct command from the doctor to the patient would stop the imitation behaviour but a simple distraction was sufficient to see imitation reappearing, despite the patient remembering what he had been told.

On the base of the above we surmise that action recognition is a two steps process:

- pragmatic recognition: the action is mentally rehersed, i.e. motor commands to perform it are extracted from visual information;
- semantic recognition: the action is classified analyzing the motor commands.

The model of action recognition we will propose is based on the assumptions indicated above.

3.5 Beyond action recognition: grasping the intention [51]

Here we present some ideas about intentions as elaborated in the paper by Iacoboni. Understanding the intentions of others while watching their actions is a fundamental building block of social behavior. The neural and functional mechanisms underlying this ability are still poorly understood. To investigate these mechanisms Iacoboni et al. used fMRI analysis. Twentythree subjects watched three kinds of stimuli: grasping hand actions without a context, context only (scenes containing objects), and grasping hand actions performed in two different contexts. In the latter condition the context suggested the intention associated with the grasping action (either drinking or cleaning). Actions embedded in contexts, compared with the other two conditions, yielded a significant signal increase in the posterior part of the inferior frontal gyrus and the adjacent sector of the ventral premotor cortex, where hand actions are represented. Thus, premotor mirror neuron area, previously thought to be involved only in action recognition, is actually also involved in understanding the intentions of others. To ascribe an intention is to infer a forthcoming new goal, and this is an operation that the motor system does automatically.

Some mirror neurons do not discriminate between stimuli of the same category (i.e., the sight of different kinds of grasping actions can activate the same neuron), but discriminate well between actions belonging to dif-

ferent categories. These properties seem to indicate an action recognition mechanism ("that's a grasp") rather than an intention-coding mechanism.

However, action implies a goal and an agent. Consequently, action recognition implies the recognition of a goal, and, from another perspective, the understanding of the agent's intentions.

More complex and interesting, so, is the problem of whether the mirror neuron system also plays a role in coding the global intention of the actor performing a given motor act. For example: Mary is grasping an apple. Why is she grasping it? Does she want to eat it, or give it to her brother, or maybe throw it away?

The same action done in two different contexts acquires different meanings and may reflect two different intentions.

In the experiment of Iacoboni et al. subjects watched three different types of movie-clips (fig. 3.8): CONTEXT, ACTION, and INTENTION, interspersed with periods of blank screen (rest condition). The CONTEXT condition consisted of two scenes with 3D objects (a teapot, a mug, cookies, a jar, etc.). The objects were arranged either as just before having had tea (the "drinking" context) or as just after having tea (the "cleaning context"). The ACTION condition consisted of a hand grasping a cup in the absence of a context on an objectless background. Two types of grasping actions were shown in the same block an equal number of times: a precision grip (the fingers grasping the cup handle) and a whole-hand prehension (the hand grasping the cup body). In the INTENTION condition, the grasping actions (also precision grip and whole hand prehension shown for an equal number of times) were embedded in the two scenes used in the CONTEXT condition, the "drinking" context and the "cleaning" contex. Here, the context cued the intention behind the action. The "drinking" context suggested that the hand was grasping the cup to drink. The "cleaning" context suggested that the hand was grasping the cup to clean up. Thus, the INTENTION condition contained information that allowed the understanding of intention, whereas the ACTION and CONTEXT conditions did not (i.e. the ACTION condition was ambiguous, and the CONTEXT condition did not contain any action).

Notably, the observation of the INTENTION and of the ACTION clips compared to those at rest yielded significant signal increase in the parieto-frontal cortical circuit for grasping.

The critical question was to test whether there are significant differences between the INTENTION condition and the ACTION and CONTEXT conditions in areas known to have mirror properties in the human brain. The INTENTION condition yielded significant signal increases compared to the ACTION condition in visual areas and in the right inferior frontal cortex, in the dorsal part of the pars opercularis of the inferior frontal gyrus. The

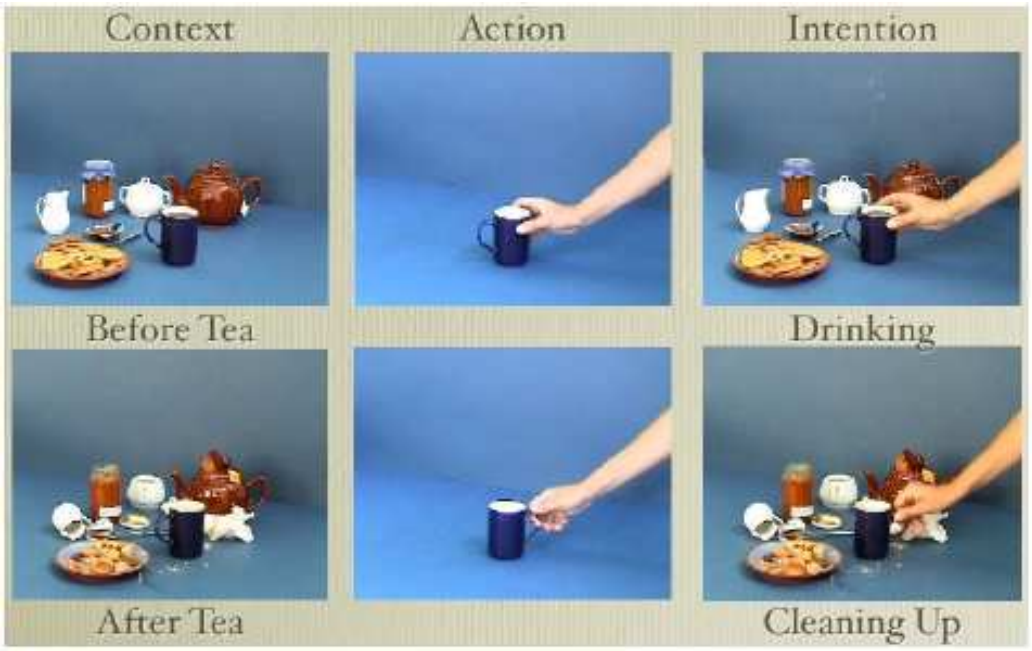

Figure 3.8: Six images taken from the context, action and intention clips.

increased activity in visual areas is expected, given the presence of objects in the INTENTION condition, but not in the ACTION condition. The increased right inferior frontal activity is located in a frontal area known to have mirror neuron properties, thus suggesting that this cortical area does not simply provide an action recognition mechanism ("that's a grasp") but rather it is critical for understanding the intentions behind others' actions.

The data of the study suggest that the role of the mirror neuron system in coding actions is more complex than previously shown and extends from action recognition to the coding of intentions.

The findings of the presented study show increased activity of the right inferior frontal cortex for the INTENTION condition, so strongly suggest that this mirror neuron area actively participates in understanding the intentions behind the observed actions.

Before accepting this conclusion, however, there are some points that must be clarified:

- as a first issue, one might argue that the signal increase observed in the inferior frontal cortex was simply due to detecting an action in any context; that is, it is the complexity of observing an action embedded in a scene, and not the coding of the intention behind actions, that

determined the signal increase;

- a second issue, closely related to the first one, is the question of canonical neurons. These neurons fire at the sight of graspable objects. Because they are also located in the inferior frontal cortex, one might be led to conclude that the increased activity observed in the intention clips was due to the presence of objects.

A strong argument against both these objections is that the activity in inferior frontal cortex is reliably different between drinking INTENTION clips and cleaning INTENTION clips, even though graspable objects were present in both conditions.

On the basis of our current knowledge of physiological properties of the inferior frontal cortex, the most parsimonious explanation of the findings reported here is that mirror neurons are the likely neurons driving the signal changes in this study.

The interpretation of these findings implies that, in addition to the classically described mirror neurons that fire during the execution and observation of the same motor act (i.e. observed and executed grasping), there are neurons that are visually triggered by a given motor act (i.e. grasping observation), but discharge during the execution not of the same motor act, but of another act, causally related to the observed act (i.e. bringing to the mouth). Neurons of this type have indeed been previously reported in F5 and referred to as "logically related" neurons. The role of these "logically related" mirror neurons was never theoretically discussed and their functions remained unclear. The work presented here allows one to attribute a functional role to these "logically related" mirror neurons and suggest that they may be part of a chain of neurons coding the intentions of other people's actions.

The stronger activation of the inferior frontal cortex in "drinking" as compared to "cleaning" INTENTION is consistent with our interpretation that a specific chain of neurons coding a probable sequence of motor acts underlies the coding of intention.

The conventional view on intention understanding is that the description of an action and the interpretation of the reason why that action is executed rely on largely different mechanisms. In contrast, the presented data show that the intentions behind the actions of others can be recognized by the motor system using a mirror mechanism. Mirror neurons are thought to recognize the actions of others, by matching the observed action onto its motor counterpart coded by the same neurons. The present findings strongly suggest that coding the intention associated with the actions of others is

based on the activation of a neuronal chain formed by mirror neurons coding the observed motor act and by "logically related" mirror neurons coding the motor acts that most likely follow the observed one, in a given context. To ascribe an intention is to infer a forthcoming new goal, and this is an operation that the motor system does automatically.

Chapter 4

The precursor of ROMOACRE

In this chapter we describe the precursor of ROMOACRE, a **RO**botic **MO**del for **AC**tion **RE**cognition. In this chapter we will refer to the precursor of ROMOACRE simply as "ROMOACRE". Instead the final version of ROMOACRE will be described in the next chapter. ROMOACRE is a neurophysiologically plausible model whose purpose is the recognition of significant human actions, like walking, running, and so on. It is neurophysiologically plausible in the sense that each stage of computation is associated with a cortical area, and in particular it reflects the mirror mechanism, using motor commands to recognise actions.
ROMOACRE is composed of the following computational stages:

1. human pose estimation from images of body silhouette;

2. evaluation of the motor commands of the action from human poses sequence;

3. recognition of the action from motor commands.

The three stages are described in detail in the following sections of this chapter. The details of image capture and preprocessing are described in the appendix A.

4.1 Human pose estimation

We define a pose as a particular configuration of body joints angles. As we will describe later in this chapter, our body model is made up of 8 monodimensional body joints angles. Let's review first the previous works about human pose estimation.

4.1.1 Previous works

In literature we have found various methods for human pose estimation. We divide them in four groups:

- the pose estimation is done from a sequence of images captured by multiple cameras;
- the pose estimation is done from single images captured by multiple cameras;
- the pose estimation is done from a sequence of monocular images;
- the pose estimation is done from a single monocular image.

The following work belongs to the first group.

- [4] A 3D reconstruction of the person's body is computed from silhouettes extracted from four cameras. In the first frame, body parts are located sequentially. The head is located first, since its shape and size are unique and stable. Other parts are found by sequential template growing and fitting. This initial estimate of body part locations, sizes and orientations is then used as a measurement for the extended Kalman filter which ensures a valid articulated body model. The same filter, with a slightly modified state and state transition matrix, is then used for tracking.

The following work belongs to the second group.

- [5] An approach for estimating 3D body pose from multiple, uncalibrated views is proposed. First, a mapping from image features to 2D body joint locations is computed using a statistical framework that yields a set of several body pose hypotheses. The recovery of 3D body pose and camera relative orientations is formulated as a stochastic optimization problem.

The following works belongs to the third group.

- [6] A probabilistic method for tracking 3D articulated human figures in monocular image sequences is presented. Within a Bayesian framework, a generative model of image appearence, a robust likelihood function based on image graylevel differences, and a prior probability distribution over pose and joint angles that models how humans move are defined. The posterior probability distribution over model parameters is represented using a discrete set of samples and is propagated over time.

- [7] A novel solution is presented that directly addresses the depth ambiguity, in which a discriminative analysis (Support Vector Machine) is extended to non-rigid human motion classification with a temporal generative motion model (Hidden Markov Model).

- [8] This work addresses the problem of tracking human body pose in monocular video including automatic pose initialization and re-initialization after tracking failures caused by partial occlusion or unreliable observation. A method is proposed which is based on data-driven Markov chain Monte Carlo that uses bottom-up techniques to generate state proposals for pose estimation and initialization.

The following works belongs to the fourth group.

- [9] In this work the authors propose a statistical formulation for 2D human pose estimation from single images. The human body configuration is modeled by a Markov network and the estimation problem is to infer pose parameters from image cues such as appearence, shape, edge, and color. From a set of hand labeled images, we accumulate prior knowledge of 2D body shapes. A data driven belief propagation Monte Carlo algorithm is proposed for efficent probabilistic inference.

- [10] This work discusses a bottom-up approach that uses local image features to estimate human upper body pose from single images in cluttered backgrounds. The method takes the image window with a dense grid of local gradient orientation histograms, followed by non negative matrix factorization to learn a set of bases that correspond to local features on the human body, enabling selective encoding of human-like features in the presence of background clutter. Pose is then recovered by direct regression.

- [11] The authors represent the human model in a phase space spanned by its different degrees of freedom and use the analysis-by-synthesis approach to match the phase space model with real images and thereby estimating the pose.

- [12] The authors describe a method based on learning for recovering 3D human body pose from single images. The approach recovers pose by direct nonlinear regression against shape descriptor vectors extracted automatically from image silhouettes. They evaluate several different regression methods: ridge regression, Relevance Vector Machine regression and Support Vector Machine regression over both linear and kernel bases. The Relevance Vector Machine regression provides much sparser

regressors without compromising performance, and kernel bases give a small but worthwhile improvement in performance.

- [13] The problem considered in this work is to take a single two-dimensional image containing a human body, locate the joint positions, and use them to estimate the body configuration and pose in three-dimensional space. The basic approach is to store a number of exemplar 2D views of the human body in a variety of different configurations and viewpoints with respect to the camera. On each of these stored views, the locations of the body joints are manually marked and labelled for future use. The test shape is then matched to each stored view, using the technique of shape context matching. Assuming that there is a stored view sufficiently similar in configuration and pose, the correspondence process will succeed. The locations of the body joints are then transferred from the exemplar view to the test shape. Given the joint locations, the 3D body configuration and pose are then estimated.

- [14] In this work the autors introduce a new algorithm that learns a set of hashing functions that efficiently index examples in a way relevant to a particular estimation task.

We tried to use some of the method described above and we chose the best performing method, k-Nearest Neighbors Weighted (k-NNW) method.

4.1.2 The k-Nearest Neighbors Weighted method [14]

In this subsection we describe the method used by ROMOACRE to perform human pose estimation, the k-Nearest Neighbors Weighted (k-NNW) method.

The task of example-based parameter estimation in vision can be formulated as follows. Input, which consists of image features (e.g. edge map, vector of responses of a filter set, edge direction histograms or vector of responses of snapshot units) computed on the original image, is assumed to be generated by an unknown parametric process $x = f(\theta)$ (e.g., θ is a vector of joint angles which represent the pose). A training set of labeled examples $(x_1, \theta_1), ..., (x_n, \theta_n)$ is provided. One must estimate θ_0 as the inverse of f for a novel input x_0. The objective is to minimize the residual in terms of the distance (similarity measure) d_θ in the parameter space.
One of the oldest techniques used for such estimation is the k-Nearest Neighbors (k-NN) method: the k-NN estimate is obtained by averaging the values

for the k training examples most similar to the input:

$$\hat{\theta}_{NN} = \frac{1}{k} \sum_{x_i \in neighborhood} \theta_i, \tag{4.1}$$

i.e. the target function is approximated by a constant in each neighborhood defined by k. This estimate is known to be consistent, and to asymptotically achieve Bayes-optimal risk under many loss functions. Note that similarity (neighborhood) is defined in terms of the distance d_x in the input space.
A natural extension to the k-NN method is the k-Nearest Neighbors Weighted (k-NNW) method, in which the neighbors are weighted according to their similarity to the query point. The k-NNW estimate is obtained by weighted average of the values for the k training examples most similar to the input:

$$\hat{\theta}_{NNW} = \sum_{x_i \in neighborhood} w_i \theta_i, \tag{4.2}$$

assuming that the weights w_i are normalized. The next section describes how k-NNW method is applied in ROMOACRE to perform human pose estimation.

4.2 Human pose estimation in ROMOACRE

Human pose estimation in ROMOACRE is performed in two stages:

1. the first stage takes as input the raw data from the image and produces as output a vector of snapshot units selective for body silhouette shapes;

2. the second stage perform human pose estimation with the k-NNW method using as weights the output of the first stage.

The two next subsections describe these two stages of the model.

4.2.1 Pose estimation: stage 1

This part of our model is inspired to the model by Giese and Poggio for the recognition of biological movements [15] and is made up of a single processing stream analogous to the ventral stream that is specialized for the analysis of form, so we call it form pathway. This form pathway comprises hierarchies of neural feature detectors that extract form features with increasing complexity along the hierarchy. The position and size invariance of the feature detectors also incresases along the hierarchy, which is completely feedforward, without

the need of top-down signals. We do not claim that such signals are not important, but without them good recognition performance can be achieved in most cases. The form pathway is composed of three computational levels.

1. The first level of the form pathway consists of local orientation detectors that models simple cells [16] in the primary visual cortex (V1). Consistent with other models of simple cells [17], these detectors are modelled as Gabor filters.

2. The second level of the form pathway contains position and scale tolerant bar detectors, which extract local orientation information. Within a limited range, thir responses are independent of the spatial position and scale of contours within their receptive fields. They might correspond to complex-like cells in area V1 [16], or to neurons that are increasingly invariant to position changes in areas V2 and V4 [18] [19]. Many neurons in areas V2 and V4 are selective for more complex form features that are similar to corners or junctions [18] [20]. Such features were not necessary to achieve sufficent selectivity of the form pathway for body silhouette shapes.

3. The third level of the form pathway contains snapshot neurons that are selective for body silhouette shapes. These model neurons are similar to view-tuned neurons in monkey inferotemporal cortex (IT) which are selective for complex shapes [21] [22] [23] and can become tuned to complex shapes through learning. Neurons with similar property in the cortex might be located in the STS of monkey and humans [24] [25] [26] [27] [28] [29]. Activity that is selective specifically for human body shapes has been found in the human lateral occipital complex [30], occipital and fusiform face areas [25] and monkey STS [31] [32].

In the next three subsections we describe in detail the three levels of the form pathway. The retina of the model is made up of 160x128 pixels and it will be filtered as described below.

Local orientation detectors

The following family of two-dimensional Gabor functions was proposed by Daugman [33] to model the spatial summation properties (of the receptive fields) of simple cells in visual cortex:

$$g_{\sigma,\gamma,\lambda,\phi,x_0,y_0,\theta} = e^{-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}} cos\left(2\pi\frac{x'}{\lambda}+\phi\right) \tag{4.3}$$

$$x' = (x - x_0)cos(\theta) + (y - y_0)sin(\theta) \tag{4.4}$$
$$y' = -(x - x_0)sin(\theta) + (y - y_0)cos(\theta) \tag{4.5}$$

where:

- σ is the standard deviation of the two-dimensional gaussian factor of the Gabor function;
- γ is the ellipticity of the gaussian factor of the Gabor function; values tipical of the receptive fields of simple cells lie between 0.2 and 1.0;
- λ is the wavelength of the cosine factor of the Gabor function;
- ϕ is the phase offset of the cosine factor of the Gabor function: values 0^0 and 180^0 correspond to "centre-on" and "centre-off" functions, while values -90^0 and 90^0 correspond to antisymmetric functions;
- x_0 is the abscissa of the centre of the Gabor function;
- y_0 is the ordinate of the centre of the Gabor function;
- θ is the orientation of the normal to the parallel stripes of the Gabor function.

In fig. 4.1 is plotted a graphic of a Gabor function. In the Gabor functions we use we set

$$\gamma = 1 \tag{4.6}$$
$$\phi = 0 \tag{4.7}$$

so we use "centre-on" Gabor functions with radial gaussian factor, while different phase shifts are crudely approximated by centering receptive fields at near locations [34].
We introduce two batteries of Gabor filters corresponding to two different scales that differ for a factor two; we consider thirtysix orientation for each of the two scales that therefore differ from each other by 5^0. We apply the filters to the images with resolution 160x128 pixels.

- scale 1:
  - $\sigma = 2.4\ pixel$;
  - $\lambda = 12\ pixel$;
  - we found that the length of the main stripe of the Gabor filter at 10% is 11 pixels;

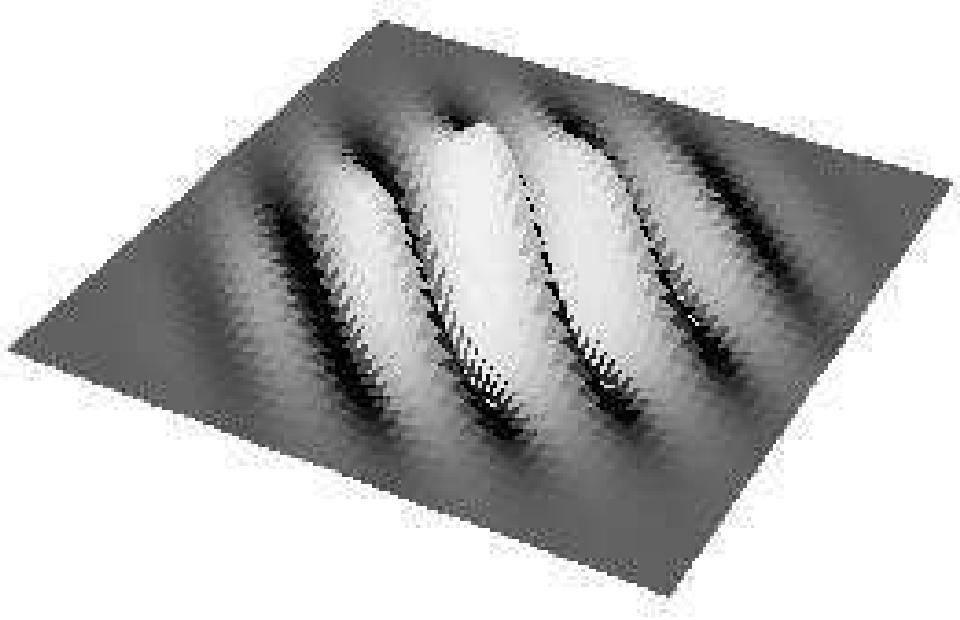

Figure 4.1: Graphic of a Gabor filter.

  - we found that the width of the main stripe of the Gabor filter at 10% is 5 pixels;
  - we apply a squared mask to the filters of this scale of 11x11 pixels;
  - we apply the filters of this scale centered on the points of a grid of step 2 pixels;
  - so we have 80x64=5120 Gabor filters for each of the 36 orientations.

- scale 2:
  - $\sigma = 4.7\ pixel$;
  - $\lambda = 18\ pixel$;
  - we found that the length of the main stripe of the Gabor filter at 10% is 21 pixels;
  - we found that the width of the main stripe of the Gabor filter at 10% is 9 pixels;
  - we apply a squared mask to the filters of this scale of 21x21 pixels;
  - we apply the filters of this scale centered on the points of a grid of step 4 pixels;

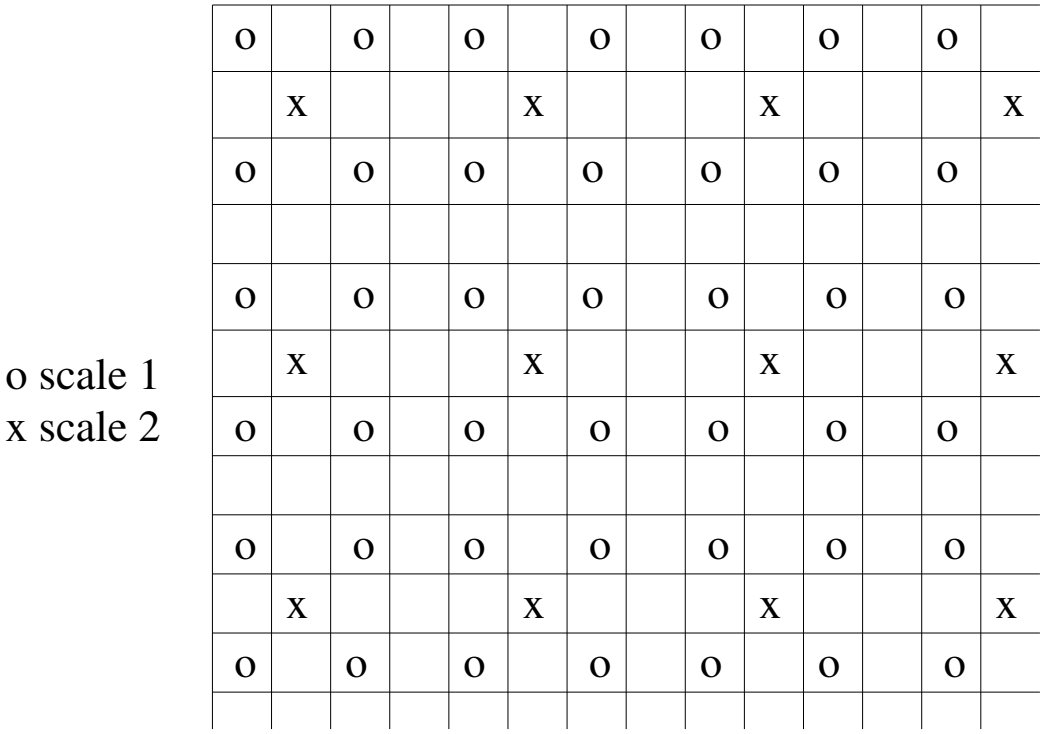

Figure 4.2: Locations of the Gabor filters in the image.

– so we have 40x32=1280 Gabor filters for each of the 36 orientations.

In fig. 4.2 are shown the locations of the Gabor filters in the image.

Tolerant bar detectors

A neurophysiologically plausible mechanism for achieving position and scale tollerance is the pooling of the responses of neurons with similar preferred orientations, but with different receptive field positions and spatial scales [35] [36] [37]. We assume that this pooling is accomplished by a nonlinear maximum operation rather than by linear summation [49] [38]. Subpopulations of complex cells in the visual cortex of cats [39] and neurons in area V4 of macaques [40] show behaviour that is compatible with a maximum computation. We compute the maximum among 4x4=16 Gabor filters of a given orientation of the scale 1 and 2x2=4 Gabor filters of the same orientation of the scale 2 (20 Gabor filters in total) as shown in fig. 4.4. The amplitute of the pooling range is of 8x8 pixels. So we obtain 160/8x128/8=20x16=320

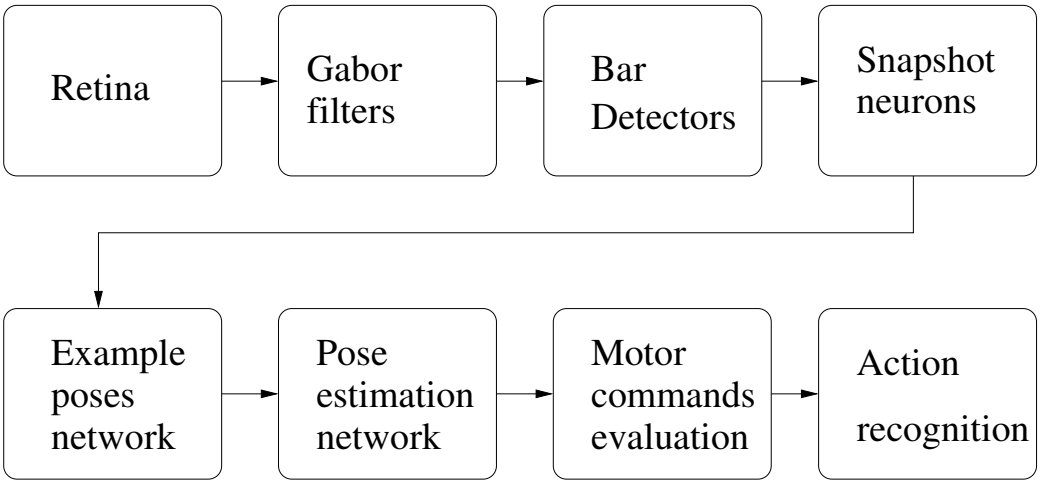

Figure 4.3: Diagram of the model.

tolerant bar detectors for each of the 36 orientations.

In fig. 4.3 is illustrated the diagram of the model.

Snapshot neurons

This level is made up of a neural network with radial gaussian basis function as shown in fig. 4.5:

- the input vector is constituted by the responses of the tolerant bar detectors, so its dimension is 320(location)x36(orientation)=11520;
- the hidden layer is constituted by thirtysix radial gaussian basis functions
$$g(x) = e^{-\frac{x^2}{2\sigma^2}}; \tag{4.8}$$
- the output layer is constituted by thirtysix snapshot neurons that are selective for body silhouette shapes.

The network learning is performed in this way:

- the central vector of the radial gaussian basis functions μ_i $(i = 1, ..., 36)$ of the same dimension of the input space are set equal to the responses of the tolerant bar detectors for the 36 captured frames;
- the standard deviations σ_i are set all equal to the average distance between the centres μ_i;

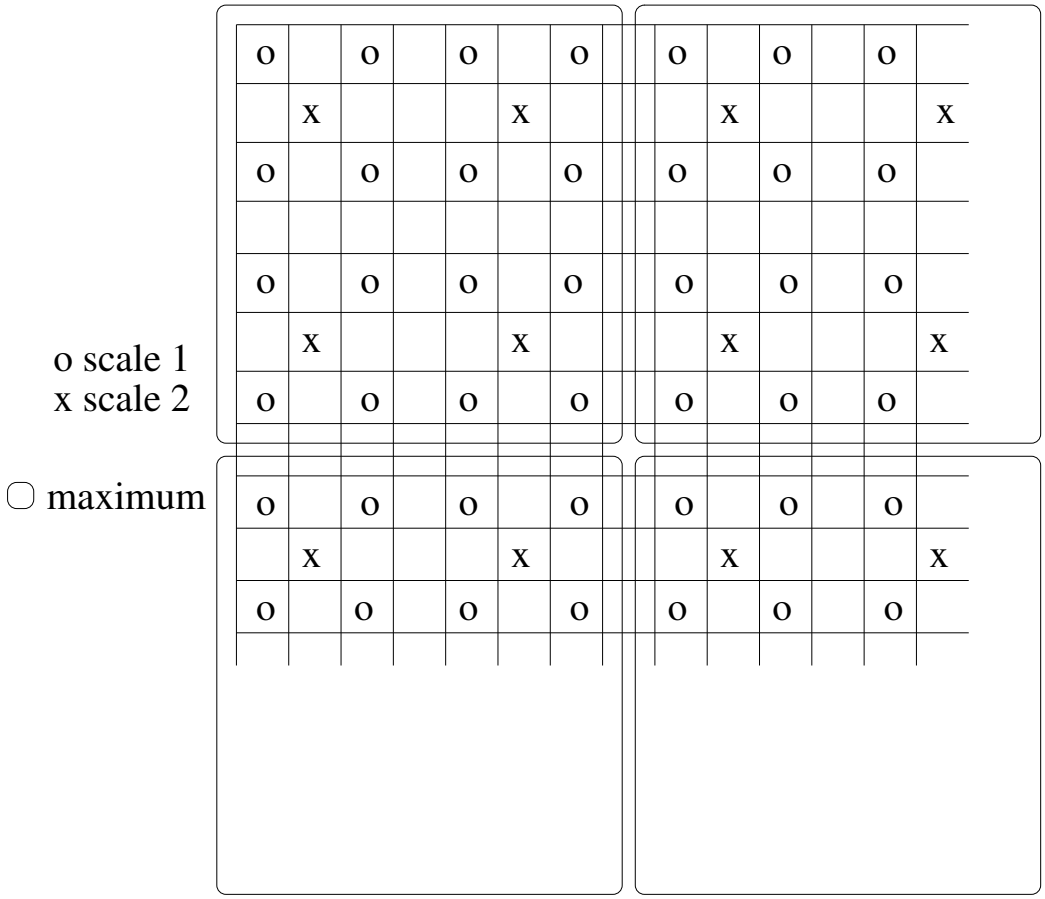

Figure 4.4: Operation of maximum among Gabor filters.

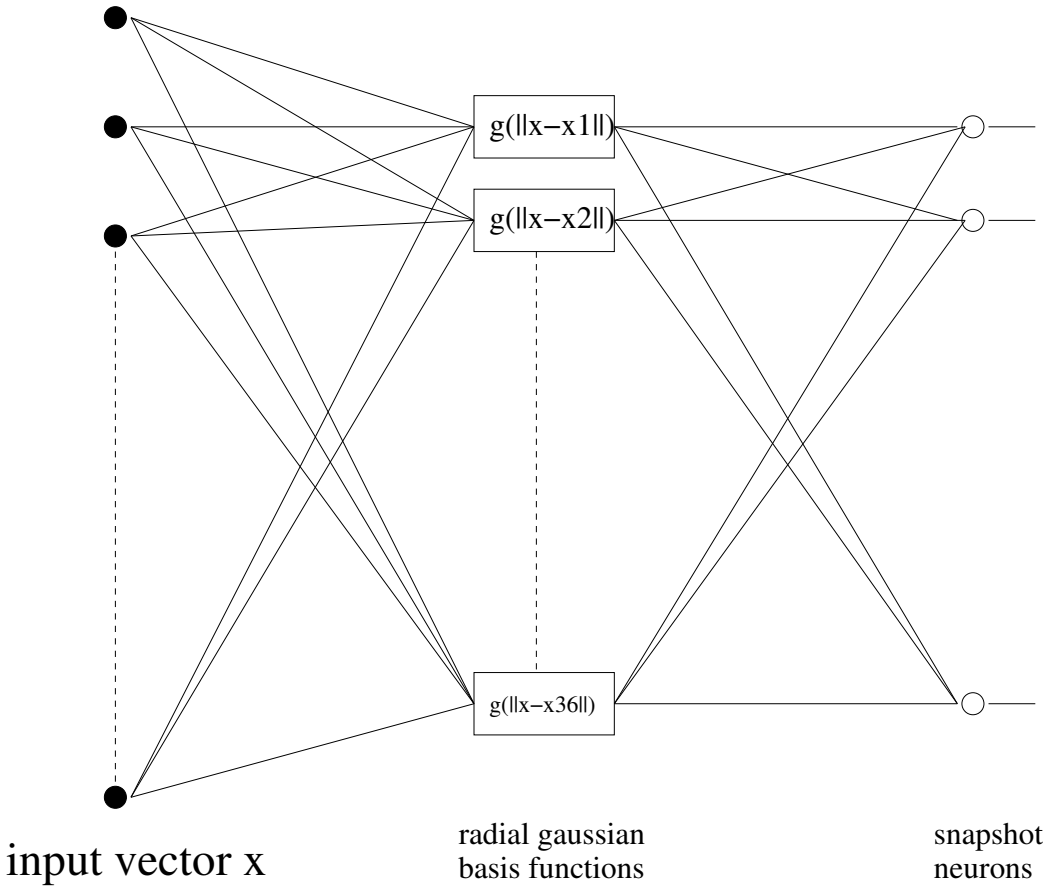

Figure 4.5: Neural network with radial gaussian basis function.

- the weights of the output layer are evaluated with supervised learning using the 36 frames.

4.2.2 Pose estimation: stage 2

Before we illustrate the second part of human pose estimation in ROMOACRE, we list the following assumptions of the model:

- we neglect the angles of the head, of the hands, and of the foot;
- we assume that the back is straight and vertical with respect to the ground;
- the limbs remain parallel to the sagittal plane.

Once we made these assumptions we can assert that the human pose is described by eight angles, that are:

- the angle of the left shoulder, i.e. the angle between the back and the left forearm, positive if the left forearm is ahead, negative if the left forearm is behind;
- the angle of the left elbow, i.e. the angle between the extension of the left forearm and the left arm;
- the angle of the left hip, i.e. the angle between the back and the left thigh, positive if the left thigh is ahead, negative if the left thigh is behind;
- the angle of the left knee, i.e. the angle between the extension of the left thigh and the left leg;
- the angle of the right shoulder, i.e. the angle between the back and the right forearm, positive if the right forearm is ahead, negative if the right forearm is behind;
- the angle of the right elbow, i.e. the angle between the extension of the right forearm and the right arm;
- the angle of the right hip, i.e. the angle between the back and the right thigh, positive if the right thigh is ahead, negative if the right thigh is behind;
- the angle of the right knee, i.e. the angle between the extension of the right thigh and the right leg;

Figure 4.6: Measurement of the shoulder angle.

In the actions we take into account that these eight angles vary in an interval of width 150^0:

- shoulders and hips: $[-75^0, +75^0]$;
- elbows and knees: $[0^0, +150^0]$.

In order to do human pose estimation with an example based method (the k-NNW method) we followed these steps.

- We showed to the subject each of the thirtysix frames and we asked him to assume the same pose he has in the frame, so we measured with an alidade the 8x36=288 angles of the body pose (figg. 4.6, 4.7, 4.8, 4.9).
- We did an angle transformation to transform the angles measured with the alidade in the body angles we have defined above.
- Each of the eight angular variables, of width 150^0, has been divided in 30 bins of width 5^0.
- The index of the occupated bin has been calculated for each of 8(angles)x36(frames)=288 angles.

In this way we produced 36 data sets of dimension 8x30=240 constituted by binary values that value all 0 except the ones that correspond to the angles values.

Figure 4.7: Measure of the elbow angle.

Figure 4.8: Measure of the hip angle.

Figure 4.9: Measure of the knee angle.

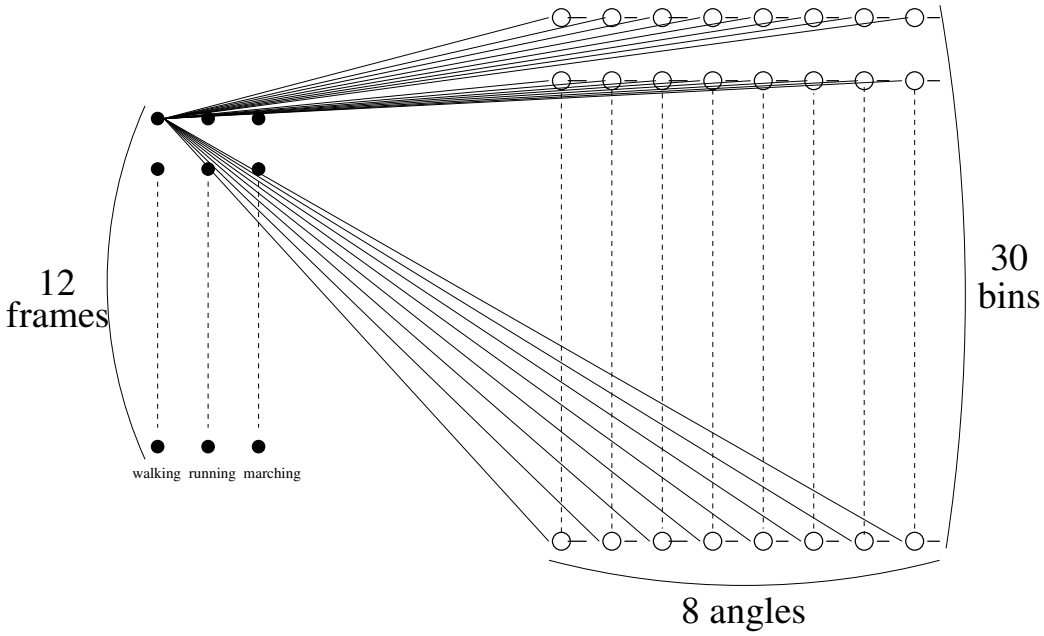

Figure 4.10: Example poses network.

Figure 4.11: First frame of the walking action.

In order to establish the correspondence between the 36 frames and the measured example poses, we use a neural network, that we call the example poses network (fig. 4.10). It is a layered feedforward network with one layer of weights:

- the input is constituted by 36 binary units that corresponds to the 36 frames;
- the output is constituted by 8(angles)x30(bins)=240 binary units;
- for the network learning we set to high the input unit that correspond to a frame and to low all the others, and set the output to the data set we obtained measuring the 8 body angles for the selected frame, so we did learning using 36 data sets: when the measured angle fall into the n-th bin all the output units are set to 0 except the n-th, which is set to 1.

When training is finished we can get the body pose associated with an example frame by setting all the input units to low except the one that corresponds to the frame: so the network outputs the body pose that the subject had in that frame.
At this point we are able to evaluate the poses of the thirtysix example frames, that are just the poses we measured. Let's see how to evaluate a generic pose (provided it is similar to some of the example poses) with the k-NNW method: we will show it with an example.

Let's consider the first two frames of the action walking (figg. 4.11, 4.12). Now let's draw a silhouette of a person with pose similar to that of these two

Figure 4.12: Second frame of the walking action.

Figure 4.13: Silhouette of the intermediate pose.

frames, i.e. an intermediate pose (fig. 4.13). As we expected the ouputs of the snapshot neurons for the first frame of walking value all 0 except the first that values 1, while for the second frame of walking they value all 0 except the second that values 1. Now let's consider the output of the snapshot neurons for the intermediate frame (tab. 4.1). We can observe that the two maximum values are the ones that correspond to the first and the second frame of the action walking: we have found the two nearest neighbours. So now we can use the 2-NNW method to evaluate the pose of the person whose silhouette is the one if fig. 4.13. Let's take the two maximum outputs among the snapshot neurons and normalize them: we obtain the weights

$$w_1 = 0.537721 \tag{4.9}$$

$$w_1 = 0.462279. \tag{4.10}$$

They are used in the implementation of the 2-NNW method, wich is shown in fig. 4.14; let's describe our implementation of pose estimation constituted by a network of five layers of units that we call the 2-NNW network.

1. We compute the output of the example pose estimation network for the two nearest neighbours (frames 1 and 2 of walking); for simplicity we limit ourselves to consider only the column of outputs relative to the left knee: we obtain that for the first frame only the 9th unit values 1 while for the second frame only the 7th unit values 1 (the left knee angle has diminished in walking as we can see in figg. 4.11, 4.12); if we put together this two sets of units we obtain the first layer of the 2-NN network.

2. The second layer of the network is constituted by thirty units, and the i-th unit receives connections only from the i-th unit of the two sets of units of the first layer respectively with weights w_1 and w_2. We obtain that all the units of the second layer value 0 except for the 9th and the 7th, that value respectively w_1 and w_2.

3. The third layer of the network is constituted by one unit that receive connections from the i-th unit of the second layer with weight i. The output value is 8.08.

4. The fourth layer of the network is constituted by thirty radial gaussian basis function: the i-th function has centre $\mu_i = i$ and standard deviation $\sigma = 1$.

Action	Frame	Output
walking	1	0.531943
walking	2	0.457312
walking	3	-0.090291
walking	4	0.006783
walking	5	0.004365
walking	6	0.000292
walking	7	0.001904
walking	8	-0.000846
walking	9	0.002122
walking	10	-0.000292
walking	11	0.000504
walking	12	0.000519
running	1	-0.003116
running	2	0.107875
running	3	-0.040614
running	4	0.012076
running	5	-0.019293
running	6	0.002468
running	7	0.000515
running	8	0.000544
running	9	0.000114
running	10	-0.000040
running	11	-0.000521
running	12	0.000656
marching	1	0.101119
marching	2	-0.074737
marching	3	-0.011283
marching	4	0.012179
marching	5	-0.014432
marching	6	0.005324
marching	7	0.008799
marching	8	-0.003484
marching	9	-0.002222
marching	10	0.002579
marching	11	0.001043
marching	12	-0.000236

Table 4.1: Output of the snapshot neurons for the intermediate frame.

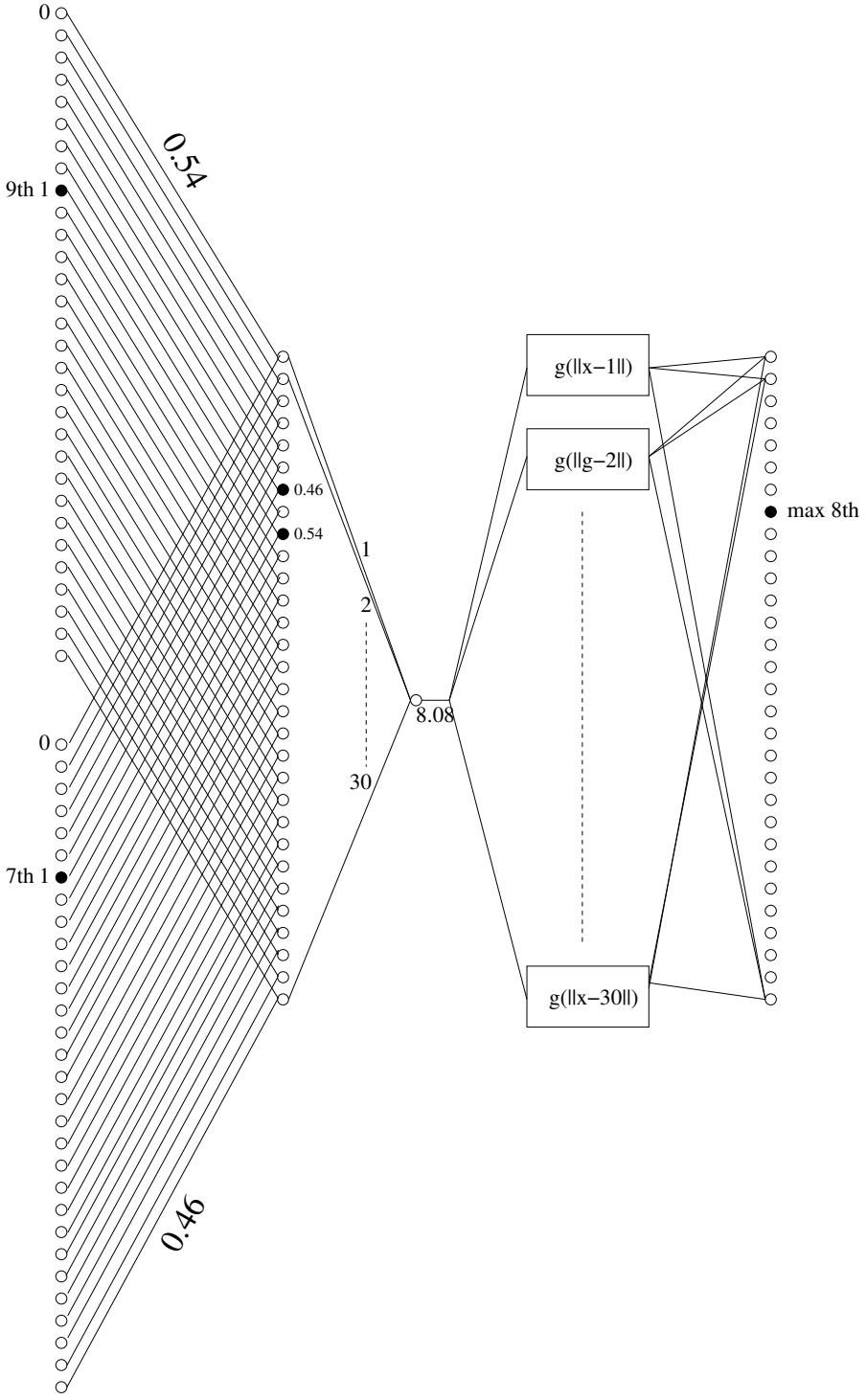

Figure 4.14: Implementation of the 2-NNW method for pose estimation.

5. The fifth layer of the network is constituted by neurons that constitute, together with the fourth layer, a radial gaussian basis functions network. This subnetwork is trained in this way: when the input (the third layer neuron) values i only the i-th output unit must value 1 and all the others must value 0.

In tab. 4.2 is shown the output of the 2-NNW network for the example we have considered; we can see that the unit with maximum value is that of the 8th unit: in the next section, that describes motor commands evaluation, we will interpret this result by stating that the occupated bin of the left knee angle for the intermediate frame is the 8th. If we compute the 2-NNW network for the eight body angle we obtain human pose estimation with an example based method.

4.3 Motor commands evaluation

In this section we describe the evaluation of the motor commands, which is done comparing the human poses at successive times obtained with the pose estimation stage. Motor commands are the difference of joint angles between successive frames. This is justified by the fact that, neglecting low level control processes, a motor command states the amplitude of the movement to be performed.

The evaluation is performed computing a network that we call the motor commands estimation network, which is shown in fig. 4.15: it is composed of three layers of units.

- Let's suppose we have estimated the human pose at time $n-1$ and at time n; for simplicity we consider the outputs of the 2-NNW network only for a single angle, say the left shoulder angle, for the frame $n-1$ and n. If we put together these two columns of units we obtain the first layer of the motor command estimation network, which is therefore composed of 2x30=60 units. We interpret the index of the maximum unit among each of the two sets as the occupated bin of the left shoulder angle respectively at time $n-1$ and n.

- All the possible variations of the maximum unit index among the two sets of thirty units are 2x(30-1)+1=59. Then we define 59 basis functions f_n that compute the maximum among each of the two sets of units and produce binary output in this way:

Unit	Output
1	-0.001117
2	0.002767
3	-0.005401
4	0.009829
5	-0.017074
6	0.031317
7	-0.069171
8	0.988623
9	0.083668
10	-0.036154
11	0.020098
12	-0.012000
13	0.007452
14	-0.004725
15	0.003045
16	-0.001973
17	0.001300
18	-0.000988
19	0.000752
20	-0.000504
21	0.000329
22	-0.000156
23	0.000194
24	-0.000021
25	0.000031
26	-0.000073
27	0.000039
28	0.000172
29	-0.000080
30	0.000015

Table 4.2: output of the snapshot neurons for the intermediate frame.

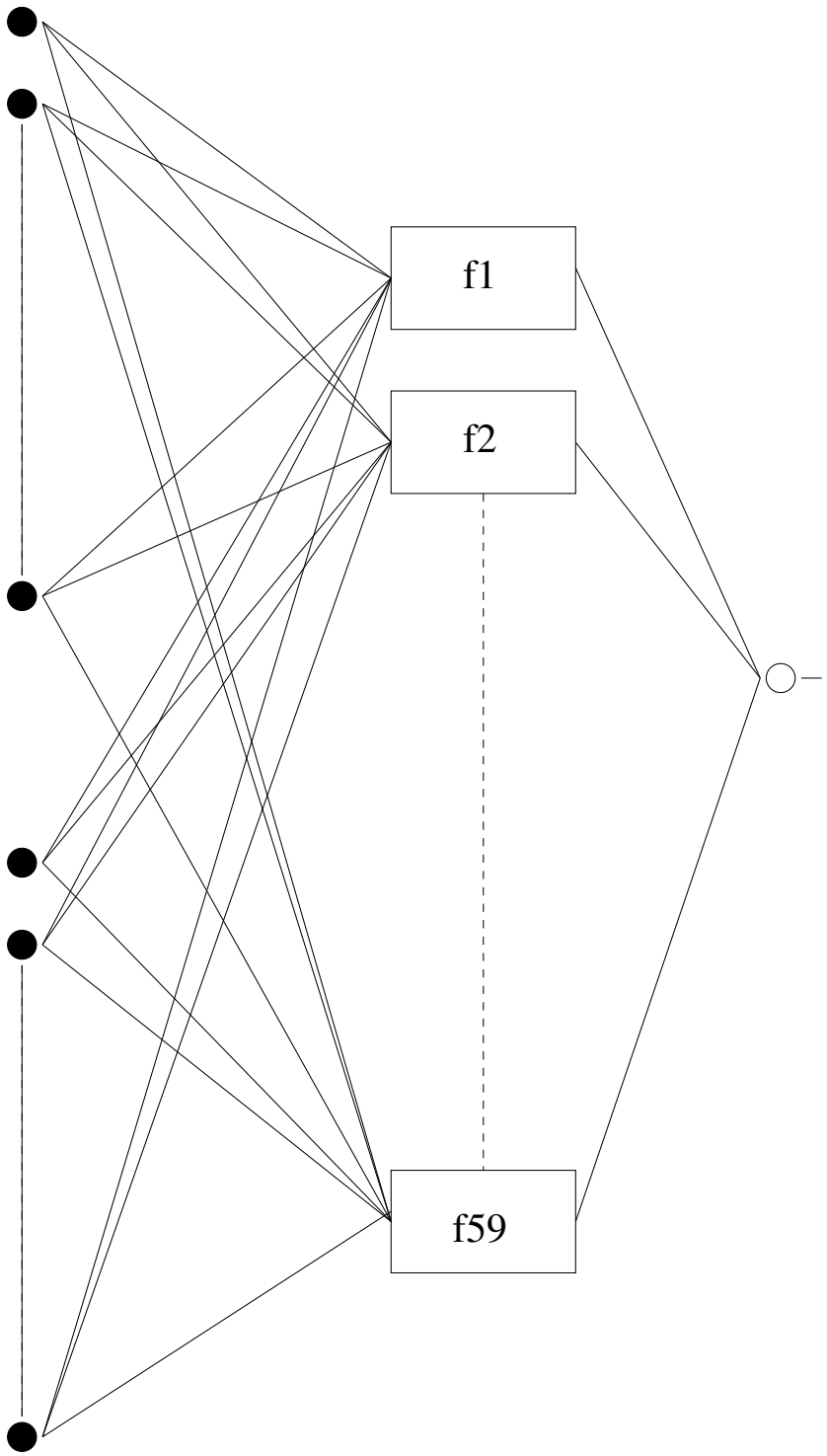

Figure 4.15: Motor commands estimation network.

  - f_1 is high only if the variation of the maximum unit index values -29;
  - f_2 is high only if the variation of the maximum unit index values -28;
  - ...
  - f_{59} is high only if the variation of the maximum unit index values +29;

- the third layer of the motor commands estimation network is constituted by a single neuron wich is trained in this way:
  - the output values -29 if only the first input is high;
  - the output values -28 if only the second input is high;
  - ...
  - the output values +29 if only the 59th input is high;

If we compute the motor commands estimation network for all the eight body angles we obtain the motor commands at time n. At time n we obtain 8x(n-1) outputs. This stage of computation reflects the behaviour of neurons directly connected to the muscle-skeletal system: MEP studies show that these neurons are active also when a human recognizes an action.
In the next section we will explain how it is possible to recognize an action by comparing the motor commands evaluated with the known motor commands of the three actions considered.

4.4 Action recognition

This stage reflects the behaviour of the mirror neurons, that are selective for the meaning of an action: let's see how it works.

- Let's consider the measured angles of the left shoulder in the action walking from time 1 to time 12; we have evaluated the index of the occupated bin of each angle; then we can evaluate the 11 variations of the bin in the time, i.e. 11 motor commands that define the vector $\vec{w}_{ls}$.
- Let's define the subvectors of dimension n-1 of $\vec{w}_{ls}$: we call them $\vec{w}_{ls}^{n,i}$.
- We consider now the n-1 outputs of the motor commands neuron for the left shoulder in the action to be recognized until time n: they define the vector $\vec{x}_{ls}(n)$

- We define the variables

$$x^i(n) = ||\vec{x}_{ls}(n) - \vec{w}_{ls}^{n,i}||. \tag{4.11}$$

- we define the gaussian function

$$g(x) = e^{-\frac{x^2}{2\sigma^2}}; \tag{4.12}$$

with $\sigma = 1$, and we evaluate the fiducial level

$$g^i_{w_{ls}}(n) = g(x^i(n)). \tag{4.13}$$

- we average over the eight body angles:

$$g^i_w(n) = \frac{g^i_{w_{ls}}(n) + g^i_{w_{le}}(n) + g^i_{w_{lh}}(n) + g^i_{w_{lk}}(n) + l \to r}{8}. \tag{4.14}$$

- we take the maximum over i to evaluate the average fiducial level of the best matching motor commands in time for the action walking:

$$g_w(n) = max\{g^i_w(n)\}. \tag{4.15}$$

- finally we evaluate also $g_r(n)$ and $g_m(n)$ with the same steps, and we normalize them:

$$\bar{g}_w(n) = \frac{g_w(n)}{g_w(n) + g_r(n) + g_m(n)} \tag{4.16}$$

$$\bar{g}_r(n) = \frac{g_r(n)}{g_w(n) + g_r(n) + g_m(n)} \tag{4.17}$$

$$\bar{g}_m(n) = \frac{g_m(n)}{g_w(n) + g_r(n) + g_m(n)}. \tag{4.18}$$

In this way we have evaluated the fiducial level of an unknown action to be a known action at any time. An action is classified to be the known action with maximum fiducial level at the final time.

In figg. 4.16, 4.17, 4.18 are shown the outputs of the model respectively for the actions walking, running, marching used for training.

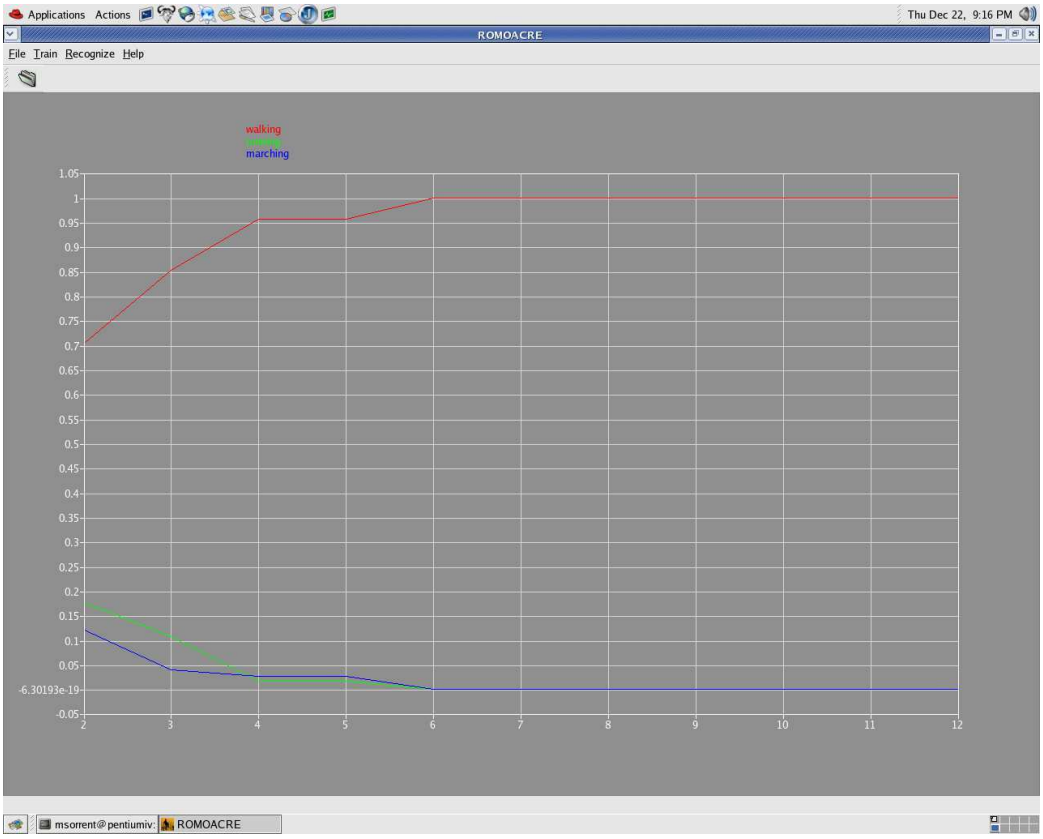

Figure 4.16: Output of the model for the action walking.

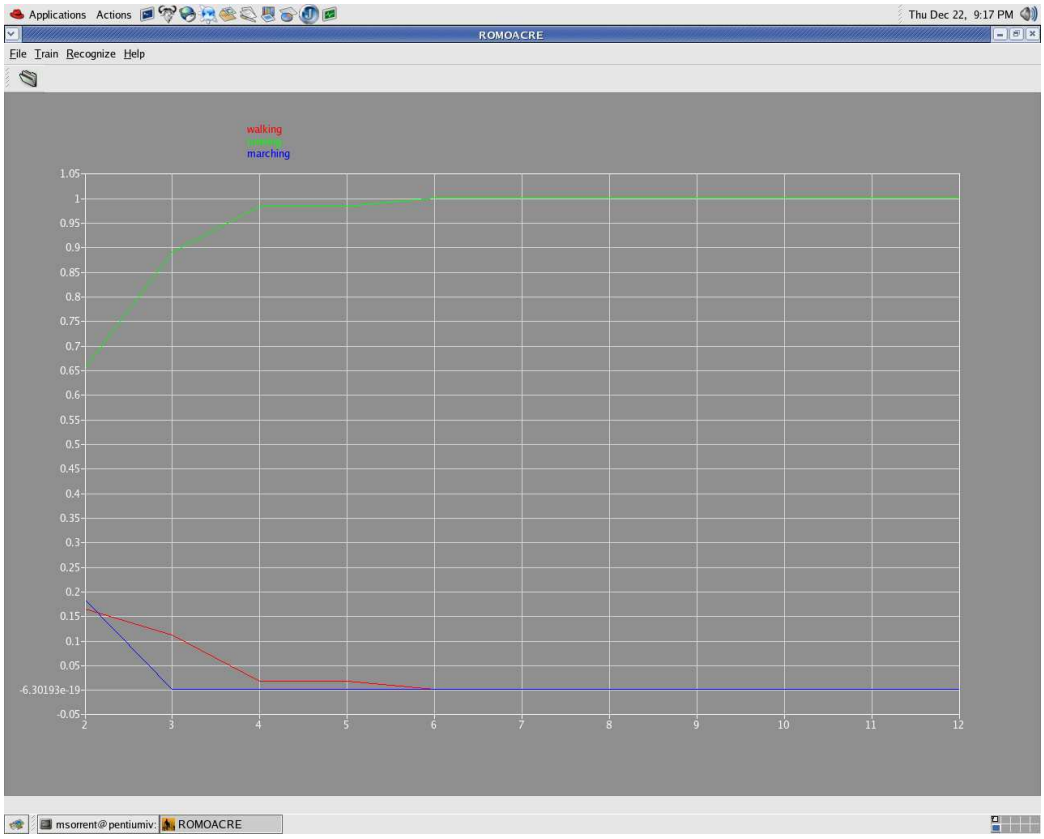

Figure 4.17: Output of the model for the action running.

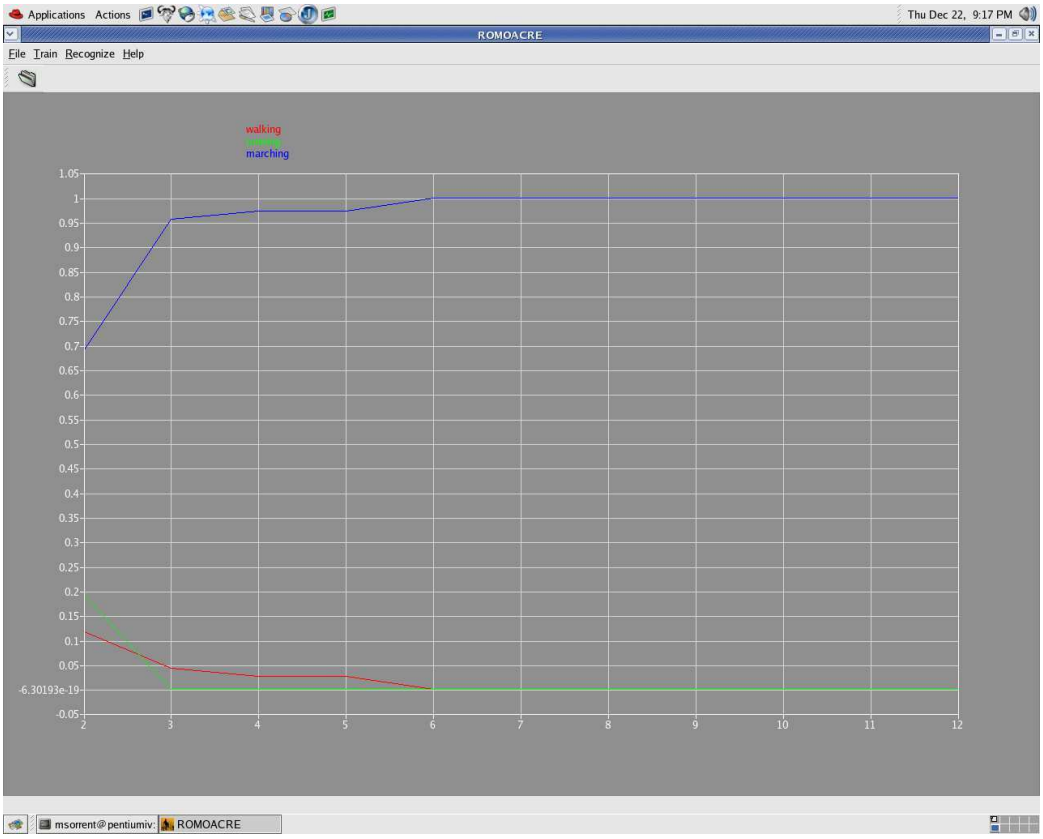

Figure 4.18: Output of the model for the action marching.

Chapter 5

ROMOACRE

In this chapter we describe ROMOACRE. Let's briefly recall the structure of ROMOACRE, which is composed, as its precursor, by the following computational stages:

1. human pose estimation from images of body silhouette;

2. evaluation of the motor commands of the action from human poses sequence (pragmatic recognition);

3. recognition of the action from motor commands (semantic recognition).

The first difference between ROMOACRE and its precursor is that in the precursor the actions are real, videos are captured and the poses are measured; in ROMOACRE, instead, actions are generated, videos are produced and poses are exported. It is better to generate the poses instead of measure them because measured poses are affected by errors.

Before we describe the three computational stages we will describe the framework used to generate actions.

5.1 Action generation with POSER

In order to generate actions we used POSER, a third-party software which allows the creation of 3D motion of the human body (fig. 5.1). With POSER is possible to create .pz3 files which contain information about the action, the point of view of the action, the actor and etc.

There are three ways to generate an action with POSER:

- modify manually the pose of the actor in each frame of the action;

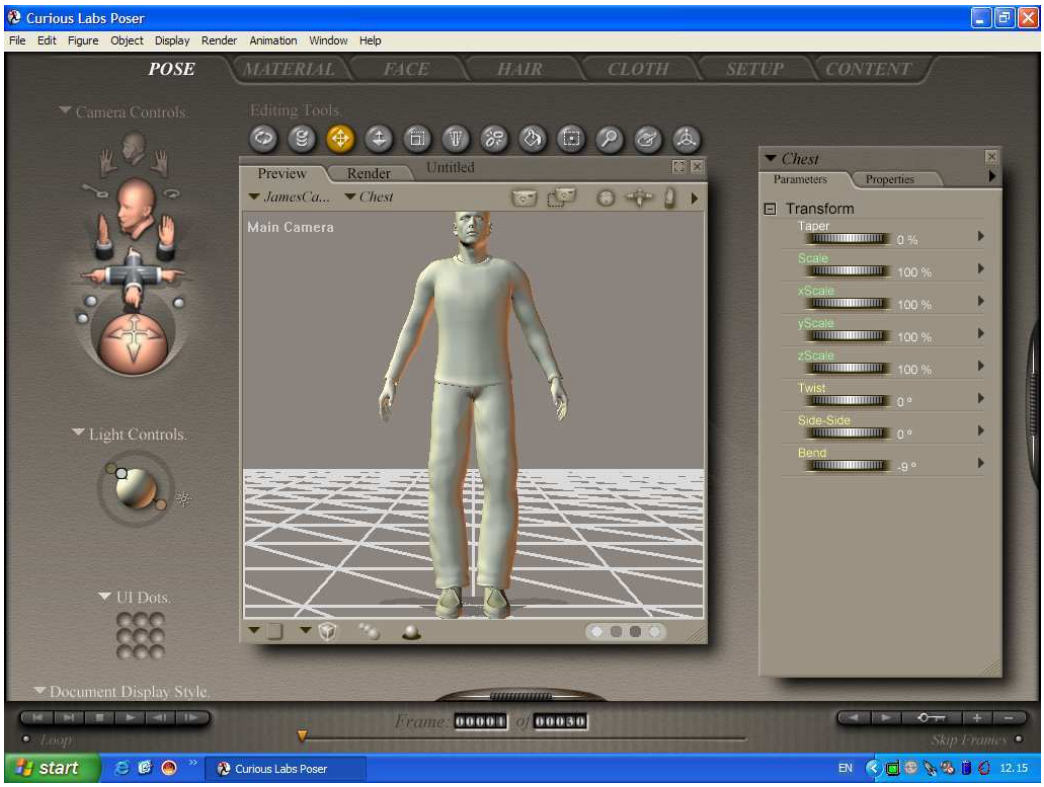

Figure 5.1: A snapshot of POSER.

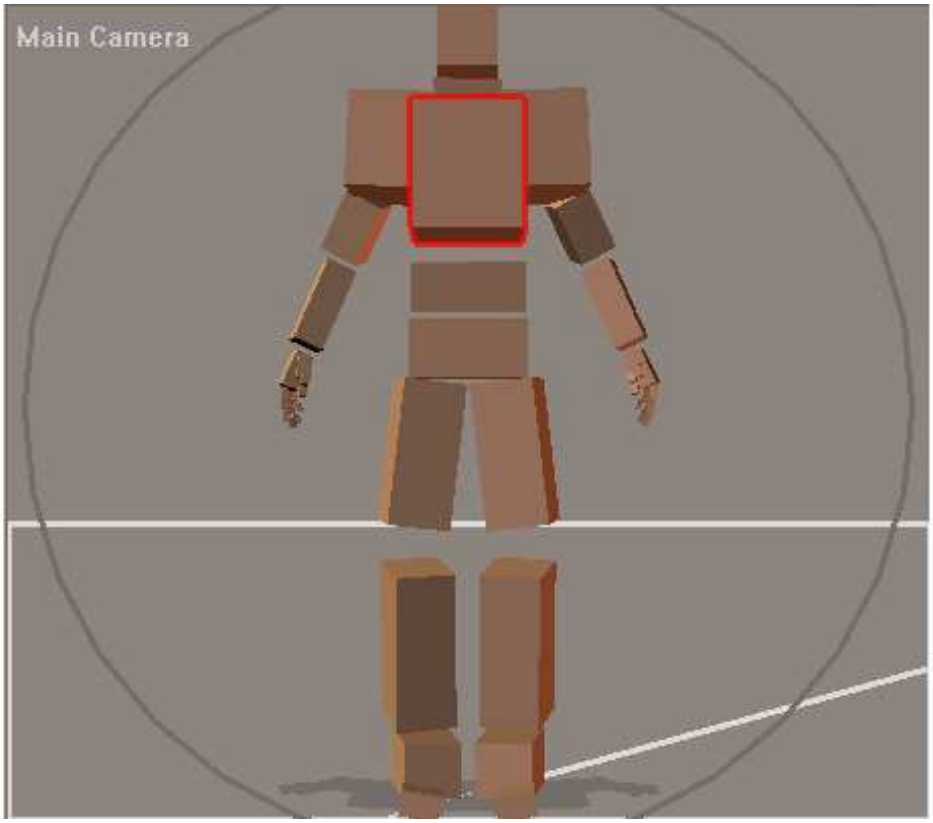

Figure 5.2: The standard body model of POSER (53 body parts).

- using a walk designer included in POSER which allows creation of actions similar to walking modifing some kinematic parameters of the action walking (e.g. head bounce) and other generic parameter (e.g. run);

- import in POSER a file which contains information about the motion of the human body. These are the BVH files (biovision hierarchical files) developed by BIOVISION.

In order to describe how we generate action with POSER we have to explain the BVH file format. This is described in APPENDIX B.

5.1.1 Action generation

In order to generate actions we performed the following steps:

- Using the walk designer we generated three actions which we call walking, running, and marching. In these actions the body model of POSER is used (fig. 5.2). It is very detailed: it includes rotation angles of the eyes, of the segments of fingers, and of toes. It models 53 body parts. The parameters of the model are 53x3+3(hip position)=162. Each action is made up of 30 frames with a frame rate of 30 frames per second;

- The BVH files of each action were exported from POSER;

- The BVH files obtained describe actions with a much too complex body model: in the image processing phase we will work with 128x128 images, and this makes impossible to extract information about fingers position. Therefore we operate in this way on the BVH files. In the hierarchy section we remove manually the too much detailed joints (e.g. fingers segments). In the motion section we set the number of frames to 15 intead of 30 and the frame time to 0.066666 instead of 0.033333. Then with a shell script and AWK (a tool which operates on formatted files) we delete the even rows from the motion section in order to obtain 15 raws and we remove the column corresponding to body parts removed in the hierarchy section. Finally we fix the position of the hip to 0 30 0 because we want the only free parameters to be the rotation angles of the joints. In this way we obtained BVH files with a body model with 19 body parts, i.e. 57 parameters;

- The BVH files obtained were imported into POSER;

- Three sets of 15 128x128 jpg images of the silhouette of the human body performing each action was produced, so we produced a total of 45 images. (figg. 5.3, 5.4, 5.5). Consequently the model analyses the silhouettes instead of natural images; we have shown in the appendix A how it is possible to transform RGB images into silhouette images.

5.2 The form pathway

In order to perform body pose estimation we need to extract a set of shape descriptors from the image. This operation is performed by the form pathway. Now we concentrate on the implementation of the form pathway. In the previous description of ROMOACRE we implemented the form pathway taking inspiration from the model of movement recognition of Giese and Poggio [15]. The form pathway was constituted by three computational levels which we summarize:

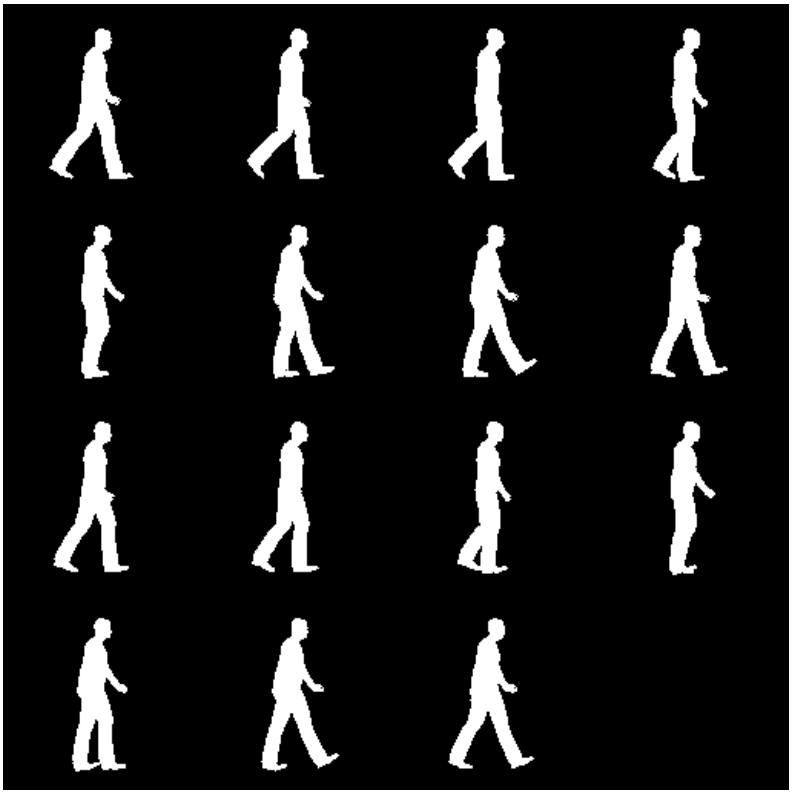

Figure 5.3: Images of the silhouettes in the action walking.

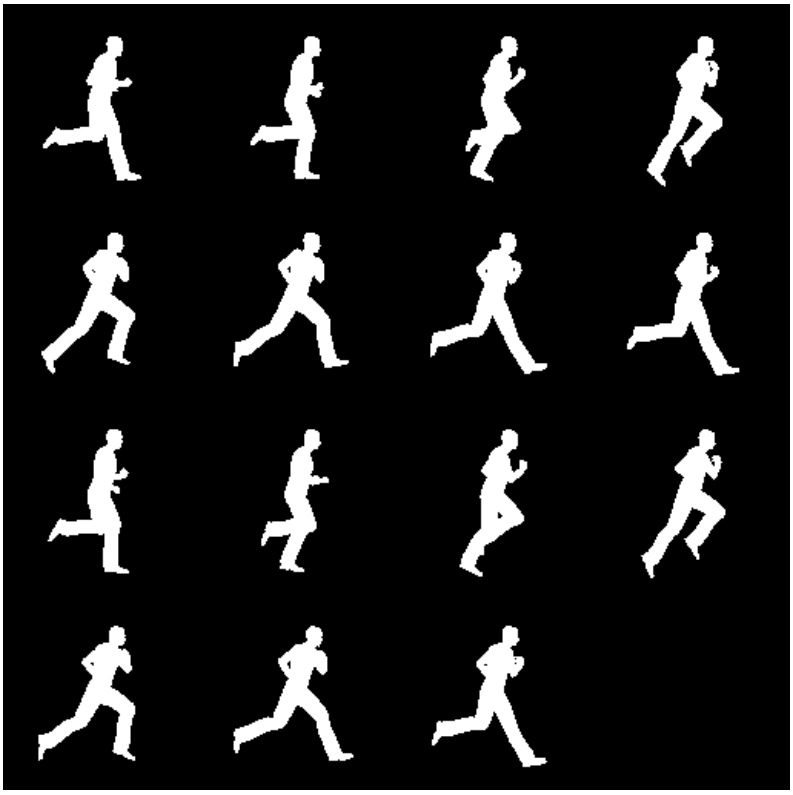

Figure 5.4: Images of the silhouettes in the action running.

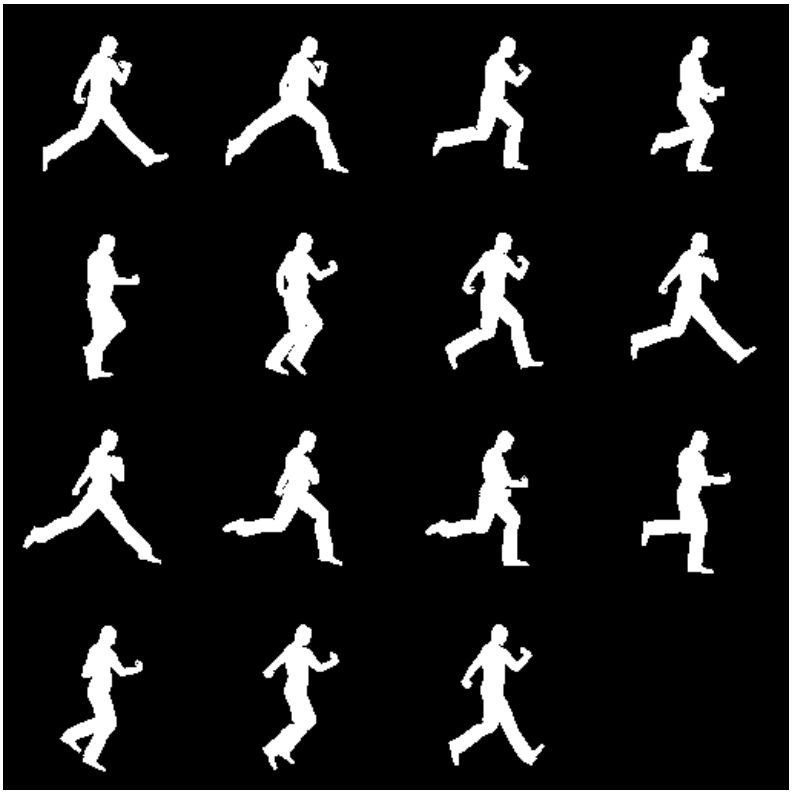

Figure 5.5: Images of the silhouettes in the action marching.

1. Local orientation detectors that model simple cells in the primary visual cortex (V1): we introduced two batteries of Gabor filters for two scales and for 36 orientations;

2. Position and scale tolerant bar detectors (V2), that were implemented computing the maximum among 4x4=16 Gabor filters of a given orientation of the scale 1 and 2x2=4 Gabor filters of the same orientation of the scale 2 (20 Gabor filters in total);

3. Neural network with gaussian radial basis functions in the hidden layer and with snapshot neurons selective for body shapes in the output layer (IT, STS, FA).

The limitation of this implementation of the form pathway is that it is only tolerant to position and scale changes but it is not invariant to such changes. The only way to obtain such invariance is trough learning, i.e. to train the network in the third level with stimuli modified in position and scale. But there is now quantitative physiological evidence [49] that view tuned units (IT) are invariant to position and scale changes, even though the stimulus was previously presented at only one scale and position.

In this chapter we present a new implementation of the form pathway (wich is a part of ROMOACRE) that shows intrinsic invariance to position and scale changes and is inspired to a model for object recognition of Riesenhuber and Poggio [49]. The MATLAB code of the form pathway is in appendix C.

5.2.1 Structure of the form pathway

The form pathway consists of a hierarchy of five levels:

1. S1 layer models simple cells;

2. C1 layer models complex cells;

3. S2 layer models composite feature cells;

4. C2 layer models complex composite cells;

5. VTU layer models view tuned cells.

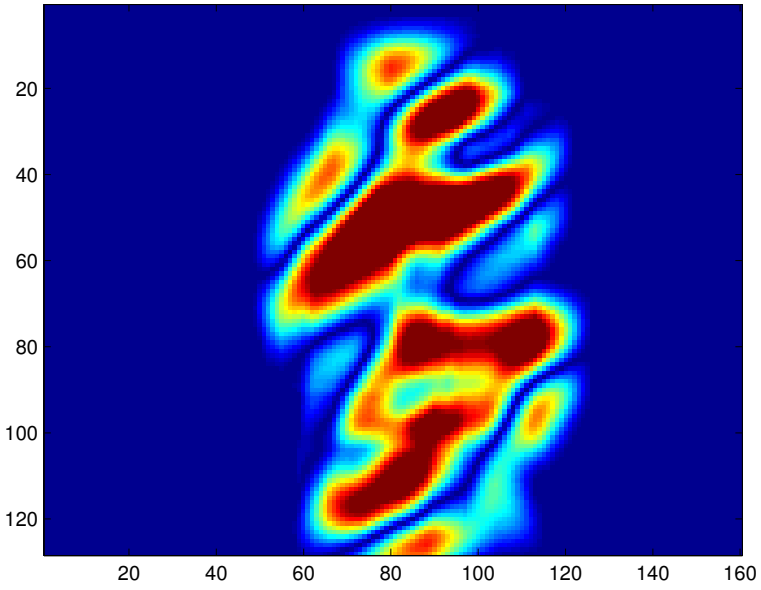

Figure 5.6: Convolution of an image with a Gabor filter.

5.2.2 S1 layer

The binary images 128x128 are densely sampled by two dimensional Gabor filters, the so called S1 units. We filter the images with filters selective for four different orientations (0^0, 45^0, 90^0, 135^0) and of twelve different sizes (7x7, 9x9, 11x11, 13x13, 15x15, 17x17, 19x19, 21x21, 23x23, 25x25, 27x27, 29x29). Such filters are sensitive to bars of different orientations, thus resempling properties of simple cells in striate cortex. Filters of each size and orientation are centered at each pixel of the input image. The filters are sum-normalized to zero and square-normalized to 1, and the result of the convolution of an image patch with a filter is divided by the power (sum of squares) of the image patch. This yields an S1 activity between -1 and 1. An example of the convolution of an image with a gabor filter is shown in fig. 5.6.

5.2.3 C1 layer

In this step filter bands are defined, i.e. groups of S1 filters of a certain size range. Within each filter band, a pooling range is defined which determines

Filter band	Filter sizes	Pooling range
1	7x7, 9x9	4x4
2	11x11, 13x13, 15x15	6x6
3	17x17, 19x19, 21x21	9x9
4	23x23, 25x25, 27x27, 29x29	12x12

Table 5.1: Filter bands and pooling ranges.

the size of the array of neighboring S1 units of all sizes in that filter band which feed into a C1 unit (corresponding to complex cells of striate cortex). Only S1 filters with the same preferred orientation feed into a given C1 unit to preserve feature specificity. In tab. 5.1 filter bands and pooling ranges are illustrated. The pooling operation that the C1 units use is the "MAX" operation, i.e. a C1 unit's activity is determined by the strongest input it receives. That is, a C1 unit responds best to a bar of the same orientation as the S1 units that feed into it, but already with an amount of spatial and size invariance that corresponds to the spatial and filter size pooling ranges used for a C1 unit in the respective filter band. Furthermore, the receptive fields of the C1 units overlap by a certain amount, given by the value of the parameter c1Overlap. We used a value of 2, meaning that half the S1 units feeding into a C1 unit were also used as input for the adjacent C1 unit in each direction. Higher values of c1Overlap indicate a greater degree of overlap.

5.2.4 S2 layer

Within each filter band, a square of 2x2=4 adjacent, nonoverlapping C1 units is then grouped to provide input to a S2 unit. There are 256 different types of S2 units in each filter band, corresponding to the 4^4 possible arrangements of four C1 units of each of four types (i.e. preferred bar orientation). The S2 unit response function is a Gaussian with mean (1,1,1,1) and standard deviation 1, i.e. an S2 unit has maximal response (we can interpret this response as a firing rate) of 1 which is attained if each of its four afferents fires at a rate of 1 as well. S2 units provide the feature dictionary of our version of the form pathway; in this case all combination of 2x2 arrangements of "bars" (more precisely, C1 cells) at four possible orientations.

5.2.5 C2 layer

To finally achieve size invariance over all filter sizes in the four filter bands and position invariance over the whole visual field, the S2 units are again pooled by a MAX operation to yield C2 units, designed to correspond to

neurons in extrastriate visual area V4 or posterior IT (PIT). There are 256 C2 units, each of which pools over all S2 units of one type at all positions and scales. Consequently, a C2 unit will fire at the same rate as the most active S2 unit that is selective for the same combination of four bars, but regardless of its scale or position.

5.2.6 VTU layer

C2 units then again provide input to the viewtuned units (VTUs), named after their property of responding well to a certain two-dimensional view of a three-dimensional object, thereby closely resembling the view-tuned cells found in monkey inferotemporal cortex (IT). The $C2 \rightarrow VTU$ connections are so far the only stage of this implementation of the form pathway where learning occurs. A VTU is tuned to a stimulus by selecting the activities of the 256 C2 units in response to that stimulus as the center of a 256-dimensional gaussian response function, yielding a maximal response of 1 for a VTU in case the C2 activation pattern exactly matches the C2 activation pattern evoked by the training stimulus. The parameter that specifies the response properties of a VTU is the standard deviation of its gaussian response function. A smaller standard deviation yields more specific tuning since the resultant Gaussian has a narrower half-maximum width. We set the standard deviation at 0.375.

5.2.7 Invariance and selectivity

Here we show the results obtained with this implementation of the form pathway.

In fig. 5.7 is shown the selectivity plot for all stimuli, i.e. the 45 VTU responses for all the 45 stimuli.

In fig. 5.8 is shown the selectivity plot for one stimulus. We can see that the level of the distractors is below 0.6.

In fig. 5.9 are shown silhouettes of various size of the same stimulus of fig. 5.8 and in fig. 5.10 is shown the plot of scale invariance. We can see that the level of the VTU is above 0.8. This indicates that there is a good balance between selectivity and scale invariance.

In fig. 5.11 is shown the same stimulus of fig. 5.8 with different location in the image, and in fig. 5.12 is shown the position invariance plot, i.e. the responses of the VTU to the stimuli shown in fig. 5.11. We observe that the response is near to 1, then there is a good balance between selectivity and position invariance.

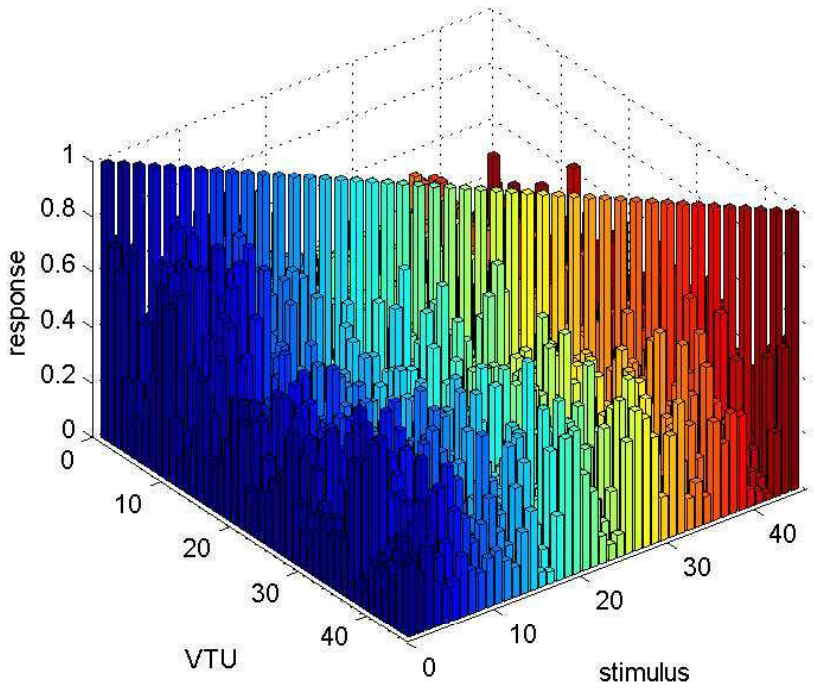

Figure 5.7: Selectivity plot for all stimuli.

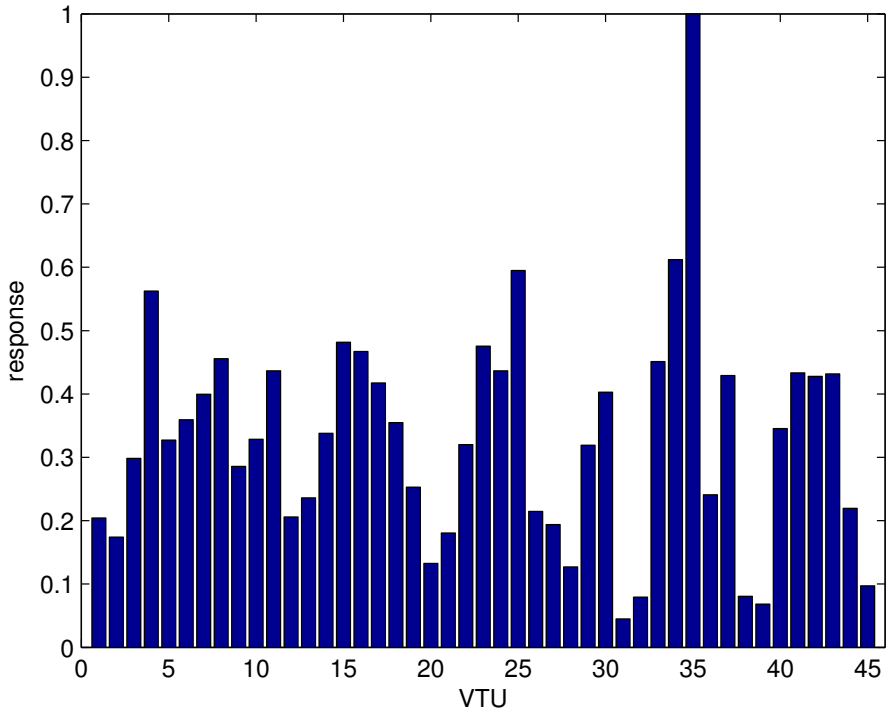

Figure 5.8: Selectivity plot for one stimulus.

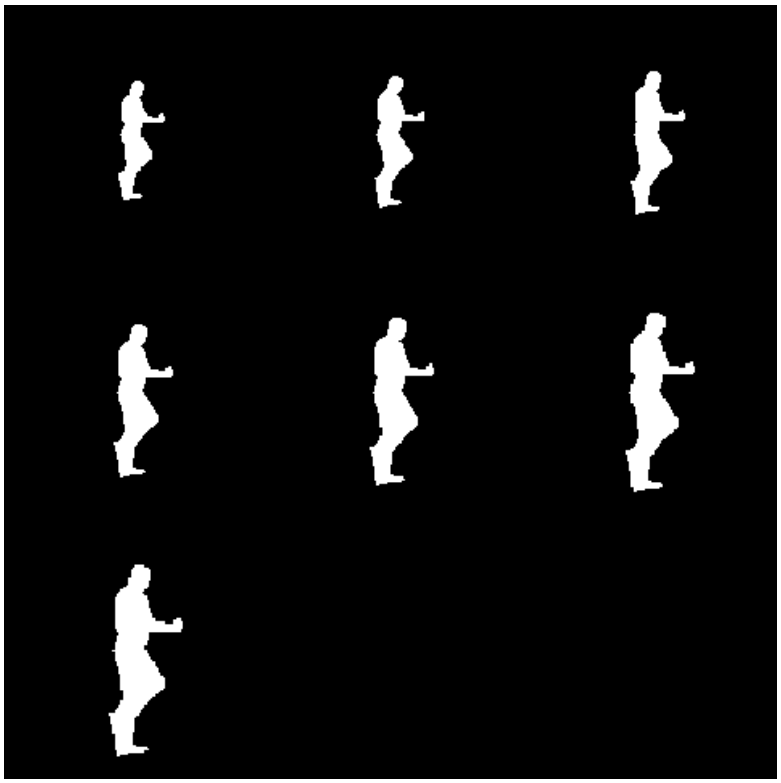

Figure 5.9: Different sizes of the same silhouette.

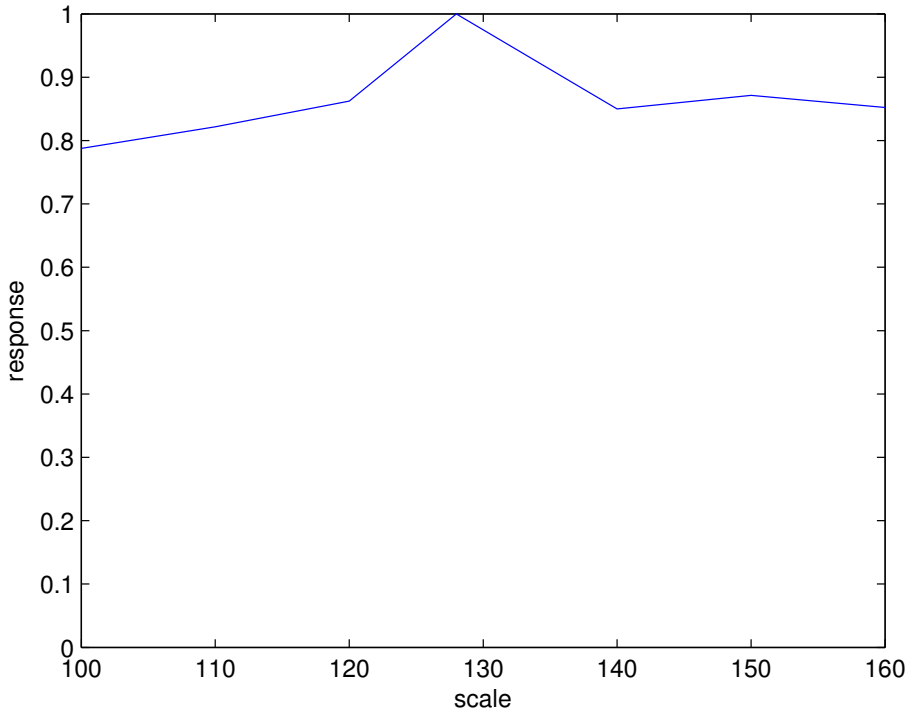

Figure 5.10: VTU responses to different scales.

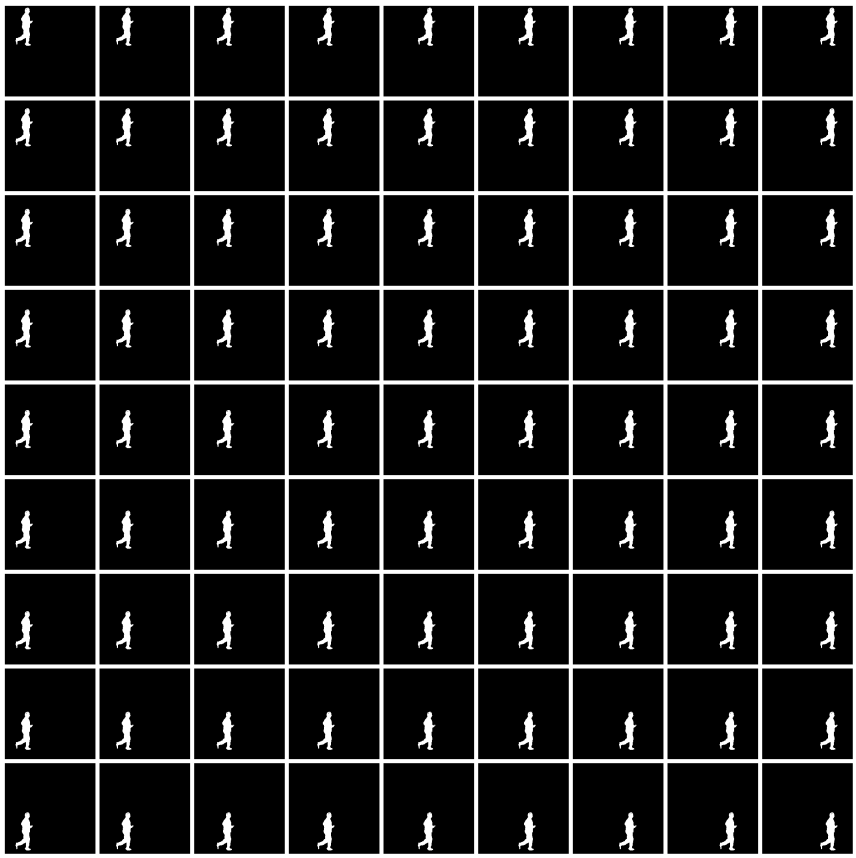

Figure 5.11: Different position of the silhouette.

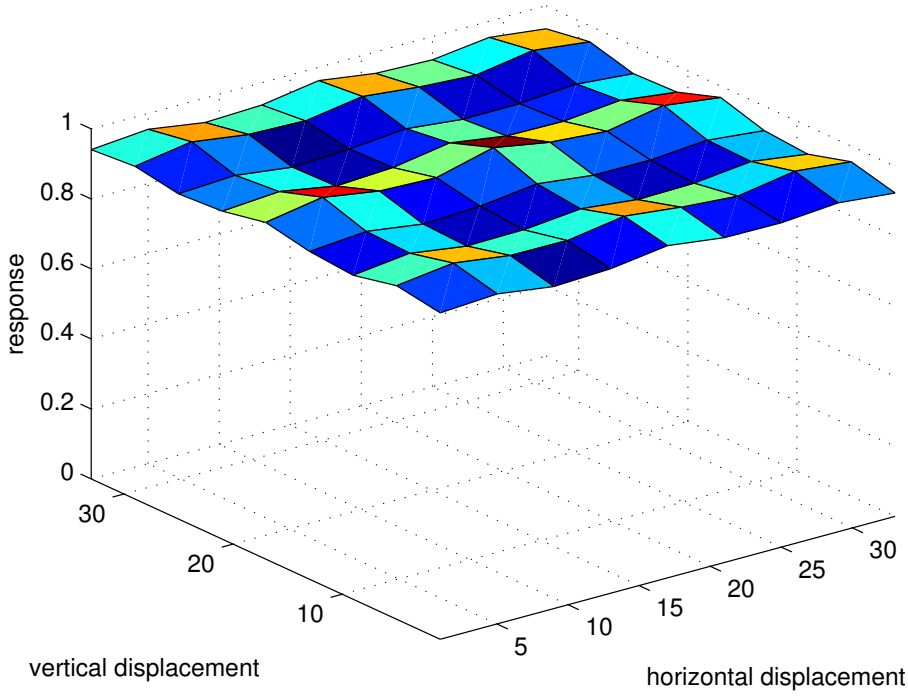

Figure 5.12: VTU responses to different position of the silhouette.

5.3 From shape descriptors to body pose

In this implementation of ROMOACRE the VTU are used only to test selectivity and invariance of the model. We don't use them to evaluate the body pose. Instead we use the C2 units as shape descriptors and we perform a regression from the space C2 (256 parameters) to the space of body pose (57 parameters). The regression method used is that of radial basis function neural network. As training set we used the 45 frames of the three actions generated. So, for each frame we extracted the C2 parameters and we know the body pose parameters from the BVH files.

5.4 Motor command evaluation

Now we have to evaluate the known motor commands for each action and the motor command of the percepted action. We cannot subtract the angles of two consecutive poses as they are Cardan angles. We can evaluate the rotation matrix for each joint, i.e. the matrix that rotates the parent joint to the child joint. The rotation matrices around each axis are:

$$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\alpha) & sin(\alpha) \\ 0 & -sin(\alpha) & cos(\alpha) \end{pmatrix}$$

$$R_y = \begin{pmatrix} cos(\beta) & 0 & sin(\beta) \\ 0 & 1 & 0 \\ -sin(\beta) & 0 & cos(\beta) \end{pmatrix}$$

$$R_x = \begin{pmatrix} cos(\gamma) & sin(\gamma) & 0 \\ -sin(\gamma) & cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The rotation order is Y Z X, so the rotation matrix is

$$R = R_x R_z R_y =$$

$$\begin{pmatrix} cos(\beta)cos(\gamma) & sin(\gamma) & sin(\beta)cos(\gamma) \\ -cos(\alpha)cos(\beta)sin(\gamma) - sin(\alpha)sin(\beta) & cos(\alpha)cos(\gamma) & -cos(\alpha)sin(\beta)sin(\gamma) + sin(\alpha)cos(\beta) \\ sin(\alpha)cos(\beta)sin(\gamma) - cos(\alpha)sin(\beta) & -sin(\alpha)cos(\gamma) & sin(\alpha)sin(\beta)sin(\gamma) + cos(\alpha)cos(\beta) \end{pmatrix}$$

Now we can evaluate the motor commands in this way: given two joint matrices at consecutive times we define a motor command as the matrix that transforms the previous joint matrix into the next joint matrix (as we already said we neglect the low level control processes):

$$M_{n-1} J_{n-1} = J_n$$

so

$$M_{n-1} = J_n J_{n-1}^{-1}$$

In this way we evaluate the known motor commands using as Cardan angles the ones in the BVH files. We are also able to extract motor commands of the perceived angles using Cardan angles produced with the RBF regression.

5.5 Pragmatic recognition

At this point the model is able to imitate the action perceived. The model will perceive the pose of the first frame. In order to produce the poses of the second frame it applies the motor command perceived matrix (for a given joint) to the joint matrix of the first frame:

$$M_1 J_1 = J_2$$

$$M_2 M_1 J_1 = J_3$$

and so on. Now using Hamilton's quaternions (associating a quaternion to a rotation) it is possible to extract from the joint matrices the Cardan angles. Finally, we insert the obtained Cardan angles in a BVH file, which defines the imitated action.

5.6 Semantic recognition

In order to perform a classification of the action we measure the match between known motor commands of each known action and perceived motor commands. There will be a matrix X such that

$$X M_k = M_p$$

where M_k is the matrix of known motor commands and M_p is the matrix of perceived motor commands. So

$$X = M_p M_k^{-1}$$

Therefore X is a rotation matrix, so it has an eigenvalue 1 to which corresponds an eigenvector that indicates the direction of the rotation axis; the other two eigenvalues are $e^{\pm i\theta}$ where θ is the rotation angle around the axis. Then

$$\theta = arccos(\frac{\lambda_2 + \lambda_3}{2})$$

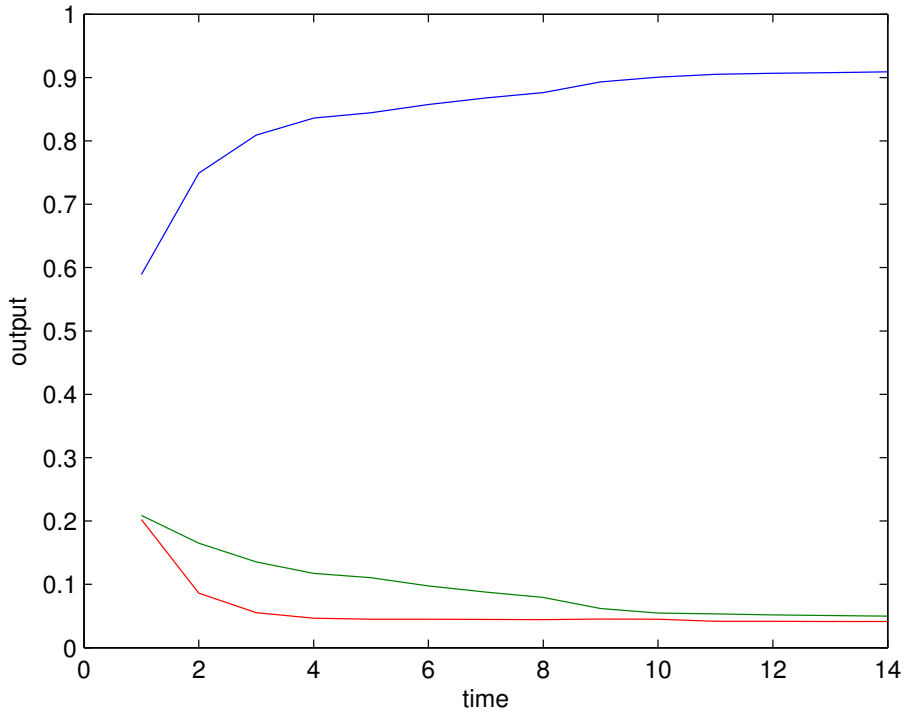

Figure 5.13: Output of the model for the action walking.

For a given action and a given joint at time n we will have a vector of eigenvalues $\theta_1, ..., \theta_n$ so we can evaluate the fiducial level for a given known action and a given joint:

$$g_n = e^{-\frac{|\vec{\theta_n}|^2}{2\sigma^2}}$$

Finally we average the fiducial levels over all joints and we normalize the fiducial levels of the three actions to 1. Finally, we have successfully tested the model. In figg. 5.13, 5.14, 5.15 are shown the outputs of the model respectively for the actions walking, running, marching. In the first plot (in which the action perceived is walking) we can see that the curve relative to walking goes to 1 while the others go to 0 as the time increases. A similar result is shown in the other two plots.

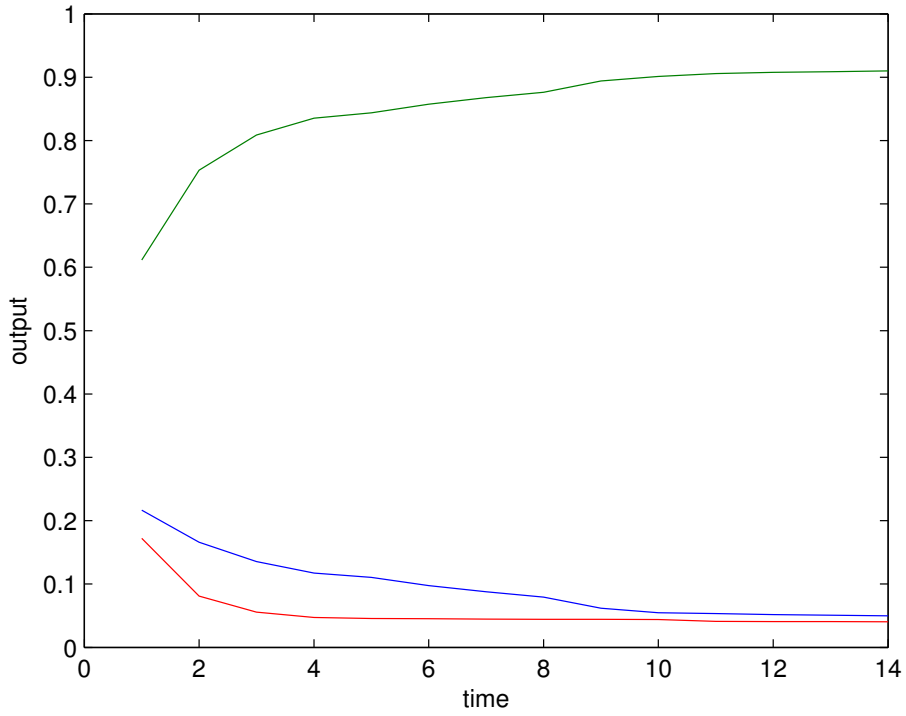

Figure 5.14: Output of the model for the action running.

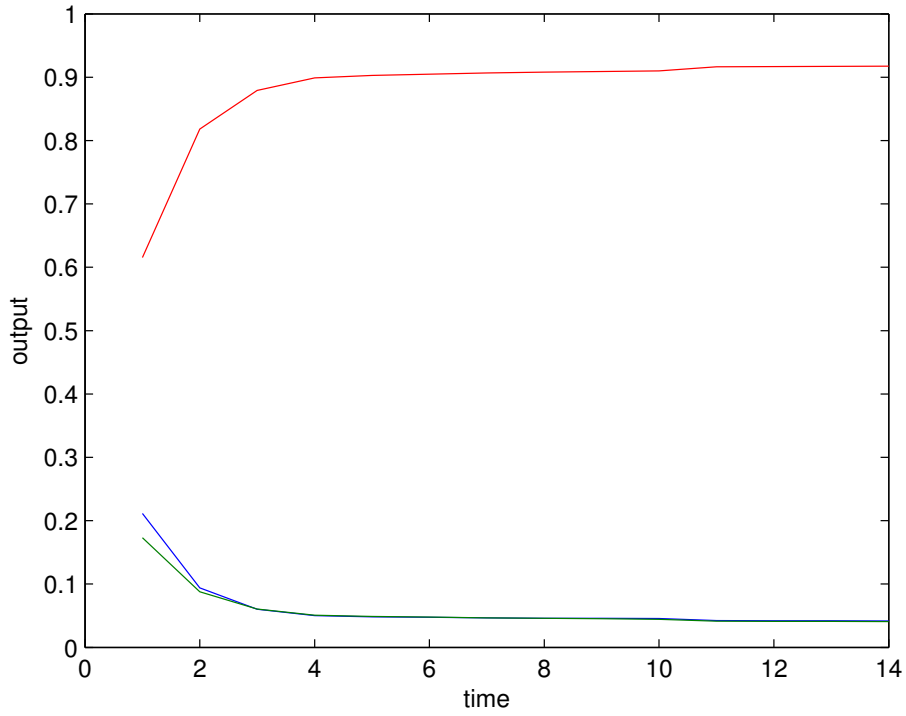

Figure 5.15: Output of the model for the action marching.

Appendix A

Here follow the details of image capture.

- Three videos were captured of a person who was walking, running, marching. The person was by profile with respect to the video camera. The person was entirely dressed in black and the background was approximatively white. The videos were captured with the camera of a video phone LG U8330. The camera resolution is $160x128$ $pixel$ and the frame rate is approximatively 20 $frame/s$. The video file format is $.3gp$.

- The three $.3gp$ files were copied on a computer with Windows XP operating system and were converted to $.avi$ compressed file format using ImTOO 3GP Video Converter.

- The three $.avi$ compressed files were uncompressed using AviToMpeg video converter.

The further steps of image preprocessing are illustrated in figg. 5.16, 5.17.

1. 160x128 RGB images are extracted from the video (fig. 5.16 top-left).

2. RGB images are converted to grayscale images (fig. 5.16 top-right).

3. Negative images are evaluated (fig. 5.16 bottom-left).

4. Negative grayscale images are converted to binary images (fig. 5.16 bottom-right).

5. The borders of the binary images are erased (filled with black, cutting operation) because we assume that the silhouette is in the central part of the visual field (fig. 5.17 top-left).

6. The four-connected objects in the images are distinguished and labeled (fig. 5.17 top-right).

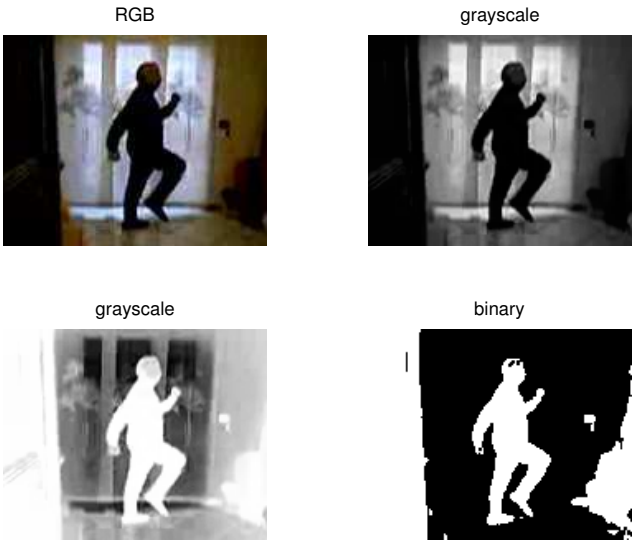

Figure 5.16: top-left: RGB image; top-right: grayscale image; bottom-left: negative image; bottom-right: binary image.

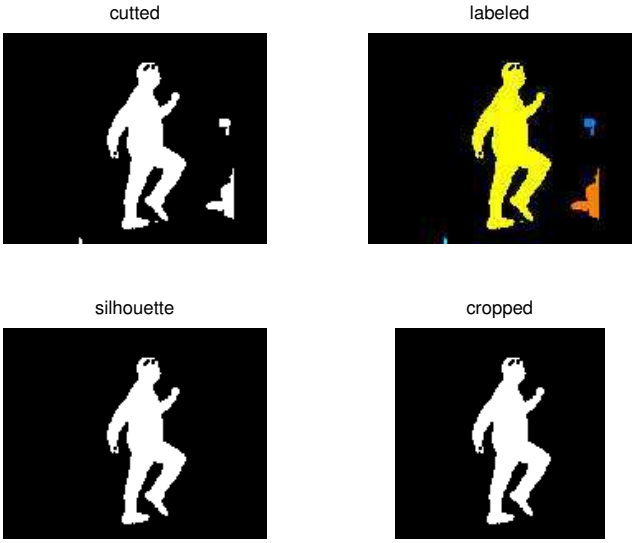

Figure 5.17: top-left: cutted image; top-right: labeled image; bottom-left: silhouette image; bottom-right: cropped image.

7. Only the object with the maximum area is taken, because it is supposed to be the silhouette (fig. 5.17 bottom-left).

8. The 160x128 images are cropped to 128x128 images in order to simplify the form pathway code (fig. 5.17 bottom-right).

So at the end of these steps we got twelve images of human silhouettes at the temporal distance of about 0.1 s for three actions: walking, running, marching.

Appendix B

Here follows a BVH example file and its description.

```
HIERARCHY
ROOT Hips
{
OFFSET 0.00 0.00 0.00
CHANNELS 6 Xposition Yposition Zposition Zrotation Xrotation Yrotation
JOINT Chest
{
OFFSET  0.00  5.21  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT Neck
{
OFFSET  0.00  18.65  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT Head
{
OFFSET  0.00  5.45  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
End Site
{
OFFSET  0.00  3.87  0.00
}
}
}
JOINT LeftCollar
{
OFFSET  1.12  16.23  1.87
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT LeftUpArm
{
```

```
OFFSET  5.54  0.00  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT LeftLowArm
{
OFFSET  0.00 -11.96  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT LeftHand
{
OFFSET  0.00 -9.93  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
End Site
{
OFFSET  0.00 -7.00  0.00
}
}
}
}
}
JOINT RightCollar
{
OFFSET -1.12  16.23  1.87
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT RightUpArm
{
OFFSET -6.07  0.00  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT RightLowArm
{
OFFSET  0.00 -11.82  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT RightHand
{
OFFSET  0.00 -10.65  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
End Site
{
OFFSET  0.00 -7.00  0.00
}
}
}
}
```

```
}
}
JOINT LeftUpLeg
{
OFFSET  3.91  0.00  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT LeftLowLeg
{
OFFSET  0.00 -18.34  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT LeftFoot
{
OFFSET  0.00 -17.37  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
End Site
{
OFFSET  0.00 -3.46  0.00
}
}
}
}
JOINT RightUpLeg
{
OFFSET -3.91  0.00  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT RightLowLeg
{
OFFSET  0.00 -17.63  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
JOINT RightFoot
{
OFFSET  0.00 -17.14  0.00
CHANNELS 3 Zrotation Xrotation Yrotation
End Site
{
OFFSET  0.00 -3.75  0.00
}
}
}
}
}
```

```
MOTION
Frames:    2
Frame Time: 0.033333
 8.03  35.01  88.36 -3.41  14.78 -164.35  13.09  40.30 -24.60  7.88
43.80  0.00 -3.61 -41.45  5.82  10.08  0.00  10.21  97.95 -23.53
-2.14 -101.86 -80.77 -98.91  0.69  0.03  0.00 -14.04  0.00 -10.50
-85.52 -13.72 -102.93  61.91 -61.18  65.18 -1.57  0.69  0.02  15.00
22.78 -5.92  14.93  49.99  6.60  0.00 -1.14  0.00 -16.58 -10.51 -3.11
15.38  52.66 -21.80  0.00 -23.95  0.00
 7.81  35.10  86.47 -3.78  12.94 -166.97  12.64  42.57 -22.34  7.67
43.61  0.00 -4.23 -41.41  4.89  19.10  0.00  4.16  93.12 -9.69 -9.43
132.67 -81.86  136.80  0.70  0.37  0.00 -8.62  0.00 -21.82 -87.31
-27.57 -100.09  56.17 -61.56  58.72 -1.63  0.95  0.03  13.16  15.44
-3.56  7.97  59.29  4.97  0.00  1.64  0.00 -17.18 -10.02 -3.08  13.56
53.38 -18.07  0.00 -25.93  0.00
```

Biovision BVH

The BVH file format was originally developed by Biovision, a motion capture services company, as a way to provide motion capture data to their customers. The name BVH stands for Biovision hierarchical data. This format mostly replaced an earlier format that they developed, the BVA format, as a way to provide skeleton hierarchy information in addition to the motion data. The BVH format is an excellent all around format, its only drawback is the lack of a full definition of the basis pose (this format has only translational offsets of children segments from their parent, no rotational offset is defined), it also lacks explicit information for how to draw the segments but this has no bearing on the definition of the motion.

Parsing the file

A BVH file has two parts: a header section which describes the hierarchy and initial pose of the skeleton and a data section which contains the motion data. Let's examine the example BVH file. The start of the header section begins with the keyword "HIERARCHY". The following line starts with the keyword "ROOT" followed by the name of the root segment of the hierarchy to be defined. After this hierarchy is described it is permissible to define another hierarchy; this too would be denoted by the keyword "ROOT". In principle, a BVH file many contain any number of skeleton hierarchies. In

practice there is often only one hierarchy.

The BVH format now becomes a recursive definition. Each segment of the hierarchy contains some data relevant to just that segment then it recursively defines its children. The line following the "ROOT" keyword contains a single left curly brace "{"; the brace is lined up with the "ROOT" keyword. The line following a curly brace is indented by one tab character; these indentations are mostly to just make the file more human readable. The first piece of information of a segment is the offset of that segment from its parent; or in the case of the root object the offset will generally be zero. The offset is specified by the keyword "OFFSET" followed by the X,Y and Z offset of the segment from its parent. The offset information also indicates the length and direction used for drawing the parent segment. In the BVH format there isn't any explicit information about how a segment should be drawn. This is usually inferred from the offset of the first child defined for the parent. Typically, only the root and the upper body segments will have multiple children.

The line following the offset contains the channel header information. This has the "CHANNELS" keyword followed by a number indicating the number of channels and then a list of that many labels indicating the type of each channel. The BVH file reader must keep track of the channel count and the types of channels encountered as the hierarchy information is parsed. Later, when the motion information is parsed, this ordering will be needed to parse each line of motion data. This format appears to have the flexibility to allow for segments which have any number of channels which can appear in any order. However, we have never encountered a BVH file that didn't have 6 channels for the root object and 3 channels for every other object in the hierarchy.

You can see that the order of the rotation channels appears a bit odd: it goes Z rotation, followed by the X rotation and finally the Y rotation. This is not a mistake, as the BVH format uses any rotation order (in the BVH files produced by POSER the order of the written rotation angles is X Z Y, but the rotation is performed in the inverse order Y Z X).

On the line of data following the channels specification there can be one of two keywords; either you will find the "JOINT" keyword or you will see the "End Site" keyword. A joint definition is identical to the root definition except for the number of channels. This is where the recursion takes place, the rest of the parsing of the joint information proceeds just like a root. The end site information ends the recursion and indicates that the current segment is an end effector (it has no children). The end site definition provides one more bit of information: it gives the length of the preceding segment just like the offset of a child defines the length and direction of its parents

segment.

The end of any joint, end site or root definition is denoted by a right curly brace "}". This curly brace is lined up with its corresponding right curly brace.

One last note about the BVH hierarchy: the world space is defined as a right handed coordinate system with the Y axis as the world up vector. Thus you will typically find that BVH skeletal segments are aligned along the Y axis.

The motion section begins with the keyword "MOTION" on a line by itself. This line is followed by a line indicating the number of frames, this line uses the "Frames:" keyword (the colon is part of the keyword) and a number indicating the number of frames, or motion samples that are in the file. On the line after the frames definition is the "Frame Time:" definition, which indicates the sampling rate of the data. In the example BVH file the sample rate is given as 0.033333, which is 30 frames a second (the usual rate of sampling in a BVH file).

The rest of the file contains the actual motion data. Each line is one sample of motion data. The numbers appear in the order of the channel specifications as the skeleton hierarchy was parsed.

Appendix C

Here follows the matlab code for S1 layer.

```
function S1 = evaluateS1 (stim,filter)

global filter_size num_filter_size num_orientation

S1_size=size(stim,1);
S1=zeros(S1_size,S1_size,num_filter_size,num_orientation);

for isize=1:num_filter_size
    xy_range=[-(filter_size(isize)-1)/2:+(filter_size(isize)-1)/2];
    [x,y]=meshgrid(xy_range);
    circle=(x.^2+y.^2 <= ((filter_size(isize)-1)/2)^2);
    circled=zeros(filter_size(isize));
    circled(:,:)=circle(:,:);
    norm=sqrt(imfilter(stim.^2,circled))+eps;
    for iorientation=1:num_orientation
        S1(:,:,isize,iorientation)=imfilter(stim,filter{isize,iorientation});
        S1(:,:,isize,iorientation)=abs(S1(:,:,isize,iorientation))./norm;
    end
end
```

Here follows the matlab code for C1 layer.

```
function C1=evaluateC1(S1)

global num_orientation C1_pooling num_band

S1_size=size(S1,1);
num_band=4;
C1=zeros(S1_size,S1_size,num_band,num_orientation);
C1_pooling=[4 6 9 12];
C1_band=[1 2;3 5;6 8;9 12];
C1_band_size=[2 3 3 4];
for iband=1:num_band
    C1xy=zeros(S1_size,S1_size,C1_band_size(iband),num_orientation);
    for i=1:S1_size
        x1=i;
        x2=min([S1_size x1+C1_pooling(iband)-1]);
        for j=1:S1_size
            y1=j;
            y2=min([S1_size y1+C1_pooling(iband)-1]);
            S1_patch=S1(x1:x2,y1:y2,C1_band(iband,1):C1_band(iband,2),:);
            C1xy(i,j,:,:)=max(max(S1_patch,[],2),[],1);
        end
    end
    C1(:,:,iband,:)=max(C1xy,[],3);
end
```

Here follows the matlab code for S2 layer.

Here follows the matlab code for the S2 layer.

```
function S2 = S2resp_zeropad (C1)

% function S2 = S2resp (C1)
% This function returns S2 responses given C1.

global num_orientation;
global C1_shift;
global S2_shift;
global S2_config;
global S2_target;
global S2_sigma;

S2_buf_size = size(C1,1) - C1_shift*(S2_config-1);
S2_buf = zeros(S2_buf_size,S2_buf_size,num_orientation,S2_config^2);

% Reshape C1 into S2_buf.
for i = 1:S2_buf_size
  for j = 1:S2_buf_size
    k = 0;
    for m = 1:S2_config
      for n = 1:S2_config
      k = k+1;
    ii = i+(m-1)*C1_shift;
    jj = j+(n-1)*C1_shift;
    S2_buf(i,j,:,k) = C1(ii,jj,:);
      end
    end
  end
end

% Resample S2_buf.
S2_range = [1:S2_shift:S2_buf_size];
S2_tmp = S2_buf(S2_range, S2_range, :,:);

% Get different configurations for S2.
% (256 x 4, for example)
idx = 0;
for l = 1:4
  for k = 1:4
```

```
    for j = 1:4
      for i = 1:4
    idx = idx+1;
    seq(idx,:) = [i j k l];
      end
    end
  end
end

for m = 1:size(seq,1)
  for n = 1:size(seq,2)
    S2_permute(:,:,m,n) = S2_tmp(:,:,seq(m,n),n);
  end
end

S2 = squeeze(sum((S2_permute-S2_target).^2,4));
S2 = exp(-S2/2/S2_sigma^2);
```

Here follows the matlab code for C2 layer.

```
function C2=evaluateC2(S2)

global num_band

C2=zeros(4,256);

for iband=1:num_band

    C2(iband,:)=max(max(S2{iband},[],1),[],2);

end
C2=max(C2,[],1);
```

Conclusions

The purpose of this work has been to develop a neurophysiological plausible model of action recognition in humans. The developed model has been called ROMOACRE (a **RO**botic **MO**del for **AC**tion **RE**cognition).

As background preliminary knowledge we relied on the studies on the central nervous system, with particular attention to cortical areas involved in visual and motor tasks, and to the related cortical circuits.

Secondly, we used the studies of the mirror mechanism, which is the neural mechanism mainly involved in action recognition.

Following that, we studied existing models of action recognition, two of which were considered of special interest for this work: the model by Giese and Poggio, and the model by Demiris and Johnson.

Having acquired this preliminary knowledge we were able to build the first version of our model for action recognition, which is called "The precursor of ROMOACRE".

Finally we developed ROMOACRE in its final release and we propose some tests of the model.

The improvements between the two models which we considered as reference models and the model we have proposed consist in the relation between single images recognition and action recognition, and in the inclusion of mirror mechanism in the architecture.

The first difference between the two considered models and ROMOACRE is the following: in the two considered models action recognition takes place as a consequence of single images recognition, while in ROMOACRE action recognition does take place through the perception of motor commands, which are considered as the difference of joint angles between successive frames (a rotation matrix in ROMOACRE). This is justified by the fact that, neglecting low level control processes, a motor command states the amplitude of the movement to be performed.

In fact in the model by Giese and Poggio there are, in the third stage of

form pathway and of motion pathway, snapshot neurons which are selective for particular silhouette shapes of the human body. So there is a stage (stage three) previous to the stage which implements action recognition (stage four) that implements image recognition.

This is also the case in the model by Demiris and Johnson, in which two states of the demonstrator (current state and target state) are fed as input to the set of inverse models and, moreover, the current state is fed as input to the set of forward models. So also in this case there is a very simple stage which implements image recognition previous to action recognition.

The same process happens in the precursor of ROMOACRE. In fact in this model the human pose estimation is a two steps process. The first step takes as input the raw data from the images and produces as output a vector of activities of snapshot neurons which are selective for different human body silhouettes. The second step performs pose evaluation with the k-NNW method, taking as weights the output of the first step. Following this, motor commands, not considered in the reference models, are evaluated and finally the action is recognized. So, also in this model action recognition is subsequent to single images recognition.

Instead the relation between single images recognition and action recognition in ROMOACRE is different. As its precursor, ROMOACRE is made up of three computational stages: human pose estimation, motor commands evaluation and action recognition (the same as its precursor). But the human pose estimation stage is different from the same stage in the precursor; in fact the pose estimation stage is divided in two steps: step one takes as input the raw data from the images and produces as output a vector of shape descriptors of the silhouette. Step two performs pose estimation through a regression between the space of shape descriptors and the space of body pose parameters. These two steps do not execute image recognition but they perform the identification of motor commands. From the sequence of these two steps and not from the recognition of single images, the semantic action recognition takes place in ROMOACRE.

A strong point of the final version of our model, which does not select for single image recognition, is its agreement with experimental data, in particular the experiment of Johansson [48]. In this famous experiment ten light points were attached on the joints of some actors and some movies of actions were produced where only the light points are visible. The observers were not able to recognize the single dot frames as a human body but they recognized the

actions watching the full dot movie. This experiment proves that single images recognition is not a necessary requirement for action recognition. This result is consistent with the architecture of ROMOACRE.

The second difference between the considered models and ROMOACRE is the inclusion of the mirror mechanism in the architecture. Giese and Poggio state that the fourth stage of their model corresponds to F5 mirror area but the mechanism they implement in order to perform action recognition is not a mirror mechanism; it resembles instead the mechanism of expected perception. On the other hand Demiris and Johnson state that they built an architecture for action recognition which includes the mirror mechanism but in their model only pragmatic recognition is performed while semantic recognition is not. In ROMOACRE and its precursor, instead, the action observed is recognized as it is reflected in the motor representation of the observer for the same action, according to what happens in the mirror mechanism.

Bibliography

[1] P. Viviani, G. McCollum, The relation between linear extent and velocity in drawing movements, *Neuroscience* 10 (1983) 211.

[2] P. Viviani, C. A. Terzuolo, Trajectory determines movement dynamics, *Neuroscience* 7 (1982) 431.

[3] F. Lacquaniti, C. Terzuolo, P. Viviani, Global metric properties and preparatory processes in drawing movements, *Preparatory States and Processes* (1983).

[4] I. Mikić, M. Trivedi, E Hunter, P. Cosman, Articulated body posture estimation from multi-camera voxel data, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1 (2001) 455.

[5] R. Rosales, M. Siddiqui, J. Alon, S. Sclaroff, Estimating 3D body pose using uncalibrated cameras, *Proceedings of the 2001 IEEE Computer Society Conference* 1 (2001) 821.

[6] H. Sidengladh, M. Black, D. J. Fleet, Stochastic tracking of 3D human figures using 2D image motion, *Part II in Lecture Notes in Computer Science* (2000) 702.

[7] H. Fei, I. Reid, Dynamic classifier for non-rigid human motion analysis, *BMVC 2004, British Machine Vision Conference* (7-9th September 2004) Kingstone University, London.

[8] M. W. Lee, R, Nevatia, Dynamic human pose estimation using Markov chain Monte Carlo Approach, *Proceedings of the IEEE Workshop on Motion and Video Computing* (WACV/MOTION'05).

[9] G. Hua, M.-H. Yang, Y. Wu, Learning to estimate human pose with data driven belief propagation, *Computer Vision and Pattern Recognition* 2 (2005) 747.

[10] A. Agarwal, B. Triggs, A local basis representation for estimating human pose from cluttered images, *To appear in Proceedings of the 7th Asian Conference on Computer Vision, 2006.*

[11] T. B. Moeslsund, E. Granum, 3D human pose estimation using 2D-data and an altenative phase space representation, *Proceedings HuMAns 2000, Hilton Head Island, South Carolina, June 16, 2000.*

[12] A. Agarwal, B. Triggs, Recovering 3D human pose from monocular images, *IEEE Transactions on Pattern Analysis & Machine Intelligence* 28 1 (2006)

[13] G. Mori, J. Malik, Estimating human body configurations using shape context matching, *Proceedings of the 7th European Conference on Computer Vision-Part III* (2002) 666.

[14] G. Shakhnarovich, P. Viola, T. Darrell, Fast pose estimation with parameter-sensitive hashing, *IEEE International Conference on Computer Vision* 2 (2003) 750.

[15] M. A. Giese, T. Poggio, Neural Mechanism for the recognition of biological movements, *NATURE REVIEWS - NEUROSCIENCE* 4 (2003) 179.

[16] D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *J. Physiol. (Lond.)*, 160 (1962) 106.

[17] J. P. Jones, L. A. Palmer, An evaluation of the two-dimensional Gabor filters model of simple receptive fields in cat striate cortex, *J. Neurophys.* 58 (1987) 1233.

[18] J. Hedge, D. C. van Essen, Selectivity for complex shape in primate visual area V2, *J. Neurosci.* 20 (2000) RC61.

[19] J. L. Gallant, J. Braun, D. C. van Hessen, Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex, *Science* 250 (19993) 100.

[20] A. Pasupathy, C. E. Connor, Response to contour features in macaque area V4, *J. Neurophysiol.* 82 (1999) 2490.

[21] N. D. Logothetis, J. Pauls, T. Poggio, Shape representation in the inferior temporal cortex of monkeys, *Curr. Biol.* 5 (1995) 552.

[22] N. K. Logothetis, D. L. Shenberg, Visual object recognition, *Annu. Rev. Neurosci.* 19 (1996) 577.

[23] K. Tanaka, Inferotemporal cortex and object vision, *Annu. Rev. Neurosci.* 19 (1996) 109.

[24] M. W. Oram, D. I. Perrett, Responses of anterior superior temporal polysensory (STPa) neurons to 'biological motion' stimuli, *J. Cogn. Neurosci.* 6 (1994) 99.

[25] E. D. Grossman, R. Blake, Brain areas active during visual perception of biological motion, *Neuron* 35 (2002) 1167.

[26] E. Bonda et al., Specific involvement of human parietal systems and the amygdala in the perception of biological motion, *J. Neurosci.* 16 (1996) 3737.

[27] R. J. Howard et al., A direct demonstration of functionnal specialization within motion-related visual and auditory cortex of the human brain, *Curr. Biol.* 6 (1996) 1015.

[28] E. Grossman et al., Brain areas involved in perception of biological motion, *J. Cogn. Neurosci.* 12 (2000) 711.

[29] L. M. Vaina et al., Functional neuroanatomy of biological motion perception in humans, *Proc. Natl. Acad. Sci. USA* 98 (2001) 11656.

[30] P. Downing, Y. Jiang, M. Shuman, N. Kanwisher, A cortical areas selective for visual processing of the human body, *Science* 293 (2001) 2470.

[31] D. I. Perrett et al., in *AI and the eye*, (1990) 181.

[32] M. W. Oram, D. I. Perrett, Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey, *J. Neurophysiol.* 76 (1996) 109.

[33] J. G. Daugman, Uncertainty relation for resolutions in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, *Journal of the Optical Society of America A* 2 (1985) 1160.

[34] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, T. Poggio, A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex, *CBCL Paper #259/AI Memo #2005-036, Massachusetts Institute of Technology, Cambridge, MA, December, 2005.*

[35] D. I. Perrett, M. W. Oram, Neurophysiology of shape processing, *Image Vis. Comput.* 11 (1993) 317.

[36] D. Hubel, T. N. Wiesel, Functional architecture of the macaque monkey visual cortex, *Proc. R. Soc. Lond. B* 198 (1977) 1.

[37] K. Fukushima, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, *Biol. Cybern.* 36 (1980) 193.

[38] B. W. Mel, J. Fieser, Minimizing binding errors using learned conjunctive features, *Neural Comp.* 12 (2000) 731.

[39] I. Lampl, M. Riesenhuber, T. Poggio, The max operation in cells in the cat visual cortex, *Soc. Neurosi. Abstr.* 30 (2001) 619

[40] T. J. Gawne, J. Martin, Response of primate visual cortical V4 neurons to two simultaneous presented stimuli. *J. Neurophysiol.* 88 (2002) 1128.

[41] M. Matelli, G. Luppino, Parietofrontal circuits for action and space perception in Macaque Monkey, *Neuroimage* 14 (2001) S27.

[42] G. Rizzolatti, G. Luppino, M. Matelli, The organization of the cortical motor system: new concepts, *Electroencelography and clinical neurophysiology* 106 (1998) 283.

[43] G. Rizzolatti, M. Matelli, Two different streams form the dorsal visual system: anatomy and function, *Exp. Brain Res.* 153 (2003) 246.

[44] G. Rizzolatti, L. Fadiga, V. Gallese, L. Fogassi, Premotor cortex and the recognition of motor actions, *Cognitive brain research* 3 (1996) 131.

[45] M. Jeannerod, The representing brain: neural correlated of motor intetion and imagery, *Beach. Brain Sci.* 17 (1994) 117.

[46] G. Rizzolatti, M. A. Arbib, Language within our grasp, *TINS* 21 (1998) 188.

[47] Y. Demiris, M. Johnson, Distributed predictive perception of actions: a biologically inspired robotics architecture for imitation and learning, *Connection science* 15 (2003) 231.

[48] G. Johansson, Visual perception of biological motion and a model for its analysis, *Percept. Psychophys.* 14 (1973) 201.

[49] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition, *Nature neurosci.* 2 (1999) 1019.

[50] Y. Demiris, G. Hayes, Imitation as dual-route process featuring predictive and learning components: a biologically-plausible computational model, *in Imitation in Animals and Artifacts Kerstin Dautenhahn and Chrystopher L. Nehaniv (eds.), Cambridge, Mass., USA: MIT Press, 2002.*

[51] M. Iacoboni et al., Grasping the intentions of others with one's own mirror neuron system, *PLoS Biol.* 3 (2005) 79.

Acknowledgments

This thesis has been possible thanks, first of all, to my tutor Giuseppe Trautteur; his trustfulness in me gave me the support I needed to carry on this work. I wish to thank also Roberto Prevete for his valuable comments and contributions to some specific aspects of my work. Thanks also to my mother Maddalena, my father Vincenzo and my brother Valerio for being so close to me during these three years. I say thanks also to Bernardo Amato, Alfonso Panico, Giovanni Placidi, and Eugenio Celentano for the special help I received from them. Many thanks also to my friends, especially to Ivo and Tiziana, for being always on my side.