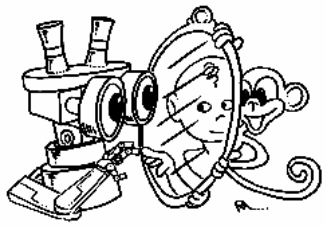

MIRROR

IST–2000-28159

Mirror Neurons based Object Recognition

Deliverable Item 3.4

Modeling of the mirror neurons representation

Delivery Date: April 30, 2003

Classification: Public

Responsible Person: Prof. Giulio Sandini – DIST, University of Genova

Partners Contributed: ALL

Short Description: This deliverable describes a formal biologically compatible model of the functioning of mirror neurons as derived from the investigation on the monkey and the current scientific understanding of the functional role of mirror neurons in action recognition. The model is mathematically described through the Bayesian formalism which fits naturally to the characteristics of problem (i.e. action representation and interpretation). The model also stands as the foundations on which to base the robotic implementation. In addition, this deliverable includes results from the first implementation and testing on a realistic data set. The questions we aimed at answering were:

What is the visual and/or motor information required to develop a representation similar to the one encountered in monkey's area F5? What sequence of learning events has to take place?

How can "visual-only" information of a motor act be used to index an extended representation, coding the learned action (this classification operation is in summary the interpretation of the firing patterns of a mirror neuron)? We propose a methodology that allows an artificial system to perform gesture recognition relying on motor information.

Project funded by the European Community under the "Information Society Technologies" Programme (1998-2002)

Content list

1. Introduction

> *"The goals of MIRROR are: 1) to realize an artificial system that learns to communicate with humans by means of body gestures and 2) to study the mechanisms used by the brain to learn and represent gestures. The biological base is the existence in primates' premotor cortex of a motor resonant system, called mirror neurons, activated both during execution of goal directed actions and during observation of similar actions performed by others. This unified representation may subserve the learning of goal directed actions during development and the recognition of motor acts, when visually perceived. In MIRROR we investigate this ontogenetic pathway in two ways: 1) by realizing a system that learns to move AND to understand movements on the basis of the visually perceived motion and the associated motor commands and 2) by correlated electrophysiological experiments" (Mirror Project Technical Annex).*

This deliverable describes a formal biologically compatible model of the functioning of mirror neurons as derived from the investigation on the monkey and the current scientific understanding of the role of mirror neurons in action recognition. The model is mathematically described through the Bayesian formalism which fits naturally to the characteristics of problem (i.e. action representation and interpretation). The model also stands as the foundations on which to base the robotic implementation. In addition, this deliverable includes results from the first implementation and testing on a realistic data set.

One of the fundamental questions raised by the discovery of mirror neurons which is also the main scientific question in the context of Mirror is to: *"investigate how "visual-only" information of a motor act (not in a self-centered coordinate frame) can be used to index the self-centered extended representation, coding the learned action (this indexing is the core of a mirror neuron) (see also the technical annex)".* A related question is that of understanding how action recognition can be facilitated by the knowledge of how to perform those same actions.

The first section of this deliverable addresses these questions by proposing a methodology that allows an artificial system to perform action recognition relying on motor information. This approach differs from the large majority of existing work on computer vision and robotics since we explicitly exploit motor information. We show that, by performing classification in motor space, the problem is simplified and implicitly provides a much larger degree of invariance to changes in the cameras' point of view.

This system is built upon a Visuo-Motor Map (VMM) that creates an association between the motor representation of the observed gestures (hand postures, in this case) and their visual appearance. The VMM is learned from observations within a supervised learning schema. In addition, the recognition system takes into account object *affordances* (see later). Tests were conducted employing a dataset collected using the project's data-glove setup. A thorough analysis of the difference in performance resulting from the use of motor or visual representations was conducted.

For the time being, the experiments on gesture recognition based on motor representations, have been simplified by employing a simplified color-based strategy to segment the hand and object from the background.

1.1. Historical perspective

The first attempt of modeling perception and action altogether was started several decades ago by Alvin Liberman, aiming at the construction of a 'speech understanding' machine (Liberman et al. 1967, Liberman and Mattingly, 1985, Liberman and Whalen, 2000). As one can easily imagine, the first effort of Liberman's team was directed at analyzing the acoustic

characteristics of spoken words, to investigate whether the same word, spoken by different subjects, possessed any common phonetic invariant. Soon Liberman and his colleagues understood that speech recognition on the basis of acoustic cues alone could not be achieved with the limited computational power available at that time. Somewhat stimulated by this negative result, they put forward the hypothesis that the ultimate constituents of speech are not sounds but rather articulatory gestures that have evolved exclusively at the service of language. Accordingly, a cognitive translation into phonology is not necessary because the articulatory gestures are phonologic in nature. This elegant idea was however strongly debated, mainly because its implementation into a real system was impossible and it only now supported by experimental evidence (Kerzel and Bekkering 2001, Fadiga et al. 2002).

Why is it that, normally, humans can visually recognize actions (or, acoustically, speech) with an approximation of about 99-100%? Why the inter-subject variability typical of motor behavior does not represent a problem for the brain while it is troublesome for machines? One possibility is that Liberman was right in saying that speech perception and speech production use a common repertoire of motor primitives that during speech production are at the basis of articulatory gestures generation, while during speech perception are activated in the listener as the result of an acoustically-evoked motor "resonance". With the only difference of the sensory modality, this sentence might be true also for other, visually perceived, actions. What, in both cases, the brain needs is a "resonant" system that matches the observed/listened actions on the observer/listener motor repertoire. Note that, an additional advantage of such an empathic system would be the capability to automatically "predict", to some extent, the future development of somebody else's actions on the basis of the observer implicit knowledge (on the same actions).

1.2. General setting

Recent neuroscience results suggest a critical role for motor action in perception. Certainly vision and action are intertwined at a very basic level. While an experienced adult can interpret visual scenes perfectly well without acting upon them, linking action and perception seems crucial to the developmental process that leads to that competence. We can construct a working hypothesis: that action is required whenever the animal (or our artifact in the following) has to develop autonomously. Further, the ability to act is also fundamental in interpreting actions performed by a conspecific. Of course, if we were in standard supervised learning setting action would not be required since the trainer would do the job of pre-segmenting the data by hand and providing the training set to the machine. In an ecological context, some other mechanism has to be provided. Ultimately this mechanism is the body itself and the ability of being the initiator of actions that by means of interaction (and under some suitable developmental rule) generate percepts informative to the purpose of learning.

Grossly speaking, a possible developmental explanation of the acquisition of these functions can be framed in terms of tracing/interpreting chains of causally related events. We can distinguish four main conceptual functions (similar to the schema of Arbib et al. (Arbib, 1981)): reaching, grasping (manipulation), "imitation" as per the mirror response, and object recognition. These functions correspond to the four levels of causal understanding introduced in Table 1. We do not delve deeper into the description here, suffices to say that they form an elegant progression of abilities which emerge out of very few initial assumptions. All that is required is the interaction between the actor and the environment, and a set of appropriate developmental rules specifying what information is retained during the interaction, the nature of the sensory processing, the range of motor primitives, etc.

Neurophysiology tells us that neurons from area F5 are involved in the control of grasping and manipulation. F5 projects then to the primary motor cortex for the fine control of movement. It has been already hypothesized that the AIP-F5 system responds to the "affordances" of the observed object with respect to the current motor abilities. Arbib and coworkers (Fagg & Arbib, 1998) proposed the FARS model as a possible description of the computation in AIP/F5. They

did not however consider how affordances can be actually learned during the interaction with the environment. Learning and understanding affordances requires a prolonged developmental time, when information about the manipulative action (motoric and visual), about the target object, and about the goal of the action is integrated. Most of the F5 neurons have a pure motor response. However, approximately a third of them is also responsive to visual cues. In particular, the analysis of the responses of neurons in F5 allowed identifying two different classes with unique responses: they were called canonical and mirror neurons respectively. In short, canonical neurons respond both when the monkey is acting – e.g. grasping a peanut – and when fixating the same object – e.g. just watching the peanut on a tray. Mirror neurons respond similarly when the animal is actively involved in performing an action – e.g. grasping the same peanut as above – but they have a quite different visual response, in fact, they respond to the observation of a similar grasping action performed by somebody else.

The first step in our developmental model requires thus the construction of the F5 "canonical" responses. The rationale is that having an understanding of the possible actions that an object "affords" simplifies the problem of recognizing when we observe someone else acting on the same object. In practice, the link between actively executing an action and just watching the action execution is not only as simple as depicted here. It is worth noting here that key element to the formation of an association between seeing an action and recognizing it is not quite the specific kinematics of the action but rather the achievement of a specific goal. Biologically, mirror neuron responses are evoked only when the action has a visible goal (the object). Neurons remain silent if the experimenter only mimics the action without a visible goal (e.g. pretending to grasp a peanut that is not there).

Consequently, the next step along this hypothetical developmental route is to acquire the F5 mirror representation. We might think of canonical neurons as an association table of grasp/manipulation (action) types with object (vision) types. Mirror neurons can then be thought of as a second-level associative map which links together the observation of a manipulative action performed by somebody else with the neural representation of one's own action. Mirror neurons bring us to an even higher level of causal understanding. In this case the action execution has to be associated with a similar action executed by somebody else. The two events do not need to be temporally close to each other and arbitrary time delays might occur.

Nature of causation	*Main path*	*Function and/or behavior*
Direct causal chain	VC-VIP/LIP/7b-F4-F1	Reaching
One level of indirection	VC-AIP-F5-F1	Grasping
Complex causation involving multiple causal chains	VC-AIP-F5-F1+STs+IT	Mirror neurons, mimicry
Complex causation involving multiple instances of manipulative acts	STs+TE-TEO+F5-AIP(?)	Object recognition

Table 1 Degrees of causal indirection, localization and function in the brain.

As we will see, the next key aspect of the model is the "presence" of motor information into the decisional process. The question we might try to answer is whether this buys us anything. Intuitively we might think of motor signals as a sort of "reference frame". If we could map everything we see (our visual input) into this motor reference frame we would gain the ability

to reason invariantly from the point of view. It would not matter the point of view visual information was gained from, the resulting action could be mapped, rotated, transformed just by transforming the motor information. Furthermore, this hypothetical learning system would have straightforward path from observation (perception) to action. Motor information would be already there. Although reenacting the same trajectory is not quite enough for imitating a goal, it is already a good starting point.

2. Gesture recognition with motor representations

From the above discussion, two core elements of the prospective mirror neurons model emerge, namely, the use of motor information (or coding) also during the recognition of somebody else's actions and the use of object affordances (we provided support for the relevance of the target object during action execution).

In practice, many objects are grasped in very precise ways, since they allow the object to be used for some specific purpose or goal. A pen is usually grasped in a way that affords writing and a glass is held in such a way that we can use it to drink. Hence, if we *recognize* the object being manipulated, then recognition immediately provides some information about the most likely grasping possibilities (expectations) and hand appearance, simplifying the task of gesture recognition.

The affordances of the object possess an attentional-like[1] property because the number of possible (or likely) events is reduced. Affordances provide expectancies that can be used to single out possible ambiguities. This has clearly to be a module of our overall system architecture.

The common approach to recognition involves comparing acquired visual features to data from a training set. Differently, our approach is based on the use a *Visual-Motor Map* (VMM) to convert such measurements to a motor space and then perform the comparison/recognition in terms of motor representations. The advantage of doing this inference in motor space is two-fold. Firstly, while visual features can be ambiguous, we were able to show that converting these features to the motor space might reduce ambiguity. Secondly, since motor information is directly exploited during this process, imitative behaviors could be trivially implemented given that all the information/signals are already available.

To use motor representations for grasp recognition, we need to define *Visuo-Motor maps* to transform visual data onto motor information. The VMM can be learnt during an initial phase of self-observation, while the robot performs different gestures and learns their visual effects. The question that remains to be addressed is that of choosing what **visual features** to use. As we will focus on the classification and imitation of coarse gestures (e.g. power grasp and precision grip), we will rely on global appearance-based image methods. Together with the prior information provided by the "canonical neurons" (or their artificial implementation thereof), appearance based methods offer an easier, fast and more robust representation than point tracking methods. In the next section we will present a Bayesian approach for a gesture recognition that includes models of the *canonical* and *mirror* neurons, using visual appearance methods.

During the self-observation phase the machine could observe its own actions together with the hand's visual **appearance.** Since the machine is the initiator of the action it could possibly relate visual consequences of enacted movements with the corresponding motor commands.

[1] *Attention* in the sense of selecting relevant information out of a possibly much larger space.

Through this process the robot might perform many gestures and learn what effect they produce on visual re-afferences and on the objects in the environment.

Once the **VMM** has been estimated, one can transform actions observed from a specific point of view into an "internal" motor description that can either be used for recognition or to elicit the corresponding gesture (in an imitation setting).

In addition to learning the VMM, **self-observation** is also crucial for learning associations between classes of objects and classes of grasp types – i.e. the "canonical" neuron representation. Through object manipulation, one can possibly learn which grasp types are successful for a certain class of objects. When observing others manipulating objects, we can learn the most likely grasp types or specific use for a given class of objects. We refer to these combinations of action, manipulator's skills, and object as a type of ***affordances*** (J.J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, 1979). For recognizing gestures (grasp types), affordances provide prior information as to which gestures are more likely given a certain object (similarly to what the response of F5 canonical neurons together with area AIP provides to the monkey).

Within this model, recognition is only based on the final phase of the gesture – that is grasping. Figure 1, for example, illustrates the appearance of the hand during the approach phase (left panel), together with the final phase of two types of grasp that were used in this work: precision grip (rightmost panel) and power grasp (center panel).

Figure 1: Hand appearance during the approach phase (left), power grasp (center) and precision grip (right).

Gesture recognition has been addressed in the computer vision community in many different ways (J. Rehg and Takeo Kanade, *Visual tracking of high DOF articulated structures: an application to human hand tracking*, ECCV, 1994), (D.M. Gavrila. *The visual analysis of human movement: A survey*. CVIU 73(1), 1999). The difficulty of hand tracking and recognition arises from the fact that the hand is a deformable, articulated object that may display many different appearances depending on the configuration, viewpoint, and/or illumination. In addition, there are frequent occlusions between hand parts (e.g. fingers).

Modeling the hand as an articulated object in the 3D space implies extracting and tracking finger-tips, fingers, and other notable points in the image. This is, in general, quite difficult depending on the viewpoint or the image acquisition conditions. To overcome this difficulty, we exploit a more **iconic representation** of the hand's shape.

In summary, the main contributions of this work are the following:

- Recognition is based on motor information, since it is invariant to the viewpoint as perhaps suggested by the existence of mirror neurons.
- Object affordances are modeled in the overall classification scheme (analogue to canonical neurons).

- The Visuomotor Map (VMM) is learned during self-observation, while the system can generate a large variety of stimuli. The hand appearance is used directly in this process, avoiding an explicit model-based kinematic reconstruction of the posture of the hand.

- We show that the use of motor features allows better and more robust classification.

The next section describes a Bayesian approach to gesture recognition that includes a biologically compatible model of the role of *canonical* and *mirror* neurons and uses visual appearance methods. The approach leads to notable classification rates while classification occurs in motor space.

2.1. A Bayesian model for canonical and mirror neurons

Gesture recognition can be modeled in a Bayesian framework, which allows naturally combining *prior* information and knowledge derived from observations (likelihood). The role played by canonical and mirror neurons will be interpreted within this setting.

Let us assume that we want to recognize (or imitate) a set of gestures, G_i, using a set of *observed* features, ***F***. For the time being, these features can either be represented in motor space or in visual space (directly extracted from images). Let us also define a set of objects, O_k, present in the scene that represent the goal of a certain grasp actions.

Prior knowledge is modeled as a probability density function, $p(Gi|O_k)$, describing the probability of each gesture given a certain object. The observation model is captured in the *likelihood function*, $p(F|G_i,O_k)$, describing the probability of observing a set of (motor or visual) features, conditioned to an instance of the pair of gesture and object. The *posterior* density can be directly obtained through Bayesian inference:

$$\begin{aligned} p(G_i|F,O_k) &= p(F|G_i,O_k)p(G_i|O_k)/p(F|O_k). \\ \hat{G}_{MAP} &= arg\max_{G_i} p(G_i|F,O_k) \end{aligned} \quad (1)$$

where $p(F|O_k)$ is just a scaling factor that will not influence the classification. The *MAP* estimate, G_{MAP}, is the gesture that maximizes the posterior density in Equation (1). In order to introduce some temporal filtering, features of several images can be considered:

$$p(G_i|F,O_k) = p(G_i|F_t, F_{t-1}, ..., F_{t-N}, O_k)$$

where F_j are the features corresponding to the image at time instant *j*. The posterior probability distribution can be estimated using a naive approach, assuming independence between the observations at different time instants. The justification for this assumption is that recognition does not necessarily require the accurate modeling of the density functions. We thus have:

$$p(G_i|F_t, ..., F_{t-N}, O_k) = \prod_{j=0}^{N} \frac{p(F_{t-j}|G_i,O_k)p(G_i|O_k)}{p(F_{t-j}|O_k)}$$

2.1.1. The role of canonical and mirror neurons

The role of canonical neurons in the overall classification system lies essentially in providing affordances modeled as the *prior* density function $p(G_i|O_k)$ that, together with evidence from the observations, will shape the final decision. This density can be estimated by the relative frequency of gestures in the training set.

Canonical neurons are also somewhat involved in the computation of the likelihood function, since it depends both on the *gesture* and the *object*, thus implicitly defining another level of association between these two. Computing the likelihood function, $p(F| G_i, O_k)$, might be difficult because the shape of the data clusters might in fact be rather complex. We modeled these data clusters as mixtures of Gaussian random variables and the Expectation-Maximization algorithm was used to determine both the number of Gaussian terms and their coefficients.

a) Visual versus motor features

An image contains a large amount of highly redundant information. This allows for the use of methods whereby the image information is compacted in lower dimensional spaces, and consequently boosting computational performance. In the following, visual features consist of projections of the original image onto linear subspaces by using Principal Components Analysis (PCA). Initial input images were compressed to a 15 dimension coefficient vector.

Rather than representing the hand as a kinematic model built from tracked relevant points at the fingers and finger tips, we coded directly the image a set of templates projected in the low-dimensional subspace. This method has the advantage of being robust and relatively fast.

Under certain reasonable assumptions motor features might correspond to proprioceptive information about the hand/arm pose/motion. In our experiments (see also Mirror- “First Year Progress Report”, PPR-1) this was obtained through the use of the Mirror data-glove based setup that records *23* joint angles of someone's hand performing a grasping action synchronized with a sequence of images of the scene (binocularly).

b) Visuo-Motor Map

The ***Visuo-Motor Map*** transforms the PCA features defined in the previous section from the visual space into the motor space:

$$VMM : \mathbf{F}^V \rightarrow \mathbf{F}^M$$

As the structure of the transformation might be complicate, it was learned by means of a set multi-layer perceptron neural network: one complete neural network with a single output for each joint angle. For each network, i , the input consists of a 15-dimensional vector $\boldsymbol{F^V}$, which contains the PCA components extracted from the imaged hand appearance. The output consists of a single unit, coding the corresponding joint angle, $\boldsymbol{F^M}$. There are *5* neurons in the hidden layer.

We assumed that $\boldsymbol{F}^{V}$ is captured across many different view points. This is “large variance” could be generated theoretically during **self-observation** since a huge variety of hand configurations can be easily displayed. Otherwise, a view-point transformation could be used to pre-transform the visual data as for example in: (M. Cabido-Lopes and J. Santos-Victor. *Visual transformations in gesture imitation: What you see is what you do,* ICRA - International Conference on Robotics and Automation, Taiwan, 2003).

In our experiments, each neural network was trained with momentum and adaptive *back-propagation* with the data pre-processed to have zero mean and unitary variance. It converges to an error of *0.01* in less than *1000* epochs. Figure 2 shows trajectories (solid-line) for a joint angle of the little finger when performing several precision grips, and estimated position of the same joint angle as predicted from image data by means of the corresponding VMM. The dashed-line in the figure shows that the trajectory reconstructed through the neural-VMM is in very close agreement with the "true" values.

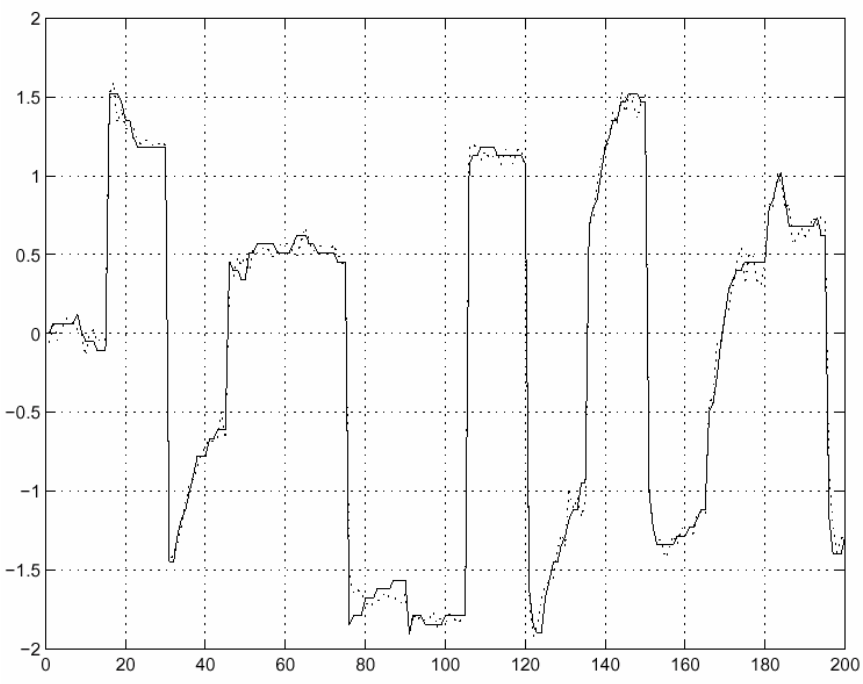

Figure 2: A sequence of several trials of a precision grip experiment. Solid line: original motor information. Dotted Line: reconstructed motor information using the Visual-Motor Map (VMM).

Even inside each grasp class variability is very large. This is due mainly to the difference in posture forced by the actual grasped objects, and it clearly illustrates how the visually observed features depend not only on the "grasp" type but also on the manipulated object (see Section 2.1.1 for discussion).

Finally, an additional aspect is that the VMM does not necessarily need to map into the hand joint space. In fact, motor features allowing a more compact representation could perhaps be used. This will be subject of further investigation in the future.

2.2. Experimental results

For the results presented here, we used a data set acquired with the data-glove setup (see Mirror – 1st Year Progress Report for details). Several subjects were asked to perform different types of grasp on a small set of objects. Each experimental trial began with the subject sitting on a chair and the hand on the table. Then, the subject was told to grasp the object that is in front of him (no specific instructions were given). The experiments include two types of grasp: **power grasp and precision grip**. Power grasp is defined when all the fingers and palm are in contact with the object. Instead, in precision grip, only the fingertips touch the object.

The three objects considered were a small sphere, a large sphere, and a box. The size of the small sphere is such that only precision grip is feasible in practice. The big sphere allows only power grasp. The box is ambiguous since it allows all possible grasps with different orientations.

Every experiment was repeated several times under varying conditions. The subject and the camera went around the table to cover a large set of points of view. Sequences were recorded binocularly although all experiments were performed monocularly. In total, we recorded the same grasp type from 6 different orientations (12 if we consider that we have a binocular vision). Motor information was recorded through the data-glove. From the 23 joint angles sequences we used only 15 values that correspond to all the joint angles of the fingers (3 for each finger). Finger's abduction and palm and wrist flexion were also available. They were not used in these experiments. Altogether the dataset contains 60 grasp sequences with 3 objects, 2 grasps with 6 different orientations.

Figure 3 shows sample images of the dataset acquired according to procedure just described. If we were to consider also the posture of the arm some of these images would not be actually realizable through self-observation only. Since we were only considering the hand, all postures in the database can be attained since moving the arm allows positioning according to all our recording viewpoints.

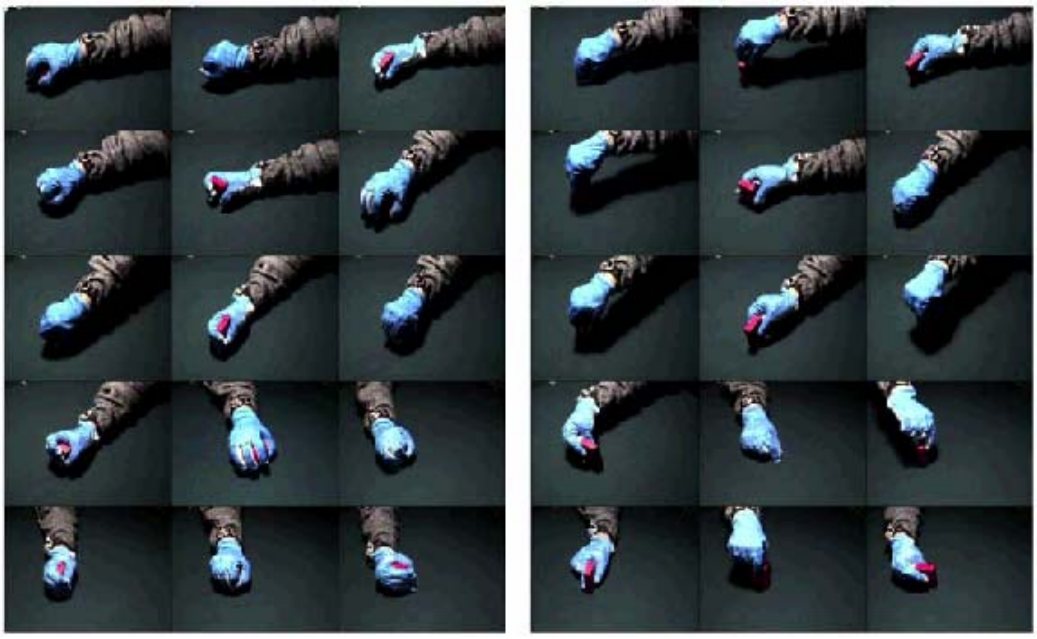

Figure 3: Data set illustrating some of the used grasp types: power (left) and precision (right). Altogether the tests were conducted using 60 sequences, from which a total of about 900 images were processed.

Every video sequence was automatically processed in order to segment the hand. First, a color-based clustering method, in the Y-Cr-Cb space, was applied to extract skin-colored pixels. The bounding box is determined based on the vertical/horizontal projections of the detected skin region. Finally, the hand is resized to a constant scale before applying the PCA. This approach yields uniformly scaled hand image regions. Figure 4 presents some segmentation results.

Figure 4: Segmentation results of scale-normalized hand regions automatically detected from color clustering.

Table 2 shows the classification rates obtained in a set of different experimental conditions. It allows us to compare the benefits of using the motor representation for classification as opposed to visual information only. The results shown correspond to the use of the ambiguous objects only when recognition is more challenging. We varied the number of viewpoints included in both the training and test sets so as to assess the degree of view invariance attained by the different methods.

In the first experiment, both the training and test sets correspond to a single viewpoint. Training was based on 16 grasp sequences, while test was done in 8 (different) sequences. The achieved classification rate was *100%*. The number of visual features (number of *PCA* components) was also tuned and the value of *5* provided good results. The number of modes (Gaussians in the mixture trained through the EM algorithm) was typically from 5 to 7.

The second experiment shows that this classifier is not able to naturally generalize to other viewpoints and/or camera positions. We used the same training set as in the first experiment, but the test-set was this time formed with image sequences acquired from four different camera positions. In this case, the classification rate is worse than random (*30%*). That is, visual information only cannot generalize well to other viewpoints.

In the third experiment, we added viewpoint variability in the training set. When sequences from all available camera positions were included in the training set the classification rate in the test set dropped to *80%*. While this is a more acceptable value, it is nevertheless significant lower from the desired 100%. This result shows that the change of the viewpoint introduces such non-trivial modifications in the hand appearance that classification errors occur.

The final experiment corresponds to the **mirror approach**. The system learns the VMM during the initial period of self-observation. Subsequently, the VMM is used to transform the (segmented) hand images into motor representation (joint angles) where the classifier is applied. High rates of classification were achieved (97%). Interestingly, the number of modes needed for the learning the conditional probabilities was only between one and two in this last case as opposed to 5-7, when recognition took place in the visual domain (experiments 1 to 3). This also clearly shows that mapping visual data into a specific motor representation helps the clustering of the data, as the latter is now intrinsically viewpoint invariant. Notice that viewpoint invariance is achieved even when the training set only contains sequences from a single viewpoint.

	Exp. I (visual)	Exp. II (visual)	Exp. III (visual)	Exp. IV (motor)
	Training			
# Sequences	16	24	64	24
View Points	1	1	4	1
Classif. Rate	100%	100%	97%	98%
# Features	5	5	5	15
# Modes	5-7	5-7	5-7	1-2
	Test			
# Sequences	8	96	32	96
View Points	1	4	4	4
Classif. Rate	100%	30%	80%	97%

Table 2. Grasp Recognition results. Notice the improvement obtained in the classification rate and viewpoint invariance due to the use of motor features.

These experiments show that a motor representation clearly describes the hand better when gesture recognition is the goal due to its inherent viewpoint independence. As only visual information is available during recognition the process greatly depends on the quality of the *VMM* approximation. Nonetheless, we believe that these results also validate the approach taken in estimating the VMM. For the case of only one camera position the quality obtained was more than acceptable with only 15 visual features. On the other hand, the use of motor features for recognition has the additional advantage of transforming imitation into a simpler problem since all the "reasoning" could be performed in motor terms. Figure 5 shows a simulated robotic hand imitating an observed gesture.

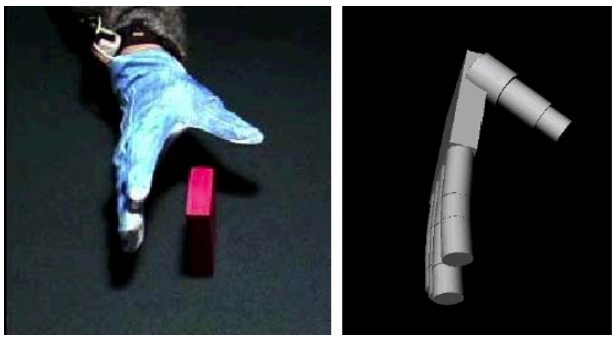

Figure 5: Reconstruction results of our model hand, obtained with the VMM

2.3. Conclusions

Although on a superficial reading it might seem that the Bayesian model encompasses all what it has to be said about mirror neurons, in fact it is substantially a supervised learning model. To relax the hypothesis of having to "supervise" the machine during training by indicating which action is which we need to remind what the evidence on mirror neurons tells us. First of all, it is plausible that the 'canonical' representation is acquired by self exploration and manipulation of a large set of different objects. F5 canonical neurons represent an association between objects' physical properties and the action they afford: e.g. a small object affords a precision grip, or a coffee mug affords being grasped by the handle. This understanding of object properties and the goal of actions is what can be subsequently factored in while disambiguating visual information. There are at least two level of reasoning: i) certain actions are more likely to be applied to a particular object – that is, probabilities can be estimated linking each action to every object, and ii) objects are used to perform action – e.g. the coffee mug is used to drink coffee. Clearly, we tend to use actions that proved to lead to

certain results or, in other words, we trace backward the link between action and effects: to obtain the effects apply the same action that earlier led to those effects.

Bearing this is mind, when observing some other individual's actions; our understanding can be framed in terms of what we already know about actions. In short, if I see someone drinking from a mug I can hypothesize a particular action (that I know already in motor terms) is used to obtain that particular effect (of drinking). This link between mirror neurons and the goal of the motor act is clearly present in the neurons' response. It is also the only possible way of autonomously learning a mirror representation. Technically speaking, the learning problem is still a supervised one but the information can now be collected autonomously. The association between the canonical response (object-action) and the mirror one (including vision of course) is made when the observed consequences (or goal) are recognized as similar in the two cases – self or others acting. Similarity can be evaluated following different criteria ranging from kinematic (e.g. the object moving along a certain trajectory) to very abstract (e.g. social consequences such as in speech).

In summary, we have presented a framework for gesture recognition based on a model for *canonical and mirror neurons* in general accordance with what is known about the physiology this specific brain area. In this model, canonical neurons provide prior information in terms of object affordances which narrows the attentional span of the system, allowing unlikely gestures or hand appearances to be immediately discarded. The fact that, despite being located in a motor area of the brain, mirror neurons are active during both the execution and recognition of an action, suggests that recognition takes place in motor terms rather than in visual space.

We proposed a Bayesian formulation where all these observations are taken into account. We described how to estimate the prior densities and likelihood functions directly from the data. A visuomotor map is used to transform image data into the motor data. The VMM is supposedly learnt during an initial period of self-observation. The use of the VMM proved to be good for classification and additionally as an extra advantage it bears the potential of simplifying gesture imitation.

Although hand posture recognition is in general quite difficult, grasp classification benefits from using this extra information. Temporal integration and object-related cues are very useful for recognition. Occlusions and ambiguous positions of the hand can also be solved if temporal information is included. The observation of a given object "conveys" information about the possible and the most probable grasp types for that object class. Expectations of the hand appearance can also be created.

The results show that it is possible to achieve almost *100%* recognition rates. Notably, the approach overcomes the need for complex schemes for detecting and tracking the fine details of the hand on the video sequences.

3. Additional references

Fadiga L, Craighero L, Buccino G, Rizzolatti G. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur J Neurosci* 2002;15:399-402.

Fagg AH, Arbib MA. Modeling parietal-premotor interactions in primate control of grasping. *Neural Netw* 1998;11:1277-303.

Kerzel D, Bekkering H. Motor activation from visible speech: evidence from stimulus response compatibility. *J Exp Psychol Hum Percept Perform* 2000;26:634-47.

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech

code. *Psychol Rev* 1967;74:431-61.

Liberman AM, Mattingly IG. The motor theory of speech perception revised. *Cognition* 1985;21:1-36.

Liberman AM, Whalen DH. On the relation of speech to language. *Trends Cogn Sci* 2000;4:187-96.

Wohlschlager A, Bekkering H. Is human imitation based on a mirror-neurone system? Some behavioural evidence. *Exp Brain Res* 2002;143:335-41.