IROS Workshop on Robot Programming by Demonstration,
Las Vegas, SA, Oct. 31st, 2003.

Motor Representations for Hand Gesture Recognition and Imitation

Manuel Cabido Lopes‡† José Santos-Victor†
{macl,jasv}@isr.ist.utl.pt

† Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisboa, Portugal

‡ Escola Superior de Tecnologia
Instituto Politécnico de Setúbal
Setúbal, Portugal

***Abstract*— We present an approach for grasp recognition and imitation based on models for canonical and mirror neurons, recently found in neurophysiological experiments. *Canonical Neurons* seem to code object affordances, e.g. possible ways of grasping. *Mirror Neurons* code goal directed tasks, like precision or power grasping of an object. The major feature of this neuron population is the use of motor information in the recognition step.**

We propose a Bayesian approach that encompasses all these aspects. Recognition is performed in the motor space and we solve the problem of getting motor information while observing another person. Our approach avoids the complexity of other approaches based on the $3D$ reconstruction of the hand from images, considering that the hand is a multi-articulated object subject to frequent occlusions.

The results obtained illustrate the benefits of designing artificial machines inspired on biological findings and hypotheses, while at the same time, offering robotics technologies as a testbed for such hypotheses.

I. Introduction

Despite being often ignored, an artificial system can retrieve a large amount of knowledge, simply by looking at other individuals, humans or robots working in the same area. In fact, similarly to human infants, a robot could learn significant information if it were able to recognize and imitate what the others are doing.

The long-term goal of our work is two-fold. On one hand, we want to develop methodologies whereby a system can learn how to perform complex tasks through imitation. On the other hand, our approach relies on recent findings in neuroscience and developmental psychology, hoping to contribute to a better understanding of the fundamental problem of how humans imitate each other and how they recognize and understand the observed behavior and actions.

This work is motivated by the recent discovery of *mirror* and *canonical* neurons [1], [2] in the F5 area of the macaque's brain. These neurons discharge during the execution of hand/mouth movements. In this paper we will focus on hand gestures, often referred to as grasp actions or grasps.

In spite of their localization in a pre-motor area of the brain, *mirror* neurons fire not only when the animal performs a specific goal-oriented grasping task, but also when observing that same action being performed by another individual. Canonical neurons [3] have the intriguing characteristic of responding when objects, that afford a *specific* type of grasp, are present in the scene, even if the grasp action is not performed or observed.

By establishing a direct connection between gestures performed by a subject and similar gestures performed by others, mirror neurons may be intimately connected to the ability to imitate found in some animal species [2], establishing an implicit level of communication between individuals.

The discovery of mirror neurons raises the fundamental question of understanding the role of motor information for "visual" gesture recognition, and how can it be facilitated by the fact that we know how to perform those gestures. This is clearly distinct from most approaches for gesture recognition, where only visual information is involved. In our paper, instead, recognition is performed in the motor space and we show that it really simplifies the problem by affording a larger degree of invariance to viewpoint modifications.

Visuo-motor representations can be acquired during extensive periods of self-observation, as well as from observing other individuals. The subject can learn how to perform various gestures and what effect they produce on the visual space and on world objects. Observation can be useful in different ways:

(i) By manipulating objects, one can learn which grasp types are successful for a certain class of objects. Also, if we observe *other* people manipulating objects, we can learn the most likely grasps or functions, for a given class of objects. We will refer to these grasp types or functions as a particular type of *affordances* [4] associated to a certain object. For recognizing gestures, affordances provide prior information as to which gestures are more likely, when acting upon a certain object class. This is a possible interpretation of the role played by *canonical* neurons in the overall process of gesture recognition and imitation.

(ii) When observing one's own gestures, the hand appearance can be estimated and directly related with the corresponding motor commands. We will refer to this association as the *Visuo-Motor Map* (VMM). Once the VMM has been estimated, one can transform views of observed gestures to motor descriptions that can either be used for recognition or to elicit the corresponding (imitated) gesture.

Grasp actions are usually partitioned into the *transport* and *grasp* phases [5]. During the transport phase, the hand moves towards the target and the grasp phase corresponds to the final segment, immediately before and after touch. It has been shown that the transport phase can change significantly, according to the particular grasp type that is performed in the end of the movement. However, it seems that this information is not used by humans for gesture recognition. Mathematically, this can be interpreted as poor (uncertain) predictive capabilities, as it is only in the final (grasp) part of the gesture that recognition takes place.

Similarly, in our work, recognition will only be based on the grasp phase of the gesture. Figure 1 illustrates the hand appearance during the approach phase, together with the final phase of two broad classes of grasps that will be used in this work: precision grip and power grasp.

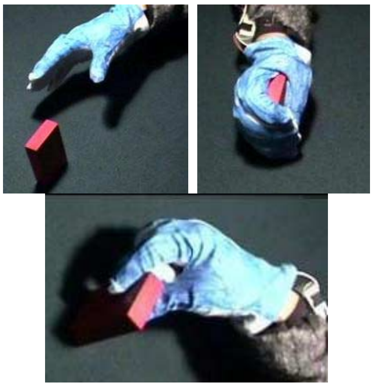

Fig. 1. Hand appearance during the approach phase (left), power grasp (center) and precision grip (right).

Gesture recognition has been addressed in the computer vision community in many different ways [6]- [11]. 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 its configuration, viewpoint 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 taken viewpoints and image acquisition conditions. To overcome this difficulty, we exploit more iconic representations for the hand shape, that are commonly believed to be used by humans when recognizing (known) gestures. Also, our approach will make use of motor information, since it is invariant to the viewpoint, as suggested by the existence of mirror neurons.

The recognition of other individuals and imitation are always intertwined and imitation mechanisms can allow better recognition. Several works suggest imitation as a very important paradigm for programming robots [12]. The imitation mechanism can be better understood if some computational models are developed that emulate the brain. An important work [13] modeled several components of the brain, presumably involved in imitation. This work was extended for the case of grasp recognition (mirror neurons) [14] and an implementation with video data was used. Although good results were obtained, the visual features used are very difficult to extract, which makes it difficult to use in real world conditions. For the case of learning motor skills, [15] presents a biologically motivated architecture. This systems works with real data and allows learning of repetitive patterns and precise movements for grasp and reaching. Several other works used biological principles in order to achieve imitation [16] - [19]. Instead of mapping different brain regions and modeling the way they function, our goal in this work is to investigate the mathematical properties of some mechanisms, hypothetically developed through evolution, that allow to recognize and imitate others. Most of the cited works, although recognizing the complexity, simplified the perception either using markers or reducing the possible postures of the demonstrator.

Imitation can be done with simple mechanisms. If the motor system is activated in order to reduce some error function derived from the visual perception, imitation emerges. This *homeostatic* behavior was used in [20], [21] in order to imitate hand trajectories.

A discussion between passive and active imitation is present in [22]. The first case relies on a perceive-recognize-reproduce sequence, while the second uses a map from perception to a set of behaviors. In the case of gesture imitation the traditional way would be to have a visual gesture classifier and then generate a similar problem. Active imitation needs a direct link from perception to action. In the cited work this two modules are mixed allowing imitation of known and unknown actions. In our work, for the case of visual features, some action generation would be necessary after the classification (passive imitation), for the case of motor features the action generation is temporally mixed with recognition (active imitation). We show, in this work, that the use of motor features allows better and more robust classification and imitation. Although for low-level imitation the map allows for imitation of unknown sequences, for grasp actions only known gestures can be imitated.

As a final comment, we would like to remark that, to consider gestures performed by the entire arm, we would need to include some sort of visual transformation to deal with the problem of viewpoint shape variance [23]. For hand movements, our approach is invariant to large variety of view points. Also, during self-observation, the system can generate a large variety of hand visual stimuli that will be used for the construction of visuo-motor maps. The viewpoint transformation for arm gestures is specifically addressed in [24].

In the next section, we will detail the main structure of our approach. In Section III we describe our Bayesian framework for grasp actions recognition and imitation, that involves models of canonical and mirror neurons. We detail how to learn the prior densities and likelihood function from data and how to estimate a visuo-motor map (VMM), using data acquired during self-observation. As suggested by studies of mirror neurons, recognition takes place in motor variables rather than visual. Finally we present some experimental results in Section IV and discuss the main conclusions in Section V.

II. Approach

Gesture recognition is, in general, a complex task [6]-[11]. Traditional approaches imply performing full $3D$ reconstruction of the hand, followed by a pose classifier. To make the 3D reconstruction, it is necessary to track the fingertips, while handling the multiple occlusions generated by the complex hand motion. State-of-the-art algorithms rely on good initial estimates and require sophisticated kinematic models of the hand.

The approach we propose here differs from other works in several ways: (i) use of object affordances in the recognition process (canonical neurons); (ii) recognition is performed in the motor space (mirror neurons) and (iii) use of global descriptors of the hand appearance.

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

This link between objects and their affordances is possibly played in the macacque's brain by the *canonical neurons* of the area F5. If two objects can be grasped in the same way, the same neurons will fire when either object is presented. The affordances of the object have thus an attention property because the number of possible (or likely) events are reduced, thus overcoming possible ambiguities. This will be the first module of our overall system architecture.

We have seen in the previous section that, in spite of their localization in a motor area of the brain, mirror neurons are also active during pure visual (recognition) tasks. When observing someone doing a familiar gesture, the same neurons, that would fire when performing this same gesture, become active. It has also been shown that lesions in the motor part of the brain do affect recognition capabilities.

This observation suggests that the motor system responsible for triggering an action is also involved when recognizing that same action, leading to the question of how to use motor information for recognition. Since during the recognition, only visual information is available, the solution lies in making a transformation from visual to the motor space, where recognition will eventually be done.

The common approach to recognition involves comparing acquired visual features to data from a training set. Instead, we will first use a *Visual-Motor Map* to convert such measurements to the motor space and then perform the comparison/recognition in terms of motor variables.

The advantage of doing this inference in the motor space is two-fold. Firstly, while visual features can be ambiguous, we show that converting these features to the motor space may reduce ambiguity. Secondly, as the motor information is directly exploited during this process, imitation can be done immediately, as all the information/signals are readily available.

To use motor representations for grasp recognition, we need to define *Visuo-Motor maps* (VMMs) 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 its visual effects.

The question that remains is that of choosing what visual features to use. As we will focus on the classification and imitation of coarse gestures (power grasp and precision grip), we will rely on global appearance-based image methods. Together with the prior information provided by the canonical neurons, 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. The approach leads to excellent classification rates and classification occurs in the motor space.

III. A Bayesian model for Canonical and Mirror Neurons

Gesture recognition can be modeled in a Bayesian framework, which allows to naturally combine *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 the motor space (as mirror neurons seem to do) or in the visual space (directly extracted from images). Let us also define a set of objects, O_k, present in the scene, that represents the goal of a certain grasp action.

The prior information is modeled as a probability density function, $p(G_i|O_k)$, 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 gesture and object. The *posterior* density can be directly obtained through Bayesian inference:

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

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 then have:

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

A. The role of canonical neurons

The role of canonical neurons in the overall classification system lies essentially in providing the 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 *object*, thus implicitly defining another level of association between these. Computing the likelihood function, $p(F|G_i, O_k)$, is more elaborated and is described in detail in Section III-B.

B. Estimating the likelihood function

As the likelihood function may correspond to a complex distribution, it will be modeled it by a Gaussian mixture, which is fitted to data points. In what follows we will describe the process of fitting a mixture model to a density, $p(x)$:

$$p(x) = \sum_{j=1}^{K} \pi_j \, p(x|j),$$

where $p(x|j) \sim N(\mu_j, \sigma_j)$, is a Gaussian distribution. For a proper probability density function, we need to ensure that $\sum_{i=1}^{K} \pi_i = 1, \; \pi_i \geq 0$.

The Expectation-Maximization (EM) algorithm can be used to estimate the parameters μ_i, σ_i, π_i that best fit the data. The main problem with this solution is the necessity of knowing in advance the number of kernels, K. In [25], [26] there is the option of modifying the number of Gaussian kernels used to best fit the data. The number of kernels can be increased during the learning process, based on a new measure designated as the total *kurtosis*, $\mathcal{K}$:

$$\mathcal{K} \triangleq \int_{-\infty}^{\infty} \left(\frac{x - \mu_j}{\sigma_j}\right)^4 \frac{p(j|x)}{\pi_j} p(x)dx - 3$$

The *kurtosis* measures how far a distribution is from a Gaussian and it is zero for a Gaussian function. If the *kurtosis* is not close to zero for a given kernel, it means that the data are not Gaussian and this kernel is split. On the other hand, the number of kernels can sometimes be reduced (merged) in order to reduce the model complexity. A "closeness" metric between two kernels, can be defined as follows:

$$d(p_1, p_2) = \frac{\prod_{x_i \in X_1} p_2(x_i) \prod_{x_i \in X_2} p_1(x_i)}{\prod_{x_i \in X_1} p_1(x_i) \prod_{x_i \in X_2} p_2(x_i)}$$

where X_i stands for the data points used for the estimation of $p_i(x)$.

Two different kernels can be merged if the distance between them is sufficiently small. At the end of this process, we have an estimate of the likelihood function directly from the data, without imposing a particular structure for the underlying distribution. An important point worth mentioning is that this method can cope with clusters that with very irregular shapes and that it automatically adapts to the shape of such clusters..

C. Mirror Neurons

The classification done by our system as several properties similar to the mirror neurons. In this section we will see how to account to some of the observations regarding this neurons into our Bayesian framework. We must first consider a Visuo-Motor Map that transforms observed visual data, to the motor representations that will eventually drive the recognition process.

1) 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, thus boosting computational performance. Our visual features consist of projections of the original image onto linear subspaces, using Principal Components Analysis (PCA). As a result, our images can be compressed to a 15 dimension coefficient vector.

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

In a real (robotic or living) system, motor features would correspond to proprioceptive information about the hand/arm pose/motion. In our experiments [27], this is obtained through the use of a data-glove that records 23 joint angles of someone's hand performing gestures.

2) Visuo-Motor Map: As referred previously, the *Visuo-Motor Map* must transform the features defined in the previous section, from the visual space to the motor space.

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

As the structure of the transformation is quite complex, it was learned with a Multi-Layer Perceptron, for each joint angle. For each network, i, the input consists of a 15-dimensional vector $\mathbf{F}^V$, which are the PCA components of the imaged hand appearance. The output consists of a single unit, coding the corresponding joint angle, $\mathbf{F}_i^M$. There are 5 neurons in the hidden layer.

We assume that $\mathbf{F}^V$ is captured across many different view points. This is possible to generate during self-observation since a huge variety of hand configurations can be easily displayed. Otherwise, some kind of view-point transformation is needed to pre-transform the visual data [24].

The VMM can lead to impossible (temporal) trajectories, as errors in input frames can cause discontinuities in the motor space. To overcome this problem, continuity is imposed in the motor data through a first-order dynamic filter.

Each 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.

It is noticeable that, even inside each grasp class, the variability is very large. This is due to the differences between the grasped objects, and illustrates how the observed features depend not only on the ”grasp” type but also on the manipulated object (see Section III-A for discussion). The dashed-line in the figure shows that the trajectory reconstructed through the neural-VMM is in a very close agreement with the ”true” values.

A final aspect worth mentioning is that the VMM can be learned very naturally during an initial phase, when a system (natural or artificial) performs hand/arm gestures and observes the (visual) consequences of such gestures. During self-observation, both proprioceptive (motor) and visual data are present and the association can be established. As an additional aspect, self-observation would allow the system to search and tune the most interesting visuo-motor features such that a more compact representation could be used.

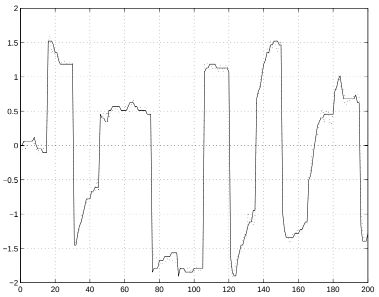

Fig. 2. A sequence of several trials of a precision grip experiment. Solid line: original motor information. Dotted Line: reconstructed motor information using the Visual-Motor Map (VMM)

IV. Experimental results

For the results presented here, we use a data set prepared at the Lira Lab, University of Genova, [27], with a specially designed experimental setup. Several subjects were asked to perform different types of grasp on different objects. The experiment begins with the subject sitting in a chair, with the hand on the table. Then, the subject is told to grasp the object that is in front of him.

The experiments include two types of grasp: power grasp and precision grip. Power grasp is defined when all the hand fingers and palm are in contact with the object. Instead, in precision grip, only the fingertips touch the object.

We considered three different objects: a small sphere, a large sphere and a box. The small sphere is sufficiently small so that only precision grip is allowed. The big sphere allows only power grasps. The box is ambiguous because it allows all possible grasps with different orientations.

Every experiment was repeated several times under varying conditions. The subject and the camera go around the table to cover a large variation of viewpoints. To record the sequences we use a stereo-pair. In total, we record the experiments from 6 different azimuths (12 if we consider the stereo-pair). In order to record the motor information, a data-glove [28], capable of recording 23 values of the hand configuration, is used. We used the first 15 values that correspond to all the joint angles (3 for each finger). Finger's abduction and palm and wrist flexion were also available but they were not used in the recognition. Altogether the data-set contains sixty grasp sequences with three objects, two grasps with six different azimuths.

Figure 3 shows sample images of the data set acquired according to process just described. Notice the multiplicity of grasps and view points. Some external observations of an arm are impossible to have when looking to one's arm. For the case of an hand this is not the case because moving the arm allows observing the hand from all viewpoints. Because of this some arm images that might appear impossible have realistic hand observations.

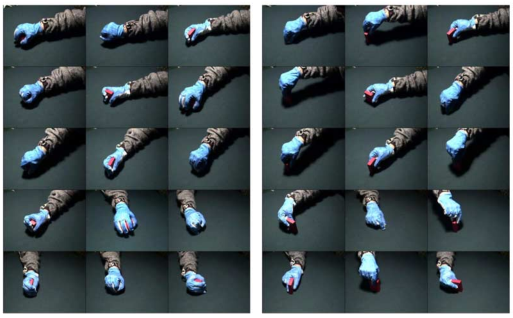

Fig. 3. Data set illustrating some of the used grasp types: power (left) and precision (right). Altogether the tests were conducted using 60 sequences, from which a total of about 900 images were processed.

Every video sequence is 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 for a constant scale before applying the PCA. This approach yields uniformly scaled hand image regions. Figure 4 presents some segmentation results.

Fig. 4. Segmentation results of scale-normalized hand regions automatically detected from colour clustering.

Table I shows the obtained classification rates. It allows us to compare the benefits of using motor representations for recognition as opposed to visual information only. The results shown correspond to the use of the ambiguous objects only, when the 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 one single view point. 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) were typically from 5 to 7.

The second experiment shows that this classifier is not able to generalizes to other view points / camera positions. We used the same training-set as in $Exp.I$, but the test-set is formed with image sequences acquired with 4 different camera positions. In this case, the classification rate is worse than random (30%).

In the third experiment, we added view point variability in the training set. When sequences from all camera positions are included in the training-set, the classification rate in the test-set drops to 80%. While this is a more acceptable value, it is nevertheless a significant drop from the desired 100%. This result shows that the view point variation introduces such challenging modifications in the hand appearance that classification errors occur.

The final experiment corresponds to the main approach proposed in this paper. The system learns a visuo-motor map during an initial period of self-observation. Then, the VMM is used to transform the (segmented) hand images to motor information, where classification is conducted. A very high degree of classification was achieved (97 %). Interestingly, the number of modes need for the learning is between 1-2 in this case as opposed to 5-7, when recognition takes place in the visual domain. This also shows that mapping visual data to motor representations, helps clustering the data, as it is now view-point invariant.

Notice that view-point invariance is achieved when the training set only contains sequences from one single view point.

TABLE I

GRASP RECOGNITION RESULTS. NOTICE THE GAIN OBTAINED IN THE CLASSIFICATION RATE AND VIEWPOINT INVARIANCE DUE TO THE USE IF MOTOR FEATURES.

	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%

These experiments show that motor representations describe the hand better, for gesture recognition, due to the inherent viewpoint independence. As only visual information is available during recognition, the process greatly depends on the *VMM*. The results also validate our approach to estimate the VMM. For the case of only one camera position the quality obtained was very good, with 15 visual features.

The use of motor features for the recognition, has the additional advantage of making imitation a straight forward process, as all the reasoning is performed in motor terms. Figure 5 shows a hand imitating an observed gesture.

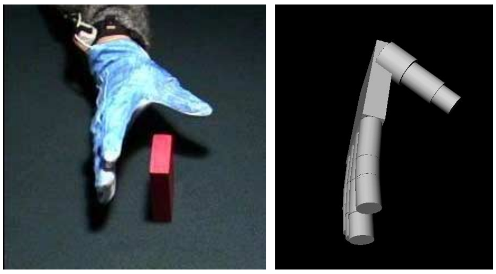

Fig. 5. Reconstruction results of our model hand, obtained with the VMM

V. Conclusions

Neurophysiology can provide many useful ideas for engineers to build more efficient artificial systems. On the other hand, designing artificial systems, grounded on such biological principles, is a valuable means of validating hypotheses or theories in biology.

In this work we propose a framework for gesture recognition based on a model for *canonical and mirror neurons*, that seem to play a fundamental role for grasp recognition or imitation in primates.

Canonical neurons provide prior information in terms of object affordances which narrows the attention span of the system, since very unlikely gestures or hand appearances can be discarded immediately . 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, suggest that recognition takes place in the motor space rather than on the visual space.

We propose a Bayesian formulation where all these observations are taken into account. We describe how to estimate the prior density and likelihood functions directly from the data. A Visuo-Motor Map is used to transform image data to the motor space, and is learnt during an initial period of self-observation. The use of the VMM is good for the classification and, as an extra advantage, gives the possibility of doing gesture imitation directly.

Although hand posture recognition is in general quite difficult, grasp classification benefits from using extra information. Temporal integration and object-related cues are very useful for recognition. Occlusions and ambiguous positions of the hand can also be solved with temporal information. 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 100% recognition rates based on this approach. Notably, we avoid using complex schemes for detecting and tracking fine details of the hand on a video sequence. Rather, we rely on the global hand appearance for this purpose.

In our opinion, the results obtained are an encouraging step in the endeavor of understanding the biological grounding of imitation and, at the same time, develop the principles to build more performing and robust machines, able to cope with complex tasks and to interact with humans.

VI. Acknowledgments

The experimental data set was prepared by Matteo Schenatti, Lorenzo Natale, Giorgio Metta and Giulio Sandini [27], Lira-Lab, University of Genova. This work was (partially) supported by the portuguese Fundação para a Ciência e Tecnologia (FCT), Programa Operacional Sociedade de Informação (POSI) in the frame of QCA III and the EU-Project IST-2000-28159, Mirror, "Mirror Neurons for Recognition", www.lira.dist.unige.it/projects/mirror/.

VII. References

[1] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Visuomotor neurons: ambiguity of the discharge or 'motor' perception? *International Journal of Psychophysiology*, 35, 2000.

[2] V.S. Ramachandran. Mirror neurons and imitation learning as the driving force behind the great leap forward in human evolution. *Edge*, 69, June 2000.

[3] A. Murata, L. Fadiga, L. Fogassi, V. Gallese, V. Raos, and G. Rizzolatti. Object representation in the ventral premotor cortex (area f5) of the monkey. *Journal of Neurophysiology*, 78(4):2226–2230, October 1997.

[4] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[5] L. Fogassi, V. Gallese, G. Buccino, L. Craighero, L. Fadiga, and G. Rizzolatti. Cortical mechanism for the visual guidance of hand grasping movements in the monkey: A reversible inactivation study. *Brain*, 124(3):571–586, March 2001.

[6] James M. Rehg and Takeo Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *ECCV (2)*, pages 35–46, 1994.

[7] Ying Wu and Thomas S. Huang. Capturing articulated human hand motion: A divide-and-conquer approach. In *ICCV (1)*, pages 606–611, 1999.

[8] Michael J. Black and Allan D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *ECCV (1)*, pages 329–342, 1996.

[9] D. M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, 1999.

[10] James M. Rehg and Takeo Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, pages 612–617, 1995.

[11] Ying Wu and Thomas S. Huang. View-independent recognition of hand postures. In *CVPR*, pages 88–94, June 2000.

[12] S. Schaal. Is imitation learning the route to humanoid robots. *Trends in Cognitive Sciences*, 3(6), 1999.

[13] Andrew H. Fagg. *A Computational Model of the Cortical Mechanisms Involved in Primate Grasping*. PhD thesis, University of Southern California, 1996.

[14] Erhan Oztop. *Modeling the Mirror: Grasp Learning and Action Recognition*. PhD thesis, University of Southern California, August 2002.

[15] M. A. Arbib, A. Billard, M. Iacoboni, and E. Oztop. Synthetic brain imaging: grasping, mirror neurons and imitation. *Neural Networks*, 13:975–997, 2000.

[16] Aude Billard and Maja J. Matarić. A biologically inspired robotic model for learning by imitation. In *International Conference on Autonomous Agents*, Barcelona, 2000.

[17] Aude Billard. Learning motor skills by imitation: A biologically inspired robotic model. *Cybernetics and Systems*, 32:155–193, 2001.

[18] Aude Billard and Stephan Schaal. Robust learning of arm trajectories through human demonstration. In *International Conference on Intelligent Robots and Systems*, pages 734–739, Maui, Hawaii, USA, 2001.

[19] Maja J. Matarić. Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics. In C. Nehaniv and K. Dautenhahn, editors, *Imitation in Animals and Artifacts*. MIT Press, 2000.

[20] P. Andry, P. Gaussier, S. Moga, J.P. Banquet, and J. Nadel. Learning and communication in imitation: An autonomous robot perspective. *IEEE Transaction on Systems, Man and Cybernetics, Part A*, 31(5):431–444, September 2001.

[21] P. Andry, P. Gaussier, and J. Nadel. From sensori-motor development to low-level imitation. In *2nd International Workshop on Epigenetic Robotics*, pages 7–15, 2002.

[22] Yiannis Demiris and Gillian Hayes. Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model. In K. Dautenhahn and C. Nehaniv, editors, *Imitation in Animals and Artifacts*. MIT Press, 2002.

[23] J.S. Bruner. Nature and use of immaturity. *American Psychologist*, 27:687–708, 1972.

[24] Manuel Cabido-Lopes and José Santos-Victor. Visual transformations in gesture imitation: What you see is what you do. In *International Conference on Robotics and Automation*, Taiwan, 2003.

[25] Paul M. Baggenstoss. Statistical modeling using gaussian mixtures and hmms with matlab. http://www.npt.nuwc.navy.mil/Csf/htmldoc/pdf/.

[26] N. Vlassis and A. Likas. A kurtosis-based dynamic approach to gaussian mixture modeling. *IEEE Trans. Systems, Man, and Cybernetics, Part A*, 29:393–399, 1999.

[27] Matteo Schenatti, Lorenzo Natale, Giorgio Metta, and Giulio Sandini. Object grasping data-set. Lira Lab, University of Genova, Italy, 2003.

[28] CyberGlove. http://www.immersion.com.