Proceedings of the 2003 IEEE International Conference on Robotics and Automation, ICRA, Taiwan, September 2003.

Visual Transformations in Gesture Imitation: what you see is what you do

Manuel Cabido Lopes José Santos-Victor

Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisbon, Portugal
{macl,jasv}@isr.ist.utl.pt

Abstract

We propose an approach for a robot to imitate the gestures of a human demonstrator. Our framework consists solely of two components: a *Sensory-Motor Map* (SMM) and a *View-Point Transformation* (VPT). The SMM establishes an association between an arm image and the corresponding joint angles and it is learned by the system during a period of observation of its own gestures. The VPT is widely discussed in the psychology of visual perception and is used to transform the image of the demonstrator's arm to the so-called ego-centric image, as if the robot were observing its own arm. Different structures of the SMM and VPT are proposed in accordance with observations in human imitation. The whole system relies on monocular visual information and leads to a parsimonious architecture for learning by imitation. Real-time results are presented and discussed.

1 Introduction

The impressive advance of research and development in robotics and autonomous systems in the past years has led to the development of robotic systems of increasing motor, perceptual and cognitive capabilities.

These achievements are opening the way for new application opportunities that will require these systems to interact with other robots or non technical users during extended periods of time. Traditional programming methodologies and robot interfaces will no longer suffice, as the system needs to learn to execute complex tasks and improve its performance through its lifetime.

Our work has the long-term goal of building sophisticated robotic systems able to interact with humans or other robots in a natural and intuitive way. One promising approach relies on imitation whereby a robot could learn how to handle a person's private objects by observing the owner's behavior, over time.

Learning by imitation is not a new topic and has been addressed before in the literature. This learning paradigm has already been pursued in humanoid robotic applications [1] where the number of degrees of freedom is very large, tele-operation [2] or assembly tasks [3]. Most published works, however, describe complete imitation systems but focus their attention on isolated system components only, while we describe a complete architecture.

We will concentrate on the simplest form of imitation that consists in replicating the gestures or movements of a demonstrator, without seeking to understand the gestures or the action's goal. In the work described in [4], the imitator can not only replicate the gestures but also the dynamics of a demonstrator, but it requires the usage of an exoskeleton to sense the demonstrator's behavior. Instead, our approach is exclusively based on vision.

The motivation to use visual information for imitation arises from the fact that many living beings - like humans - resort to vision to solve an extremely large set of tasks. Also, from the engineering point view, video cameras are low-cost, non invasive devices that can be installed in ordinary houses and that provide an enormous quantity of information, specially if combined with domain knowledge or stereo data.

Interestingly, the process of imitation seems to be the primary learning process used by infants and monkeys during the first years of life. Recently, the discovery of the *mirror neurons* in the monkey's brain [5, 6] has raised new hypotheses and provided a better understanding of the process of imitation in nature. These neurons are activated both when a monkey performs a certain action and when it sees the same action being performed by a demonstrator or another monkey.

Even if the role of these neurons is not yet fully understood, a few important conclusions can nevertheless be drawn. Firstly, mirror neurons clearly illustrate the intimate relationship between perception and action. Secondly, these neurons exhibit the remarkable ability of "recognizing" certain gestures or actions when seen from very different perspectives (associating gestures performed by the demonstrator to the subject's own gestures).

One of the main contributions of this paper is related to this last observation, that is illustrated in Figure 1. We propose a method that allows the system to "rotate" the image of gestures done by a demonstrator (allo-image) to the corresponding image (ego-image) that would be obtained if those same gestures were actually performed by

the system itself. We call this process the *View-Point Transformation* (VPT). Surprisingly, in spite of the importance given to the VPT in psychological studies [7], it has received very little attention from other researchers in the field of visual imitation.

Figure 1: Gestures can be seen from very distinct perspectives. The image shows one's own arm performing a gesture (ego-image) and that of the demonstrator performing a similar gesture (allo-image).

One of the few works that dealt explicitly with the VPT is [8]. However, instead of considering the complete arm posture, only the mapping of the end-effector position is done. The map between the allo and ego image is performed using epipolar geometry, based on a stereo camera pair.

Other studies addressed this problem in an implicit and superficially way. A mobile robot capable of learning the policy followed by another mobile vehicle is described in [9]. Since the system kinematics is very simple, the VPT corresponds to a transformation between the views of the two mobile robots. This is achieved in practice by delaying the imitator's perception until it reaches the same place as the demonstrator, without focusing the process of VPT. The work described in [10] has similar objectives to our own research and allows a robot to mimic the "dance" of an Avatar. However, it does not address the VPT at all, and a special invasive hardware was used to perform this transformation. Instead, we present a simple architecture for imitation which carefully addresses the fundamental process of *View-Point Transformation*.

The VPT allows the robot to map observed gestures to a canonical point-of-view. The final step consists in transforming these mapped features to motor commands, which is referred to as the *Sensory-Motor Map* (SMM). Our complete architecture for imitation is shown in Fig. 2.

The Sensory-Motor Map can be computed explicitly if the parameters of the arm-hand-eye configuration are known a priori but - more interestingly - it can be learned

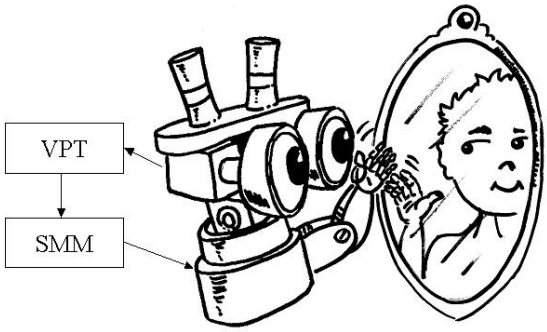

Figure 2: The combination of the Sensory-Motor Map and the View-Point Transformation allow the robot to imitate the arm movements executed by another robot or human.

from observations of arm/hand motions. Again, biology can provide relevant insight. The Asymmetric Tonic Neck reflex [11] forces newborns to look to their hands, which allows them to learn the relationship between motor actions and the corresponding visual stimuli.

Similarly, in our work the robot learns the SMM during an initial period of self-observation, while performing hand/arm movements. Once the SMM has been estimated, the robot can observe a demonstrator, use the VPT to transform the image features to a canonical reference frame and map these features to motor commands through the SMM. The final result will be a posture similar to that observed.

Structure of the paper

In Section 2, we present the models used throughout this work, namely the arm kinematics and the camera/eye geometry. Section 3 is devoted to the definition and estimation of the *Sensory-Motor Map*. In Section 4 we describe how the system performs the *View-Point Transformation*. In Section 5 we show how to use these elementary blocks to perform imitation and present experimental results. In Section 6, we draw some conclusions and establish directions for future work.

2 Modeling

Throughout the paper we consider a robotic system consisting of a computer simulating an antropomorphic arm and equipped with a real web camera. This section presents the models used for the camera and robot body.

2.1 Body/arm kinematics

The anthropomorphic arm is modeled as an articulated link system. Fig. 3 shows the four arm links: L_1 - forearm, L_2 - upper arm, L_3 - shoulder width and L_4 - body height.

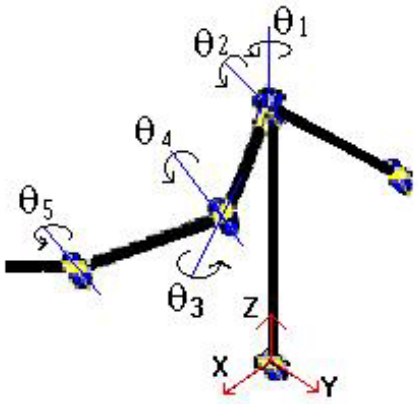

Figure 3: Kinematic model of the human arm.

It is further assumed that the relative sizes of these links are known, e.g. from biometric measurements: $L_1 = L_2 = 1$, $L_3 = 1.25$ and $L_4 = 2.5$.

2.2 Camera/eye geometry

An image is a projection of the $3D$ world whereby depth information is lost. In our case, we will retrieve depth information from a single image by using knowledge about the body links and a simplified, orthographic camera model.

We use the scaled orthographic projection model that assumes that the image is obtained by projecting all points along parallel lines plus a scale factor. Interestingly, such approximation may have some biological grounding taking into account the scale-compensation effect in the human vision [12] whereby we normalize the sizes of known objects irrespective to their distances to the eye.

Let $\mathbf{M} = [X\ Y\ Z]^T$ denote a 3D point expressed in the camera coordinate frame. Then, with an orthographic camera model, $\mathbf{M}$ is projected onto $\mathbf{m} = [u\ v]^T$, according to:

$$\begin{aligned} \mathbf{m} &= \mathcal{P}\mathbf{M} \\ \begin{bmatrix} u \\ v \end{bmatrix} &= s \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \end{aligned} \quad (1)$$

where s is a scale factor that can be estimated placing a segment with size L fronto-parallel to the camera and measuring the image size $l (s = l/L)$.

For simplification, we assume that the camera axis is positioned in the imitator's right shoulder with the optical axis pointing forward horizontally. With this specification of the camera pose, there is no need for an additional arm-eye coordinate transformation in Equation (1).

3 Sensory-Motor Map

The *Sensory-Motor Map* (SMM) defines a correspondence between perception and action. It can be interpreted in terms of forward/inverse kinematics for the case of robotic manipulators. The SMM can be used to predict the image resulting from moving one's arm to a certain posture. In our case, the SMM will allow the system to determine the arm's joint angles that correspond to a given image configuration of the arm.

In the context of imitation, the SMM can be used with different levels of ambiguity/completeness. In some cases, one wants to replicate exactly someone else's gestures, considering all the joint angles. In some other cases, however, we may want to imitate the hand pose only, while the position of the elbow or the rest of the arm configuration is irrelevant. To encompass these possibilities, we have considered two cases: the *full arm SMM* and the *free-elbow SMM* that will be described in the following sections. Finally we describe how the system can learn the SMM during a period of self-observation.

3.1 Full-Arm SMM

We denote the elbow and wrist image coordinates by $\mathbf{m_e}$ and $\mathbf{m_w}$, the forearm and upper arm image length by l_1 and l_2 and $\theta_{i=1..4}$, the various joint angles. We then have:

$$[\theta_1, \cdots, \theta_4] = \mathcal{F}_1(\mathbf{m_e}, \mathbf{m_w}, l_1, l_2, L_1, L_2, s)$$

where $\mathcal{F}_1(.)$ denotes the SMM, L_2/L_1 represents the (known) length of the upper/forearm and s is the camera scale factor.

The computation of this function can be done in successive steps, where the angles of the shoulder joint are determined first and used in a later stage to simplify the calculation of the elbow joint's angles.

The inputs to the SMM consist of features extracted from the image points of the shoulder, elbow and wrist; the outputs are the angular positions of every joint. The shoulder pan and elevation angles, θ_1 and θ_2 can be readily obtained from image data as:

$$\begin{aligned} \theta_1 &= f_1(\mathbf{m_e}) = \arctan(v_e/u_e) \\ \theta_2 &= f_2(l_2, L_2, s) = \arccos(l_2/sL_2) \end{aligned}$$

Once the system has extracted the shoulder angles, the process is repeated for the elbow. Before computing this second set of joint angles, the image features undergo a set of transformations so as to compensate the rotation of the shoulder:

$$\begin{bmatrix} u'_w \\ v'_w \\ \xi \end{bmatrix} = \mathcal{R}_{zy}(\theta_1, \theta_2) \left(\begin{bmatrix} u_w \\ v_w \\ \sqrt{s^2 L_1^2 - l_1^2} \end{bmatrix} - \begin{bmatrix} u_e \\ v_e \\ 0 \end{bmatrix} \right) \quad (2)$$

where ξ is not used in the remaining computations and $\mathcal{R}_{zy}(\theta_1, \theta_2)$ denotes a rotation of θ_1 around the z axis followed by a rotation of θ_2 around the y axis.

With the transformed coordinates of the wrist we can finally extract the remaining joint angles, θ_3 and θ_4:

$$\begin{aligned} \theta_3 &= f_3(\mathbf{m}'_{\mathbf{w}}) = \arctan(v'_w/u'_w) \\ \theta_4 &= f_4(\mathbf{m}'_{\mathbf{w}}, L1, s) = \arccos(l'_1/sL_1) \end{aligned}$$

The approach just described allows the system to determine the joint angles corresponding to a certain image configuration of the arm. In the next section, we will address the case where the elbow joint is allowed to vary freely.

3.2 Free-Elbow SMM

The *free-elbow* SMM is used to generate a given hand position, while the elbow is left free to reach different configurations. The input features consist of the hand image coordinates and the depth between the shoulder and the hand.

$$[\theta_1, \theta_2, \theta_4] = \mathcal{F}_2(\mathbf{m}_{\mathbf{w}}, {}^rdZ_w, L_1, L_2, s)$$

The elbow joint, θ_3, is set to a comfortable position. This is done in an iterative process aiming at maintaining the joint positions as far as possible from their limit values. The optimal elbow angle position, $\hat{\theta}_3$ is chosen to maximize:

$$\hat{\theta}_3 = \arg\max_{\theta_3} \sum_i (\theta_i - \theta_i^{limits})^2$$

while the other angles can be calculated from the arm features. Again, the estimation process can be done sequentially, each joint being used to estimate the next one:

$$\begin{aligned} \theta_4 &= \arcsin\left(\frac{{}^rx_h^2 + {}^ry_h^2 + {}^rz_h^2}{2} - 1\right) \\ \theta_1 &= 2\arctan\left(\frac{b_1 - \sqrt{b_1^2 + a_1^2 - c_1^2}}{a_1 + c_1}\right) \\ \theta_2 &= 2\arctan\left(\frac{b_2 - \sqrt{b_2^2 + a_2^2 - c_2^2}}{a_2 + c_2}\right) + \pi \end{aligned}$$

where the following constants have been used:

$$\begin{aligned} a_1 &= \sin\theta_4 + 1 \\ b_1 &= \cos\theta_3 \cos\theta_4 \\ c_1 &= -{}^ry_h \\ a_2 &= \cos\theta_4 \cos\theta_2 \cos\theta_3 - \sin\theta_2(1 + \sin\theta_4) \\ b_2 &= -\cos\theta_4 \sin\theta_3 \\ c_2 &= {}^rx_h \end{aligned}$$

3.3 Learning the SMM

In the previous sections we have derived the expressions of the full-arm and free-elbow SMMs. However, rather than coding these expressions directly we adopted a learning approach whereby the system learns the SMM by performing arm movements and observing the effect on the image plane.

The computation of the SMM can be done sequentially: estimating the first angle, which is then used in the computation of the following angle and so forth. This fact allows the system to learn the SMM as a sequence of smaller learning problems.

This approach has strong resemblance to the development of sensory-motor coordination in newborns and young infants, which starts by simple motions that get more and more elaborate as infants acquire a better control over motor coordination.

In all cases, we use a *Multi-Layer Perceptron* (MLP) to learn the SMM, i.e. to approximate functions $f_{i,i=1..4}$. Table 1 presents the learning error and illustrates the good performance of our approach for estimating the SMM.

θ_1	θ_2	θ_3	θ_4
$3.6e^{-2}$	$3.6e^{-2}$	3.6	3.6

Table 1: Mean squared error (in deg.^{2}) for the each joint in the *full-arm SMM*

Ideas about development can be further exploited in this construction. Starting from simpler cases, de-coupling several degrees of freedom, interleaving perception with action learning cycles are developmental "techniques" found in biological systems.

4 View-Point Transformation

A certain arm gesture can be seen from very different perspectives depending on whether the gesture is performed by the robot (self-observation) or by the demonstrator.

One can thus consider two distinct images: the *egocentric* image, I_e, during self-observation and the *allocentric* image, I_a, when looking at other robots/people. The *View-Point Transformation* (VPT) has the role of aligning the allo-centric image of the demonstrator's arm, with the ego-centric image, as if the system were observing its own arm.

The precise structure of the VPT is related to the ultimate meaning of imitation. Experiments in psychology show that imitation tasks can be ambiguous. In some cases, humans imitate only partially the gestures of a demonstrator (e.g. replicating the hand pose but having a different arm configuration, as in sign language), use a different arm or execute gestures with distinct absolute orientations [13]. In some other cases, the goal consists in mimicking someone else's gestures as completely as possible, as when performing dancing or dismounting a complex mechanical part.

According to the structure of the chosen VPT, a class of imitation behaviors can be generated. We consider two different cases. In the first case - 3D VPT - a complete three-dimensional imitation is intended. In the second case - 2D VPT - the goal consists in achieving coherence only in the image, even if the arm pose might be different. Depending on the desired level of coherence (2D/3D) the corresponding (2D/3D) VPT allows the robot to transform the image of an observed gesture to an equivalent image as if the gesture were executed by the robot itself.

4.1 3D View-Point Transformation

In this approach we explicitly reconstruct the posture of the observed arm in 3D and use fixed points (shoulders and hip) to determine the rigid transformation that aligns the allo-centric and ego-centric image features: We then have:

$$I_e = \mathcal{P}\, T\, Rec(I_a) = VPT(I_a)$$

where T is a 3D rigid transformation and $Rec(I_a)$ stands for the 3D reconstruction of the arm posture from allo-centric image features. Posture reconstruction and the computation of T are presented in the following sections.

4.1.1 Posture reconstruction

To reconstruct the 3D posture of the observed arm, we will follow the approach suggested in [14], based on the orthographic camera and articulated arm models presented in Section 2.

Let M_1 and M_2 be the 3D endpoints of an arm-link whose image projections are denoted by $\mathbf{m_1}$ and $\mathbf{m_2}$. Under orthography, the X, Y coordinates are readily computed from image coordinates (simple scale). The depth variation, $dZ = Z_1 - Z_2$, can be determined as:

$$dZ = \pm\sqrt{L^2 - \frac{l^2}{s^2}}$$

where $L = \|M_1 - M_2\|$ and $l = \|m_1 - m_2\|$.

If the camera scale factor s is not known beforehand, one can use a different value provided that the following constraint, involving the relative sizes of the arm links, is met:

$$s \geq \max_i \frac{l_i}{L_i} \quad i = 1..4 \tag{3}$$

Fig. 4 illustrates results of the reconstruction procedure. It shows an image of an arm gesture and the corresponding 3D reconstruction, achieved with a single view and considering that s and the arm links proportions were known.

With this method there is an ambiguity in the sign of dZ. We overcome this problem by restricting the working volume of the arm. In the future, we will further address this problem and several approaches may be used: (i) optimization techniques to fit the arm kinematic model to the image; (ii) explore occlusions to determine which link is in the foreground; or (iii) use kinematics constraints to prune possible arm configurations.

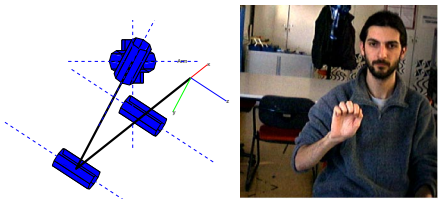

Figure 4: Left: Reconstructed arm posture. Right: Original view.

4.1.2 Rigid Transformation (T)

A 3D rigid transformation is defined by three angles for the rotation and a translation vector. Since the arm joints are moving, they cannot be used as reference points. Instead, we consider the three points in Fig. 3: left and right shoulders, (M_{ls}, M_{rs}) and hip, M_{hip}, with image projections denoted by $(m_{ls}, m_{rs}, m_{hip})$. The transformation T is determined to translate and rotate these points until they coincide with those of the system's own body.

The translational component must place the demonstrators right shoulder at the image origin (which coincide's with the system's right shoulder) and can be defined directly in image coordinates:

$$t = -{}^a m_{rs}$$

After translating the image features directly, the remaining steps consist in determining the rotation angles to align the shoulder line and the shoulder-hip contour. The angles of rotation along the z, y and x axes, denoted by ϕ, θ and ψ are given by:

$$\begin{aligned} \phi &= \arctan\left(v_{ls}/u_{ls}\right) \\ \theta &= \arccos\left(u_{hip}/L_4\right) \\ \psi &= \arccos\left(v_{hip}/L_3\right) \end{aligned}$$

Hence, by performing the image translation first and the 3D rotation described in this section, we complete the process of aligning the image projections of the shoulders and hip to the ego-centric image coordinates.

4.2 $2D$ View-Point Transformation

The $2D$ VPT is used when one is not interested in imitating the depth variations of a certain movement, alleviating the need for a full 3D transformation. It can also be seen as a simplification of the 3D VPT if one assumes that the observed arm describes a fronto-parallel movement with respect to the camera.

The $2D$ VPT performs an image translation to align the shoulder of the demonstrator (${}^{a}m_s$) and that of the system (at the image origin, by definition). The VPT can be written as:

$$VPT({}^{a}m) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} [{}^{a}m - {}^{a}m_s] \quad (4)$$

and is applied to the image projection of the demonstrator's hand or elbow, ${}^{a}m_h$ or ${}^{a}m_e$.

Notice that when the arm used to imitate is the same as the demonstrator, the imitated movement is a mirror image of the original. If we use a simple identity matrix in Equation (4) then the movement will be correct. At the image level both the 2D and 3D VPTs have the same result but the 3D posture of the arm is different in the two cases.

From the biological standpoint, the 2D VPT is more plausible than the 3D version. In [13] several imitation behaviors are presented which are not always faithful to the demonstrated gesture: sometimes, people do not care about usage of the correct hand, depth is irrelevant in some other cases, movements can be reflections of the original ones, etc. The 3D VPT might be more useful in industrial facilities where gestures should be reproduced as exactly as possible.

5 Experiments

We have implemented the modules discussed in the previous sections to build a system able to learn by imitation. In all the experiments, we use a web camera to observe the demonstrator gestures and a simulated robot arm to replicate those gestures.

We start by describing the approach used for hand-tracking before presenting the overall results of imitation. The position of the shoulder is assumed to be fixed. In the following sections we shall discuss about the procedures for doing imitation.

5.1 Hand Color Segmentation

To find the hand in the image we use a color segmentation scheme, implemented by a feed-forward neural network with three neurons in the hidden layer. As inputs we use the hue and saturation channels of HSV color representation. The training data are obtained by selecting the hand and the background in a sample image. After color classification a *majority* morphological operator is used. The hand is identified as the largest blob found and its position is estimated over time with a Kalman filter. Figure 5 shows a typical result of this approach.

5.2 Gesture Imitation

The first step to achieve imitation consists in training the system to learn the Sensory-Motor Map as described in Section 3.3. This is accomplished by a neural network that estimates the SMM while the system performs a large number of arm movements.

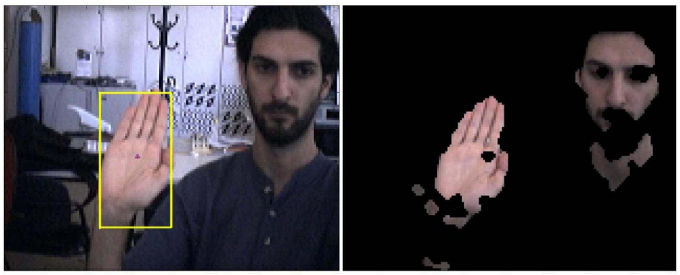

Figure 5: Skin color segmentation results.

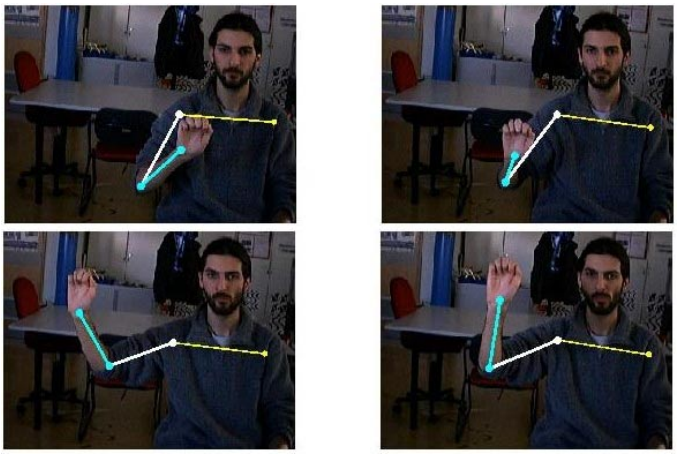

Figure 6: The quality of the results can be assessed by the coincidence of the demonstrator gestures and the result of imitation.

The imitation process consists of the following steps: (i) the system observes the demonstrator's arm movements; (ii) the VPT is used to transform these image coordinates to the *ego-image*, as proposed in Section 4 and (iii) the SMM generates the adequate joint angle references to execute the same arm movements.

Figure 6 shows experimental results obtained with the $3D$-VPT with the learned SMM (full-arm). To assess the quality of the results, we overlaid the images of the executed arm gestures (wire frame) on those of the demonstrator. The figure shows that the quality of imitation is very good.

Figure 7 shows results obtained in real-time (about 5 Hz) when using the $2D$ VPT and the *free-elbow* SMM. The goal of imitating the hand gesture is well achieved but, as expected, there are differences in the configuration of the elbow, particularly at more extreme positions.

These tests show that encouraging results can be obtained with the proposed framework under realistic conditions.

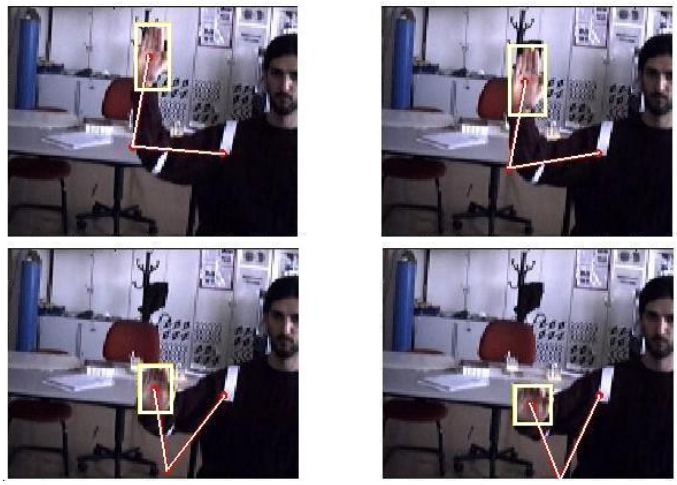

Figure 7: Family of solutions with different elbow angles, while the hand is faithfully imitated.

6 Conclusions and future work

We have proposed an approach for learning by imitation that relies exclusively on visual information provided by a single camera.

One of the main contributions is the *View-Point Transformation* that performs a "mental rotation" of the image of the demonstrator's arm to the *ego-image*, as if the system were observing its own arm. In spite of the fundamental importance of the VPT in visual perception and in the psychology of imitation [7], it has received little attention by researchers in robotics.

We described two different VPTs needed for 3D or 2D imitation. The *View-Point Transformation* can have an additional interest to *Mirror Neurons* studies, by providing a canonical frame of reference that greatly simplifies the recognition of arm gestures.

The observed actions are mapped into muscles torques by the *Sensory-Motor Map*, that associates image features to motor acts. Again two different types of SMM are proposed, depending on whether the task consists of imitating the entire arm or the hand position only. The SMM is learned automatically during a period of self-observation.

Experiments conducted to test the various sub-systems have led to encouraging results, thus validating our approach to the problem.

Besides improvements on the feature detection component using shape and kinematic data, future work will focus on the the understanding of the task goals to enhance the quality of imitation.

References

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

[2] J. Yang, Y. Xu, and C.S. Chen. Hidden markov model approach to skill learning and its application to telerobotics. *IEEE Transations on Robotics and Automation*, 10(5):621–631, October 1994.

[3] T. G. Williams, J. J. Rowland, and M. H. Lee. Teaching from examples in assembly and manipulation of snack food ingredients by robot. In *2001 IEEE/RSJ, International Conference on Intelligent Robots and Systems*, pages 2300–2305, Oct.29-Nov.03 2001.

[4] Aaron D'Souza, Sethu Vijayakumar, and Stephan Schaal. Learning inverse kinematics. In *International Conference on Intelligent Robots and Systems*, Maui, Hawaii, USA, 2001.

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

[6] Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Visuomotor neurons: ambiguity of the discharge or 'motor' perception? *International Journal of Psychophysiology*, 35, 2000.

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

[8] Minoru Asada, Yuichiro Yoshikawa, and Koh Hosoda. Learning by observation without three-dimensional reconstruction. In *Intelligent Autonomous Systems (IAS-6)*, 2000.

[9] A. Billard and G. Hayes. Drama, a connectionist architecture for control and learning in autonomous robots. *Adaptive Behaviour*, 7(1), 1999.

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

[11] G. Metta, G. Sandini, L. Natale, and F.Panerai. Sensorimotor interaction in a developing robot. In *First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 18–19, Lund, Sweden, September 2001.

[12] Richard L. Gregory. *Eye and Brain, The Psychology of Seeing*. Princeton University Press, Princeton, New Jersey, 1990.

[13] Philippe Rochat. Ego function of early imitation. In Andrew N. Meltzoff and Wolfgang Prinz, editors, *The Imitative Mind*. Cambridge University Press, 2002.

[14] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80, 2000.