A Grid-based Approach to the Body Correspondence Problem in Robot Learning by Imitation

L. Molina-Tanco, J.P. Bandera, J.A. Rodríguez, R. Marfil, A. Bandera and F. Sandoval

Grupo de Ingeniería y Sistemas Integrados, Dpto. Tecnología Electrónica
ETSI Telecomunicación, Universidad de Málaga, 29071-Málaga (Spain)
jpbandera@uma.es

Summary. In the learning by imitation framework, a module is required to translate the visually perceived behaviour of the demonstrator's body to a representation where the imitator perceives its own movements, either visually, propioceptively, or both. This transformation must take into account the different body configurations of imitator and demonstrator. In this work, a solution to translation of the visually perceived end-effector motion is presented. The proposed approach is included in a system which allows a robot to learn the behaviour of a human demonstrator from stereo visual information. Real-time results are presented and discussed.

Key words: Learning by imitation, body correspondence problem, visual perception.

1 Introduction

Learning by observation and imitation constitute two important mechanisms for learning social behaviour in humans and other animal species, e.g. dolphins, chimpanzees and other apes [3]. Inspired by nature and in order to speed up the learning process in complex motor systems, imitation arises as a powerful tool to improve the learning process [9]. Recent work has demonstrated that learning by observation and imitation can help robots to quickly learn new skills and tasks from natural human instruction and few demonstrations [4, 8].

One of the main problems that must be solved by a learning by observation system is the translation of the perceived behaviour of the demonstrator to an imitator's self-centered reference frame. If the imitator can perceive visually its own behaviour, the translation can be formulated as a view-point transformation problem [1, 5] . However a complex visuo-motor mapping module must be used to map the view-point transformed motion into motor commands. An alternative approach is to

assume a kinematic model to estimate the movements of the demonstrator. The computed joint angles are used to directly drive the robot motors [6]. These and other approaches to imitation [8] often assume that demonstrator and imitator present the same body configuration.

Our humanoid robot HOAP-1 by Fujitsu presents a large number of degrees of freedom, which nevertheless are not sufficient to consider its body similar to that of a human demonstrator. Therefore, the previous assumption is not applicable in our case. This difficulty is often circumvented by imitating only the behaviour of the demonstrator's end-effectors [6, 8]. However, the translation of the perceived demonstrated motion must be done carefully to respect joint limits and avoid collisions, while still performing imitation. In this paper, we present a method to translate motion that can be employed when demonstrator and imitator present different body configurations. The key idea is to quantize the locations that the end-effectors can reach using three-dimensional grids. Self-collisions and joint limits sculpt these grids into a complex shape which is learnt a priori in a body babbling stage. Imitation is reduced to choosing the cell of the imitator's grid that corresponds to the perceived demonstrator cell.

The paper is organised as follows: Section II is a brief description of the whole learning by imitation system. This description is necessary to justify some principles adopted to finally implement the proposed motion translation module. Section III describes the motion translation module and analyzes the two different criteria which can be chosen to map imitator and demonstrator behaviours. Section IV shows some experimental results and, finally, conclusions and future work are presented in Section V.

2 Overview of the learning by imitation architecture

Fig. 1 shows a perspective of the learning by imitation framework in which the module proposed in this paper is integrated. The architecture can be divided into two major modules related to visual perception and active imitation. The stereo visual perception system is the responsible of the detection and tracking of the head and hands of the demonstrator. This system works without special devices or markers, using an attention mechanism which drives attention towards skin color regions. A model-based pose estimation method based on inverse kinematics, joint limit enforcement and collision avoidance [2] reduces noise in the estimation of the three-dimensional location of head and hands. The module presented in this paper transforms the estimated hand positions to reference frames local to each arm of the robot.

The active imitation module provides the robot with two modes of imitation. Thus, it allows the robot to replicate the behaviour of a demonstrator without seeking to understand it (action level imitation or mimicking [5]). Additionally, it allows memorization of observed behaviours, and recognition when a recently perceived behaviour corresponds to one of the previously memorized ones ("true imitation" or program level imitation [5]).

Joint angles are extracted through the use of a kinematic model of the robot body. This model includes a set of constraints that limit the robots movements and avoids collisions between the different body parts of the robot [2]. The body model also determines the 3D space which contains the poses that the robot can achieve. Finally, in order to recognize previously memorized behaviours, the active imitation system includes a behaviour comparison module that uses a dynamic programming to make a decision on whether a behaviour has been previously observed.

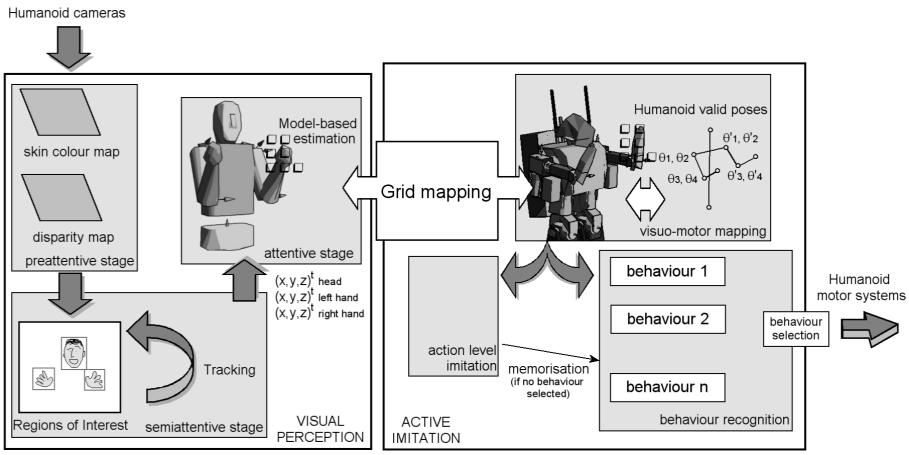

Fig. 1. Overview of the learning by imitation architecture

3 Grid-based approach to solving the correspondence problem

In any form of imitation, a correspondence has to be established between demonstrator and imitator. When the imitator body is very similar to that of the demonstrator, this correspondence can be achieved by directly mapping the corresponding body parts. Thus, Lopes and Santos-Victor [5] propose two different view-point transformation algorithms to solve this problem when the imitator can visually perceive both the demonstrator's and its own behaviour. However, the similarity between the two bodies is not always sufficient to adopt this approach. In these cases, it is not possible to establish a simple one-to-one correspondence between the coordinates of their corresponding body parts [7]. However, a robot might imitate a human behaviour successfully even without having the same number and type of joints in its head and neck or arms and hands as the human whose behaviour it emulates. Thus, Sauser and Billard [8] describe a model of a neural mechanism by which an imitator can map movements of the end-effector performed by other agents onto its own frame of reference. Their work is based on the mapping between observed and achieved

subgoals, where a subgoal is defined as to reach a similar relative position of the arm end-effectors or hands. Our work is based on the same assumption.

In this paper, the mapping between observed and achieved subgoals is defined by using three-dimensional grids, associated to each demonstrator and imitator hand. Fig. 3 shows the grids associated to the end-effectors of the demonstrator and imitator. These grids are internally stored by the robot and can be autonomously generated from the human body and robot's models. The grid provides a quantization of the demonstrators range of motion and its cells can be related to the cells of the imitators grid. This relation is not a one-to-one mapping, as the robots end-effectors are not able to reach all the positions that the humans hands can reach. There is a many-to-one correspondence between the position coordinates of corresponding human hands and robot end-effectors. Thus, two main problems have to be solved: i) how to perform re-scaling of the observed behaviour to the imitators end-effectors and ii) how to estimate the function that determines the egocentric (imitator) cell associated to an observed allocentric (demonstrator) cell.

The problem of the re-scaling can be solved by hand [8], translating the demonstrator's observed pose to a final resulting pose that lies within the imitator's range of motion. When a one-to-one mapping between the body parts of demonstrator and imitator is possible, the robot can solve the re-scaling problem by itself. Thus, Lopes and Santos-Victor [5] propose a sensory-motor module that allows autonomous learning of the motor commands to reach a certain pose, related to the observed one by the view-point transformation.

In this work, the grids that define the human demonstrators and the robot imitators range of motion are autonomously acquired in a previous body babbling stage. Two strategies are presented below to perform the mapping between the observed allocentric cell and the corresponding egocentric cell: Uniform scale mapping (USM) and non-uniform scale mapping (NUSM). Both strategies produce a look-up table that establishes a suitable many-to-one mapping between the cells in the demonstrator's grid and the cells of the imitator's grid.

3.1 Uniform scale mapping

The length of a stretched arm for both the demonstrator and the imitator gives the maximum diameters of the corresponding grids. The relation between these diameters automatically provides a re-scaling factor. If both arms presented the same ranges of motion, mapping the behaviour of the demonstrator's arm into that of the imitator's would be a matter of selecting the same cell in both grids. USM starts by supposing that this is the case. The cell that was reached by the demonstrator is checked for existence in the imitator's grid. If the cell is not present, the closest valid cell in the imitator's grid is chosen instead.

3.2 Non-uniform scale mapping.

USM may distort the quality of the imitated behaviour if a large part of it is performed in an area that the robot cannot reach. In particular, the end-effectors of our

HOAP-1 robot cannot reach the area closest to the shoulders, as the elbows cannot bend beyond 90 degrees. NUSM models the region around the human model shoulder as a sphere, and the region around the robot shoulder as a region within two concentric spheres. NUSM defines a transformation between both regions which is applied to all demonstrator cells.

Figure 2 illustrates the difference between both strategies using simple two-dimensional grids. Greyed areas represent the cells that the robot cannot reach. Colour is employed to indicate how the cells are mapped between the grid associated to the human model (demonstrator) and the grid associated to the robot model (imitator). The following section describes in detail how the grids are obtained and present a qualitative comparison.

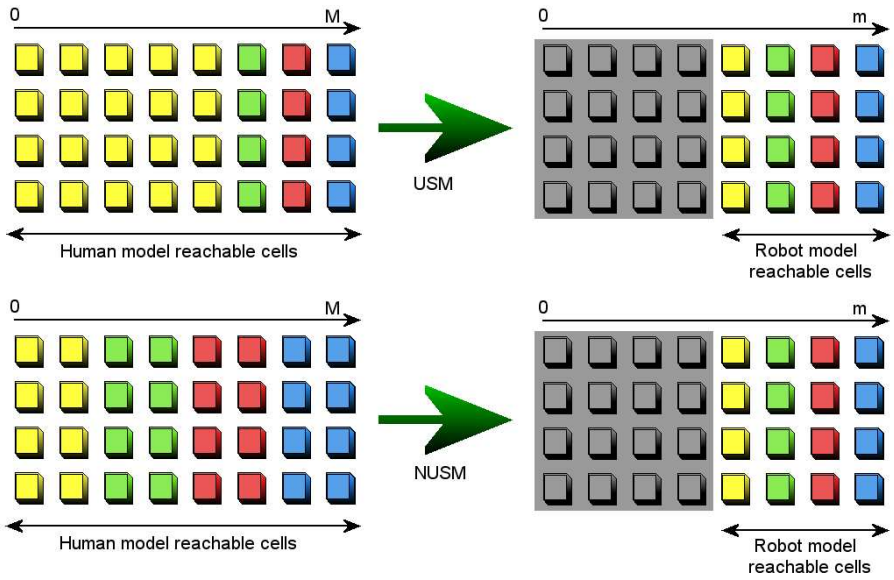

Fig. 2. The two strategies (USM and NUSM) presented in this paper to match end-effector positions reached by the human model (demonstrator) and the robot model (imitator). See text for details.

4 Experimental results

4.1 Autonomous grid learning

In order to generate the three-dimensional grids which quantize the range of motion of the imitator and demonstrator models, a body babbling action is required to determine reachable end-effector positions. This action can be performed automatically,

by randomly selecting poses from uniform three-dimensional grids that envelop the model. Two grids are generated, one for each arm. Each grid is centered in its corresponding shoulder and, after the autonomous exploration process, will contain only the positions that the corresponding end-effector is able to reach.

The initial grid is a $(d \times d \times d)$ box, where d is the length of the model arm when stretched. The space within the box is quantized to a limited number of positions (grid cells), which the method can explore exhaustively. Poses that do not produce collisions and are reached within joint limits are selected as valid ones.

Once the subset of all valid cells is extracted from an uniform grid, it conforms the definitive grid associated to the arm of the demonstrator or the imitator. Fig. 3 shows the grids associated to human and HOAP-1. Valid cells represent a small subset of the complete uniform grid. For the human model, only an average 20% of the cells are reachable by the end-effectors. For the HOAP-1 model, less than 10% of the cells are valid.

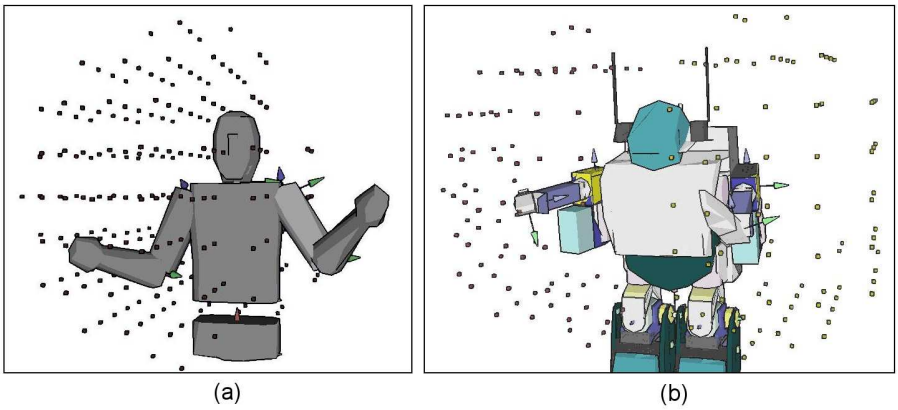

Fig. 3. a) Grid associated to human right arm, for a cell size of 15 cm; and b) Grids associated to HOAP-1 left (yellow) and right (red) arms, for a cell size of 45 mm.

The cell size is a parameter of this exploration process that is selected in order to bound the computational resources. If this size is too large, each cell defines a great volume of the range of motion. This can provoke that the posterior active imitation module confuses different behaviours. On the other hand, if the the volume defined by the cell is too small, more computational resources are needed. The following section presents imitation results with two cell sizes for the grids associated to the demonstrator, 15 cm and 5 cm, which result in a total of 223 and 6533 valid cells. These would also be the number of entries in the lookup tables that define the mapping between the demonstrator and the imitator cells.

4.2 Qualitative evaluation of the mimicking ability

The proposed system has been tested in our learning by imitation framework (Section 3), whose goal is to make a Fujitsu HOAP-1 imitates a human demonstrator. The robot uses the visual information provided by a pair of head-mounted stereo cameras to perform gesture tracking and recognition. The baseline of this stereo system is 28 mm. The whole system runs on a 3 GHz Pentium IV PC, that sends the joint angles to the robot. These angles are read by robot's own PC and sent to motor controllers via USB. The process is performed in real time, at 25 frames per second. The experiments consisted of imitation and learning of a set of diving signals. Each signal conveys a message to fellow divers, and the meaning of this message depends on a sequence of positions and movements.

It is not the intention of this paper to analyze this learning by imitation framework, but to show the performance of the grid-based solution to the correspondence problem. Thus, the first experiments reported here modify mapping algorithm and cell sizes and compare resulting HOAP-1 trajectories.

Fig. 4 shows different codifications of the '(I want to) go Up' ([GoUp]) gesture. This gesture consists in moving up and down the right hand. Each dot represents a cell visited during the movement. Figs. 4.a and 4.b present results of USM and NUSM respectively, using a cell size of 15 cm in the human model grids and a cell size of 45 mm in the robot grids. Figs. 4.c and 4.d use the same algorithms, with cell size of 5 cm for the demonstrator and 15 mm for the imitator. As depicted, the reduction in cell size produces an increment in the density of visited cells, so that the resulting HOAP-1 movement is smoother and closer to demonstrator's trajectory.

The second set of experiments fixed cell size, and alternated between the two different mapping algorithms presented in this paper. Fig. 5 depicts the cell trajectories in the XZ (frontal) plane, using USM and NUSM algorithms. As Fig. 5.b shows, the density of cells USM trajectories are coarser than those generated by NUSM, which produce motions with more detail. Besides, NUSM tends to separate the end-effector positions from the robot, avoiding the non-reachable space next to the shoulders. This emphasizes the movements, resulting in a better subjective recognition of the imitated behaviour.

5 Conclusions

In learning by imitation, the correspondence problem can be solved by a direct one-to-one mapping between the coordinates of the corresponding body parts of demonstrator and imitator. In our case, the imitator is a humanoid robot HOAP-1 whose shape and degrees-of-freedom do not allow us to make this direct one-to-one mapping. Similarly to other authors, we choose to circumvent this difficulty by formulating the correspondence problem as that of matching observed and achieved subgoals, where a subgoal is defined as to reach a similar relative position of the imitator end-effectors or hands. However care must still be taken with joint-limits and self-collisions when transforming the perceived end-effector trajectories.

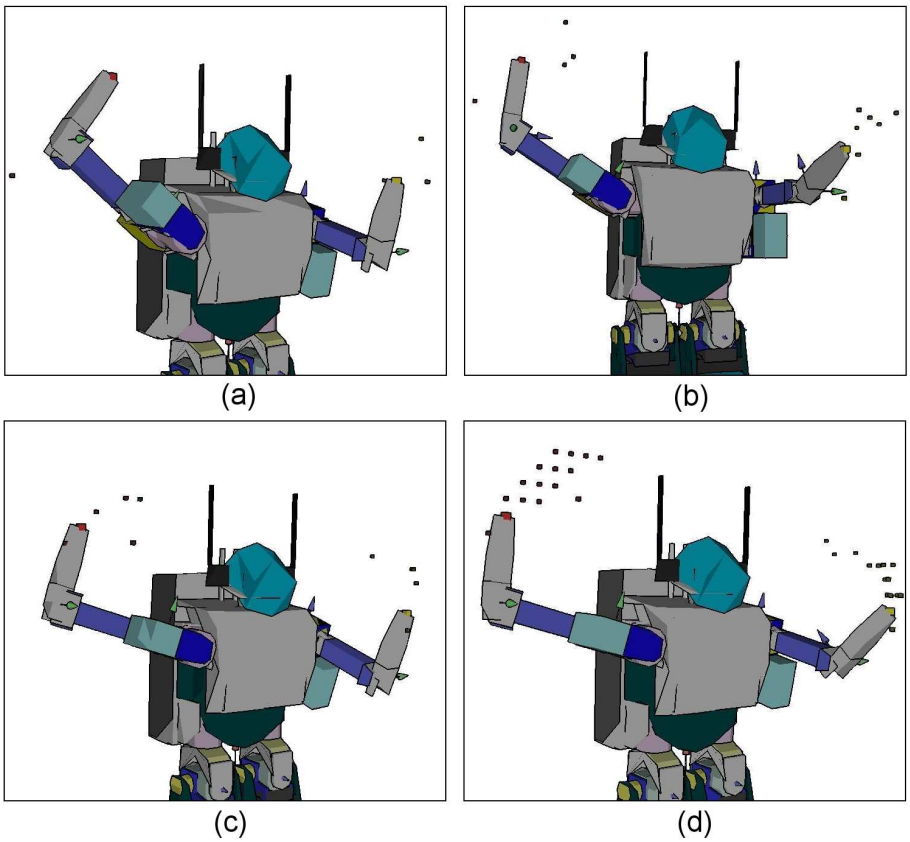

Fig. 4. Cells visited during imitation of [GoUP] gesture: a) USM. Cell size 45 mm; b) NUSM. Cell size 45 mm; c) USM. Cell size 15 mm; and d) NUSM. Cell size 15 mm

In this paper we introduce a simple approach based on quantizing both the perceived hand positions of the demonstrator and end-effector positions reachable by the robots using three-dimensional grids. A look-up table stores a mapping between cells of each grid. This is a many-to-one mapping, as there are less cells in the grid associated to the robot. We present two strategies to build this look-up table. The two strategies have been tested in an imitation scenario in which demonstrator and imitator present different body configurations. Initial experiments suggest that movements are mapped correctly, as imitated actions are recognized as similar to demonstrated ones. Also, the comparison between the two strategies points out that, although a generic uniform mapping is a valid option, a non uniform mapping is a better solution when the relation between demonstrator's and imitator's reachable spaces can be roughly established.

Our future work will address the use of these grids to help in parameterizing perceived and executed movements. Grids with different cell sizes appears as a promis-

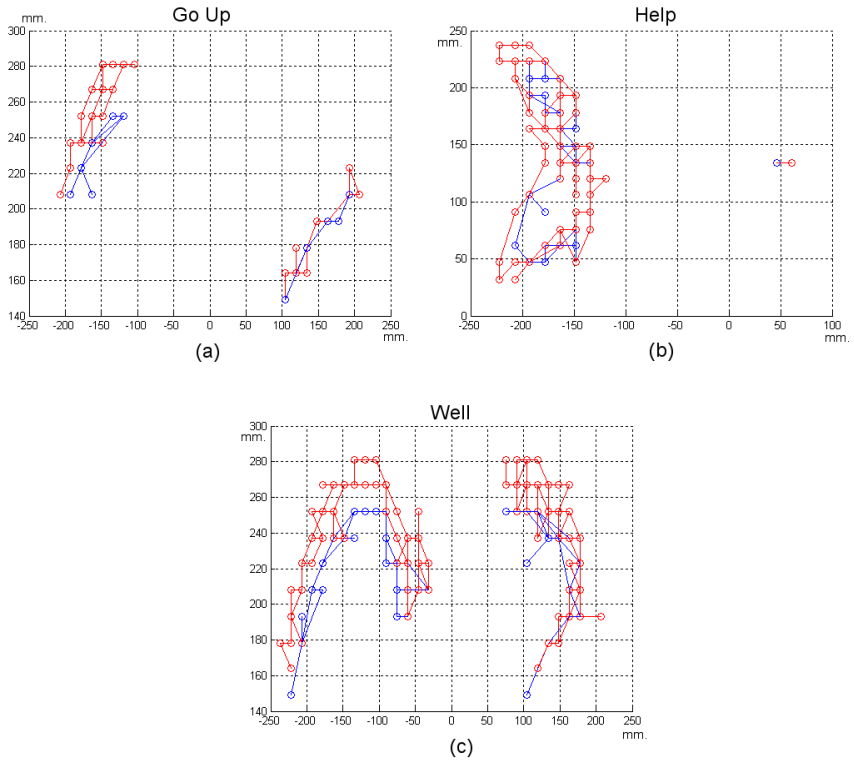

Fig. 5. XZ Trajectories of HOAP-1 arms during imitation of different human gestures.

ing tool not only to recognize a particular behaviour, but also to refine perception and imitation.

Acknowledgements

This work has been partially granted by the Spanish Ministerio de Educación y Ciencia (MEC) and FEDER funds Project no. TIN2004-05961 and by the European Robotics Research Network (EURON) Project VISOR.

References

1. Asada M., Yoshikawa Y., and Hosoda. K. (2000). Learning by observation without three-dimensional reconstruction, Intelligent Autonomous Systems (IAS-6): 555-560.

2. Bandera J.P., Marfil R., Molina-Tanco L., Bandera A. and Sandoval F. (2005) Model-based Pose Estimator for Real-time Human-Robot Interaction. Proc. of the Third International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2005), Singapore.
3. Dautenhahn K. and Nehaniv C. (2002) Imitation in animals and artifacts. MIT Press, Cambridge.
4. Demiris J. and Hayes G.M. (2002) Imitation as a dual-route process featuring predictive and learning components: A biologically plausible computational model. In: Dautenhahn K. and Nehaniv C. (eds) Imitation in animals and artifacts. MIT Press: Cambridge.
5. Lopes M. and Santos-Victor J. (2005) Visual learning by imitation with motor representations. IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics 35(3): 438-449.
6. Molina-Tanco L., Bandera J.P., Marfil R. and Sandoval F. (2005) Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Alberta, Canada: 1808-1813.
7. Nehaniv C. and Dautenhahn K. (2005) The correspondence problem in social learning: What does it mean for behaviors to “match” anyway?. In: Hurley S. (ed) Perspectives on imitation: from cognitive neuroscience to social science. MIT Press, Cambridge.
8. Sauser E. and Billard A. (2005) View sensitive cells as a neural basis for the representation of others in a self-centered frame of reference. In Proc. of the Third International Symposium on Imitation in Animals and Artifacts, Hatfield, UK: 119-127.
9. Schaal S. (1999) Is imitation learning the route to humanoid robots?. Trends in Cognitive Sciences 3(6): 233-242.