Learning postures through an imitation game between a human and a robot

S. Boucenna[1], E. Delaherche[1], M. Chetouani[1], P. Gaussier[2]
[1] ISIR, UPMC, UMR 7222, Paris, France, [2] ETIS, CNRS UMR 8051, ENSEA, Cergy-Pontoise University
{boucenna,delaherche}@isir.upmc.fr, mohamed.chetouani@upmc.fr, gaussier@ensea.fr

***Abstract*—In this paper, we investigate a sensory-motor architecture allowing a robot to learn to recognize postures. The learning is performed without a teaching signal that associates a specific posture with the robot's motor internal state. Our architecture assumes that the robot initially performs postures, then the human imitates them. An on-line learning scheme without an explicit reward or ad-hoc detection mechanism or a formatted teaching technique is proposed. Investigations on how a "naive" system can learn to imitate correctly another person's posture during a natural interaction motivate the current research work.**

I. Introduction

In previous study [Boucenna et al., 2010], we showed that a robotic head can autonomously develop the facial expression recognition without having a teaching signal. Our starting point was a mathematical model that shows that, if the baby uses a sensory motor architecture for the recognition of a facial expression, then the parents must imitate the baby's facial expression to allow on-line learning [P. Gaussier, 2004]. The baby-mother interaction is usually considered as a relevant framework. A newborn or infant has a set of expressions that are linked with his/her own internal state, for example crying and a sad face when he/she needs food or a happy face after being fed. [Meltzoff and Moore, 1977] show that infants between 12 and 21 days of age can imitate both facial expressions and gestures. Learning to recognize facial expressions without a teaching signal that allows the association between what is perceived by the robot and his internal state is challenging (e.g., the vision of "happy face" and an internal emotional state of happiness [G. Gergely, 1999]). This issue is investigated through robotic experiments [Boucenna et al., 2010]. In this paper, we show that as the facial expression recognition problem, posture recognition can autonomously be learned with the help of the sensory-motor architecture.

To test our model, the following experimental protocol was adopted: In the first phase of the interaction (learning phase), the robot produces a random posture (4 basic postures and a neutral posture) for 2s; then, the robot returns to a neutral posture for 2s to avoid human misinterpretation of the robot posture (the same procedure is used in psychological experiments). The human subject is asked to imitate the robot. After this first phase, which lasts between 2 and 3 min according to the subject's "patience", the robot must imitate the posture of the human partner (The robot can produce only the postures learned during the learning phase).

II. Sensory-motor architecture

A simple sensory-motor architecture allows the robot to learn, recognize and imitate postures. The visual processing allows to extract local views, then each local views are learned by the SAW (Self Adaptive Winner Takes All) group (visual features), and the LMS group learns the association between the visual features and the postures (Figure 1). This architecture can solve the posture recognition problem if and only if we assume that the robot produces first posture according to his/her internal state and that next the human partner imitates the posture of the robot allowing in return the robot to associate these postures with his/her internal state.

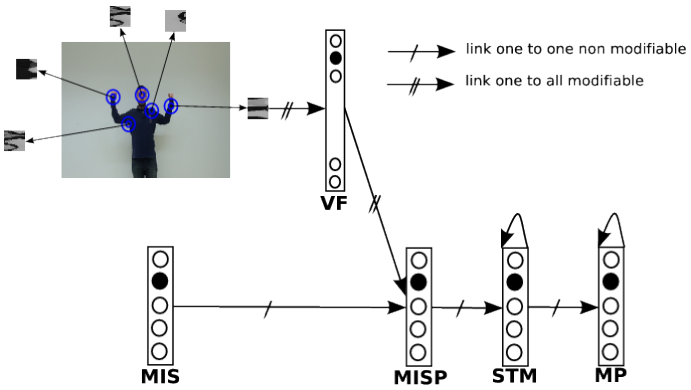

Fig. 1. The architecture for posture recognition and imitation. The visual processing allows for the sequential extraction of the local views. This visual system is based on a sequential exploration of the image focus points. A gradient extraction is performed on the input image. A convolution with a Difference Of Gaussian (DOG) provides the focus points. Last, the local views are extracted around each focus point. The LMS group learns the association between the visual features and the postures.

In order to obtain a limited number of local views to be learned and analyzed a Difference of Gaussian (DoG) filter is applied on the gradient image to determine stable focus points associated to angular or curved areas (these points remain stable from small perspective or scale variations). An inhibition of return allows to explore sequentially the image around each focus point. Local views are next obtained after a log polar transform of the input image centered on each focus point (this increases the robustness of the extracted local views to small rotations and scale variations).

In a first phase of the experiment (learning phase), the robot produces random postures and analyzes the images grabbed

from its CCD camera. For each postural state, a winner takes all mechanism is used to store and recognize the explored local views. The extracted local view around each focus point is learned and recognized by a group of neurons SAW using a k-means variant that allows online learning and real-time functions [Kanungo et al., 2002].

Of course, there is no constraint on the selection of the local views (no person detection). This scenario means that many distractors can be present (there are local views in the background). This scenario also means that any of these distractors can be learned on SAW. The person/ground discrimination can be learned because the local views in the background will not be statistically correlated with a given posture. In an interaction, the distractors are present for all of postures, and their correlation with a specific posture will tend to zero. Only the local views on the person correlated with a given robot posture will be reinforced.

The use of the Widrow and Hoff rule (derived from a least mean square (LMS) optimization) will learn correctly if, during the period that is allowed for the exploration of one image, sufficient focus points can be found on the person. In our neural network, the LMS (Least Mean Square) associates the activity of the visual features SAW with the robot posture. Neuronal activities correlated with the robot postures (its motor command) are reinforced through a classical conditioning mechanism [Widrow and Hoff, 1960]. During the learning phase, the unconditional stimulus is the signal generated by the robot itself to trigger a specific posture. As a result, the robot learns to discriminate the features associated to a given posture from the distractor in the image (objects in the background but also non relevant parts of the human body). In typical conditions, views associated to different parts of the body are found (these views are mainly centered on the arms, the head and leg).

III. Preliminary Results

The online learning can involve specific problems regarding real-time learning because the human reaction time to the robot postures is not immediate. This time lag can greatly disturb the learning process. If the robot learns the first images, which are associated with the human's previous posture, then the previous posture is unlearned. The presentation time of a given posture must be long enough for the first images to have expired. In spite of this problem, the incremental learning is robust as we increase the number of human partners (up to 7 human partners). The success rate is 75% when the robot learned with 7 human partners for the recognition of 5 specific postures (Figure 2). Yet, the introduction of a short term memory (STM) at the motor level (for the triggering of posture) allows to obtain quite convincing results at the price of a small hysteresis in the decision process. In spite of this hysteresis, the system maintains a real-time interaction because the system can analyse 10 images per second (the system can analyze up to 10 local views on each image).

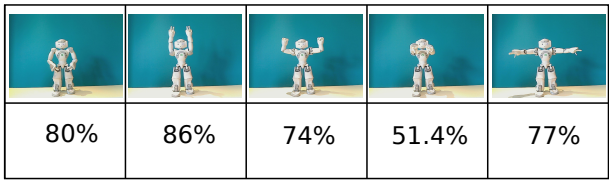

Fig. 2. The success rate for each posture. A total of 7 persons interacted with the robot. During the learning phase, these humans imitate the robot, and then the robot imitates them. To perform the statistical analyses, each image was annotated with the response of the robot head. The annotated images were analyzed, and the correct correspondence was checked.

IV. Conclusion

The goal of these experiments is to show that our architecture can be generalized to several tasks (for example, facial expression recognition [Boucenna et al., 2010], joint attention [Boucenna et al., 2011], or posture recognition). Our approach can enable autonomous learning through interaction, if the robot produces and, then the human partner imitates the robot. We can suggest the robot/human system is an autopoietic system [Mataruna and Varela, 1980] in which the imitation is an important element to maintain the interaction and to allow the learning of more and more complex skills.

In future studies, we want to build an architecture that integrate several capabilities (e.g., emotion, posture, and voice recognition) to improve the human robot communication [Delaherche et al., 2012].

Acknowledgments

This work was supported by the UPMC "Emergence 2009" program and the European Union Seventh Framework Programme under grant agreement n288241.

References

[Boucenna et al., 2010] Boucenna, S., Gaussier, P., Andry, P., and Hafemeister, L. (2010). Imitation as a communication tool for online facial expression learning and recognition. *IROS*, pages 5323–5328.

[Boucenna et al., 2011] Boucenna, S., Gaussier, P., and Hafemeister, L. (2011). Development of joint attention and social referencing. *IEEE International Conference on Development and Learning - ICDL*, 2:1–6.

[Delaherche et al., 2012] Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony : A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*. to appear.

[G. Gergely, 1999] G. Gergely, J. W. (1999). Early socio-emotional development: contingency perception and the social-biofeedback model. *In P. Rochat, (Ed.), Early Social Cognition: Understanding Others in the First Months of Life*, pages 101–136.

[Kanungo et al., 2002] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:881–892.

[Mataruna and Varela, 1980] Mataruna, H. and Varela, F. (1980). *Autopoiesis and Cognition: the realization of the living*. Reidel, Dordrecht.

[Meltzoff and Moore, 1977] Meltzoff, A. N. and Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78.

[P. Gaussier, 2004] P. Gaussier, K. Prepin, J. N. (2004). Toward a cognitive system algebra: Application to facial expression learning and imitation. *In Embodied Artificial Intelligence*, pages 243–258.

[Widrow and Hoff, 1960] Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *IRE WESCON*, pages 96–104, New York. Convention Record.