A developmental approach to grasping

Lorenzo Natale, Giorgio Metta and Giulio Sandini

LIRA-Lab, DIST, University of Genoa
Viale Causa 13, 16145, Genova Italy
email: {nat, pasa, sandini}@liralab.it

Abstract

In this paper we describe a developmental path which allows a humanoid robot to initiate interaction with the environment by grasping objects. Development begins with the exploration of the robot's own body (control of the head and arm, identification of the hand) and moves afterward to the external world (reaching and grasping). A final experiment is reported to illustrate how these simple behaviors can be integrated to start autonomous exploration of the environment. In fact we believe that for an active system the capacity to act is not a mere arrival point but it is rather required in order for the system to further develop by acquiring and structuring information about its environment.

Introduction

If the first interaction with the environment happens through vision it is only by acting that we are able to discover certain properties about the entities populating the external world. For example by applying different actions on an object we can probe it for properties like weight, rigidity, softness and roughness, but also collect information about its shape. Furthermore we can carry out further exploration to learn how an object behaves when certain actions are applied to it or, in a similar way, how we can handle it to achieve a particular goal (tool use).

Besides, autonomous agents can exploit actions to actively guide exploration. For an artificial system like a robot this can be extremely useful to simplify learning. For instance the system can identify a novel object in a set and grasp it, bring it closer to the cameras (so to increase the resolution), rotate it, squeeze it, and eventually drop it after enough information has been acquired. Exploration in this case is easier because it is initiated by the agent in a self-supervised way. This does not only mean that the agent has direct control on the exploration procedure, but also that it can establish a causal link between its actions and the resulting perceptions. While holding and rotating an object, for example, its appearance, the tactile sensation coming from the hand, along with the torque sensed at the wrist, can be associated to the position of the fingers around the object and to its orientation. Similarly, affordances can be explored by trying to grasp the object in different ways and discovering what kind of actions can be performed with it.

The ability to manipulate objects emerges relatively early in children during development; for instance at three months infants start reaching for objects to grasp them and bring them to their mouth; nine months-old babies are able to control the fingers to perform different grasp types (precision grip, full palm grasp, (von Hofsten, 1983)). Maybe it is not by chance that infants learn to grasp way before they can speak or walk. However, even the simplest form of grasp (with the full hand open as newborns do in the first months of their life) is not a trivial task. It involves at least the ability to control gaze, to move the arm to reach a particular position in space, pre-shape the hand and orient it according to the object's size and orientation. In addition, the impact with the object must be predicted to correctly plan the pre-shaping of the hand (von Hofsten et al. 1998). In infants all these motor competences are not present at birth; rather they are acquired during development by exploiting an initial set of innate abilities which allow them to start the exploration of their body and environment.

In this paper we present a possible developmental path for a humanoid robot mimicking some aspects of infant development. In our artificial implementation we divided this process in three phases. The first phase concerns learning a body self-image; the robot explores the physical properties of its own body (e.g. the weight of the arm, the visual appearance of the hand) and basic motor abilities (e.g. how to control the head to visually explore the environment). We call the second stage learning to interact; here the robot starts active exploration of the external world and learns to perform goal directed actions on objects (mainly reaching and grasping). Finally the third phase involves learning about objects and others; the robot's previous experience is used to create expectations on the behavior of other entities (objects as well as intentional agents).

It is important to stress that these classification is not meant to be strict. These three stages in fact are not actually present in the robot; all modules "grow" at the same time; the maturation of each part allows the overall system to perform better but at the same time it increases the possibility of other parts to develop. Thus for instance the ability of the head to perform saccade allows the system to fixate objects and start reaching for them. Arm movements,

although not accurate, in turn allow the system to initiate interaction and improve based on its own mistakes.

The third phase is perhaps the most critical and challenging one as it leads to the development of advanced perceptual abilities. In previous work we have addressed at least some aspects related to this phase (Natale, Rao and Sandini 2002, Fitzpatrick et al. 2003). In this we focus on the first two phases of the developmental process of the robot: learning a body-schema and learning to act.

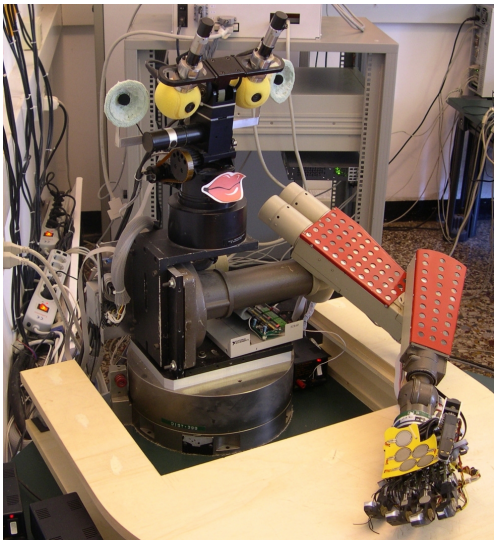


Figure 1 The robotic setup: the Babybot

The robotic setup

The robotic setup is an upper torso humanoid robot composed of a five dof head, a six dof arm and a five-fingered hand (Figure 1). Two cameras are mounted on the head; they can pan independently and tilt together on a common axis. Two additional degrees of freedom allow the head to pan and tilt on the neck. The arm is an industrial manipulator (Unimate Puma 260); it is mounted horizontally to closer mimic the kinematics of a human arm. The hand has seventeen joints distributed as follows: four joints articulate the thumb, whereas index, middle, ring and little fingers have three phalanges each. The fingers are underactuated to reduce the overall number of motors employed. Thus two motors allow the thumb to rotate and flex while two motors are connected to the index finger; finally the remaining fingers are linked and form a single virtual finger that is actuated by two motors only. Intrinsic compliance in all joints allows passive adaptation of the hand to the object being grasped. Magnetic and optic encoders provide position feedback from all phalanges. As far as the sensory system is concerned, the robot is equipped with two cameras, two microphones, and a three axis gyroscope mounted on the head. Tactile feedback is available on the hand; a force sensor allows measuring force and torque at the wrist. Finally proprioceptive feedback is available from the motor encoders. More details about the robot can be found in (Natale 2004).

Learning a body-map

The physical interaction with the environment requires a few prerequisites. To grasp an object the robot must be able to direct gaze to fixate a particular region of the visual field, program a trajectory with the arm to bring it close to the object and eventually grasp it. Although reaching in humans is mostly ballistic, localization of the hand is required to perform fine adjustments at the end of the movement, or, in any case, during learning. We previously addressed the problem of controlling the head to perform smooth pursuit and saccades towards visual and auditory targets (Natale, Metta and Sandini 2002, Metta 2000, Panerai, Metta and Sandini 2002). Here we focus the discussion on the second aspect, that is learning to localize the arm end-point and to segment it out from the rest of the world.

It is known that in humans and primates the brain maintains an internal representation of the body, the relative positions of the limbs, their weight and size. This body-schema is used for planning but, maybe more importantly, also to predict the outcome of an ongoing action and anticipate its consequences. Prediction and anticipation are important aspects of cognition because they extend our ability to understand events by matching our perception and expectations.

Graziano and colleagues (Graziano 1999, Graziano et al. 2000) found neurons in the primate's motor cortex (area 5) which code the position of the hand in the visual field. Tested under different conditions these neurons had receptive fields coding the position of the hand in space; in particular some of them showed to be driven by visual information (that is they fired when the hand was visible), whereas others happened to relay on proprioceptive feedback only (they fired even in those cases when the hand was covered with a barrier).

In infants self-knowledge appears after a few months of development; for instance five-months-old infants are able to recognize their own leg movements on a mirror (Rochat and Striano 2000). But what are the mechanisms used by the brain to build such representation? Pattern similarities between proprioceptive and other sensory feedbacks are cues that could be used to disambiguate between the external world and the body. Indeed, experimental results on infants corroborate the hypothesis that perception of *intermodal form* actually plays a dominant role in the development of self-recognition (Rochat and Striano 2000).

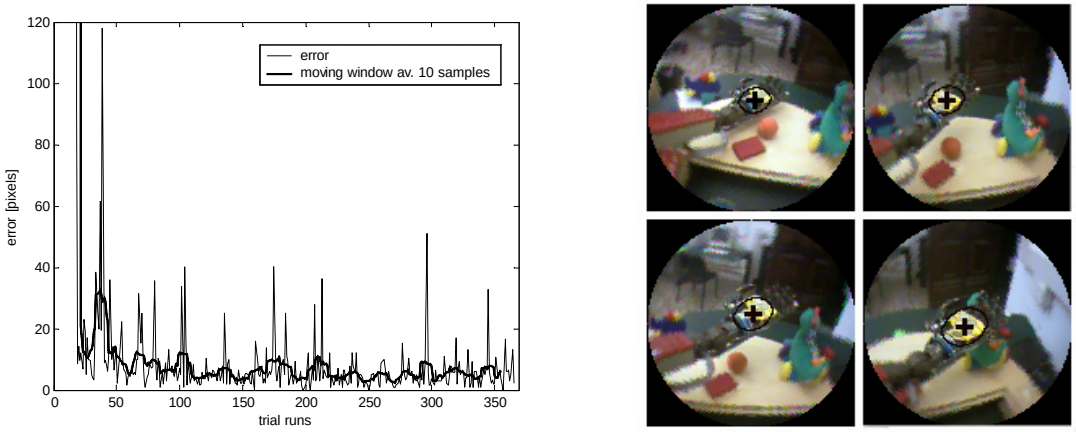


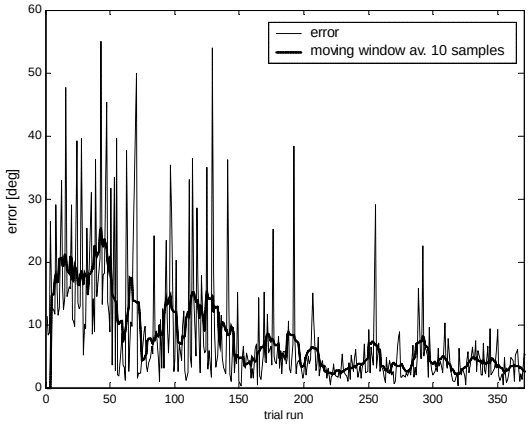
Figure 2 Learning the hand localization. Left: average error in pixels during learning. Right: result of the localization at the end of learning (robot's point of view, left eye).

The problem of learning a body-schema has been addressed in robotics as well. Yoshikawa et al. (Yoshikawa at al. 2003) exploited the idea that the body is invariant with respect to the environment; in their work proprioceptive information is used to train a neural network to segment the arms of a mobile robot. In the case of Metta and Fitzpatrick (Metta and Fitzpatrick 2003) the robot moved the arm in a repetitive way and optic-flow was computed to estimate its motion in the visual field. Cross-correlation between visual and proprioceptive feedback was then used to identify those part of the image which were more likely to be part of the arm end-point. Similarly, in our case the robot moves the wrist to produce small periodic movements of the hand. A simple motion detection algorithm (image difference with adaptive background estimation) is employed to compute motion in the visual field; a zero-crossing algorithm detects the period of oscillation for each pixel. The same periodic information is extracted from the proprioceptive feedback (motor encoders). Pixels which moved periodically and whose period was similar to the one computed in the motor feedback are selected as part of the hand. Instead, the algorithm segments out uncorrelated pixels (e.g. someone walking in the background). The segmentation is a sparse pixel map; a series of low-pass filters at different scale is sufficient to remove outliers and produce a dense binary image.

By using this segmentation procedure the robot can learn to detect its own hand. In particular it builds three models: a color histogram, and two forward models to compute the position and size of the hand in the visual field based on the current arm posture. The latter are two neural networks which provide the expected position, shape and orientation of the hand given the arm proprioceptive feedback. The color histogram is independent (at least to a certain extent) of the orientation and position of the hand and can easily computed out of a single trial. However by accumulating the result of successive experiments it is possible to reduce the noise and increase the accuracy of the histogram. The forward models are trained as follow: the segmentation procedure is repeated several times thus randomly exploring different arm postures. For each trial the center of mass of the segmentation is extracted and used as a training sample for the first neural network. Additional shape information is extracted by fitting a parametric contour on the segmented regions; a good candidate for this purpose is the ellipse because it captures orientation and size of the hand. Accordingly a second neural network is trained to compute the ellipse parameters which fit the hand in the visual field given the current arm posture.

The color histogram gives a statistical description of the color of an object and can be used to spot regions of the image that are more likely to contain the hand. However, the histogram alone is easily fooled by objects that have similar colors. By putting together the contributions of the two neural networks it is possible to reduce the ambiguities and identify precisely the hand in the visual field. Figure 2 reports the result of learning and shows the result of the localization for a few different arm postures.

Overall the hand detection system can be employed in different ways. Since its output is expressed in a retinocentric reference frame the x,y, coordinate of the hand can be sent directly to the controller of the head which can track it as the arm moves in space (see Figure 2). In the next section we will see how this coordinated behavior might be exploited to learn how to reach visually identified objects. Another possibility is to make the robot look at its hand to explore an object that has grasped. This feature may prove helpful especially in case the robot is endowed with foveated vision. Finally by addressing the forward models with desired joint values (a virtual arm position), the robot can predict what will be the position of the hand for a given arm posture; in other words the same mapping used for the hand localization can convert the hand trajectory from joint space to retinal coordinates.



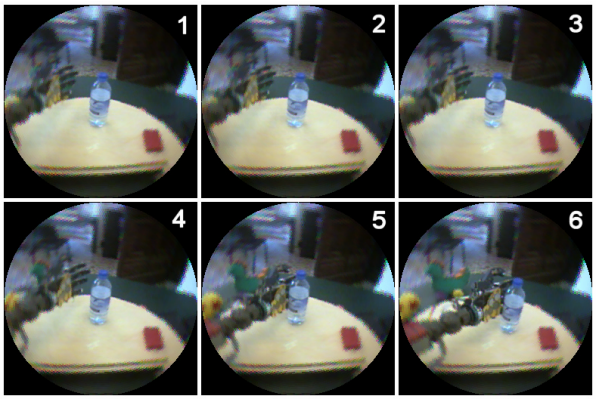


Figure 3 Learning to reach. Left: error during learning, joint angle (squared root of the sum squared error of each joint, in degrees). Right: an exemplar sequence after learning (robot's point of view, left eye).

Learning to reach

Two problems need to be solved to successfully reach for a location in space; the first one is the kinematic transformation between the target position and the corresponding arm posture whereas the second one concerns how to actually generate the motor commands required to achieve that particular posture (inverse dynamics and trajectory generation). In this section we focus on the first problem, that is how to learn the transformation required to compute the joint configuration to reach a specific point in space. Let us assume that the robot is already fixating the target. In this case the fixation point implicitly defines the target for reaching; besides, if the correct angle of vergence has been achieved, the posture of the head defines univocally any position in the three dimensional space (in polar form distance, azimuth and elevation). To solve the task the robot needs the following mapping:

$$\boldsymbol{q}_{arm} = f(\boldsymbol{q}_{head}) \qquad (1.1)$$

where $\boldsymbol{q}_{head}$ is a vector which represents the head posture (target point) and $\boldsymbol{q}_{arm}$ is the corresponding arm joint vector.

Thus reaching starts by first achieving fixation of the object; $\boldsymbol{q}_{head}$ is then used to address the mapping of equation (1.1) and recover the arm motor command $\boldsymbol{q}_{arm}$. Interestingly, the procedure to learn the reaching map is straightforward if we relay on the tracking behavior that was described in the previous section. At the beginning (before starting to reach) the robot explores the workspace by moving the arm randomly while tracking the hand; each pair of arm-head posture defines a training sample to learn equation (1.1) (the reaching map). After enough samples are acquired, the robot can use the reaching map and start performing reaching. However, exploration of the workspace and actual reaching do not need to be separate. If the map is properly initialized (for instance with three values distributed at the center, left and right with respect to the robot) exploration can be achieved by adding noise to the output of the map and activating the tracking of the hand to estimate the actual position of the arm. Properly initialization of the map is required to assert that the value sent to the arm is always meaningful (and safe); the noisy component guarantees the exploration of the workspace. As learning proceeds and new samples are collected the amount of noise (its variance) is progressively reduced to zero to achieve precise reaching. In the experiment reported here, the two methods were interleaved. After reaching the robot performed a few random movements while tracking the hand (the noise in this case had a Gaussian distribution with mean value of 0 degrees and standard deviation of 5 degrees). This strategy is beneficial because it allows to collect more than a single training sample for each reaching trial; besides, in this way, the exploration is biased toward those regions of the space where reaching occurs more often (usually in the part of the workspace in front of the robot).

Once the final arm posture is retrieved from the map it is still necessary to plan a trajectory to achieve it. For this purpose a linear interpolation is carried out between the current and final arm position; the arm command is thus applied in "small steps". The actual torque is computed using a PD controller employing gravity compensation (for details see (Natale 2004)). The complete control schema is reported below (Figure 4).

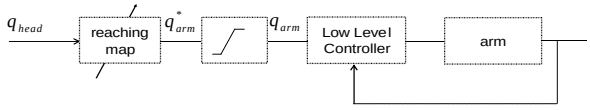


Figure 4 Reaching control schema.

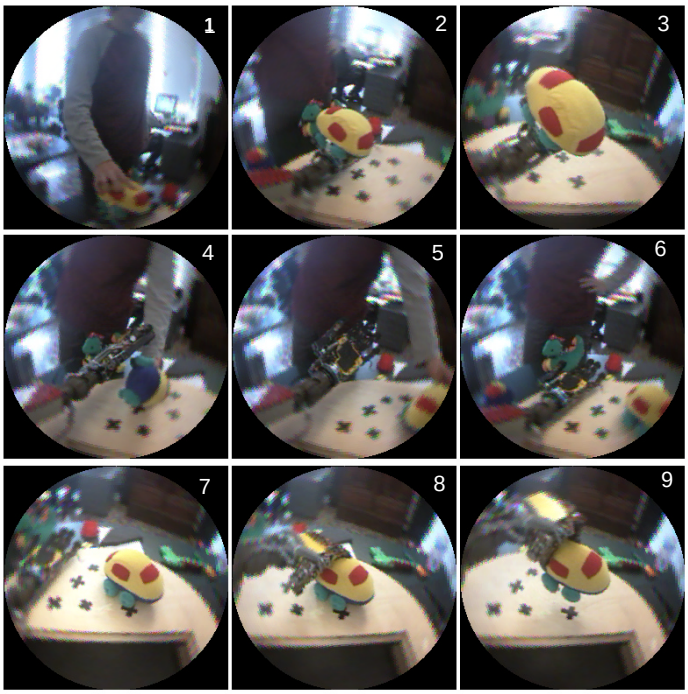


Figure 5. Grasping sequence (robot's point of view, left eye). At frame 1 a human places a toy in the robot's palm. Tactile feedback initiates a clutching action of the hand around the toy, while at the same time, the robot begins moving the eyes to fixate the hand (frames 2 and 3). Once fixation has been achieved a few frames at the center of the cameras are captured to train the object recognition algorithm (frame 3); at frame 4 the toy is released. The robot then starts to search for the toy to grasp it. The object is hence localized, fixated and finally grasped (frames 6-9).

Grasping an object on the table

We present now an experiment to show a possible integration of the modules described in the previous sections. The experiment is still preliminary but it is useful to introduce and illustrate the direction we pursue in our research.

With reference to Figure 5, the experiment starts when an object is placed in the palm of the robot (frame 1). The pressure on the palm elicits a grasping action; the fingers flex toward the palm to close around the object. At this point the robot brings the object close to the eyes while maintaining fixation on the hand (frame 2 and 3). When the object is fixated a few frames are captured at the center of the cameras to train an object recognition algorithm (the details of the object recognition are not relevant here, for a description see (Fitzpatrick 2003)). After the object recognition algorithm is trained the object is released (frame 4). Among the other objects the robot can now spot the one it has seen before, fixate it and finally grasp it (frames 5-9). Haptic information is used to detect if the grasp was successful (mainly the shape of the hand at the end of the grasp); if failure is detected the robot starts looking for the object again and performs another trial, otherwise it waits until another object is placed on the palm.

A few aspects need to be explained in greater detail. The hand motor commands are always preprogrammed; the robot uses three given primitives to close the hand after pressure is detected on the palm, and to pre-shape and flex the fingers around the object during active grasping. The correct positioning of the fingers is achieved by exploiting passive adaptation and the intrinsic elasticity of the hand (see (Natale 2004, Natale, Metta and Sandini 2004). The arm trajectory in also in part preprogrammed to approach the object from above, increasing the probability of success. This is obtained by including waypoints in the joint space relative to the final arm posture (the latter is retrieved from the map as described in the previous section). No other knowledge is required by the robot to perform the task, as the object to be grasped is not known in advance.

Discussion

In this paper we have proposed a possible developmental path for a humanoid robot. The experiments described focus on the steps which allow the robot to learn to reach objects on a table. The knowledge initially provided to the robot consists of a set of stereotyped behaviors, basic

perceptual abilities and learning rules. No prior knowledge about the objects to be grasped is assumed.

The robot starts by learning to control and recognize its own body (control of gaze and arm, hand localization); the first experiment showed how it is possible to build a model of the hand to allow the robot to detect and distinguish it in the environment. In the second experiment we describe how this knowledge can be used to learn to interact with the environment by reaching for objects. A few points are worth stressing. First, learning is online and it is not separated from the normal functioning of the robot. Second, all stages of development are required and equally important. Thus, reaching cannot start (and improve) if gaze is not controlled or the robot has not learnt to localize the hand.

Learning to act is an essential requirement to start interaction with the environment. By properly moving the arm in the workspace, in fact, the robot can try simple actions like pushing or pulling an object on a table. Even this simple form of interaction proves sufficient for developing more sophisticated perceptual abilities. This was shown in some of our previous works (Natale, Rao and Sandini 2002, Fitzpatrick et al. 2003, Metta and Fitzpatrick 2003) where we illustrated how a humanoid robot can learn to push/pull an object in different directions or even imitate pushing/pulling actions performed by another agent (a human). This stresses once more how important is the physical interaction between the agent and the world during ontogenesis; the motor repertoire of actions that is discovered and learnt by the agent in fact constitutes a reference system that can be used to map events that happen in the environment thus adding meaning to them. For this reason we believe that learning to act is at the basis of higher level functions like action/event recognition, interpretation and imitation.

Finally, in the third experiment, we show how the motor and perceptual abilities developed in these initial stages can be integrated meaningfully. The resulting behavior allows the robot to autonomously acquire visual and haptic information about objects in the environment (another experiment is this direction is reported in (Natale, Metta and Sandini 2004)).

Acknowledgments

The work described in this paper has been supported by the EU Projects ADAPT (IST 2001-37173), MIRROR (IST-2000-28159) and RobotCub (IST 2004-004370).

References

Fitzpatrick, P., 2003. From First Contact to Close Encounters: A Developmentally Deep Perceptual System for a Humanoid Robot. Ph.D. diss., Massachusetts Institute of Technology.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S. and Sandini, G., 2003. Learning About Objects Through Action: Initial Steps Towards Artificial Cognition, IEEE International Conference on Robotics and Automation (ICRA 2003), Taipei, Taiwan.

Graziano, M.S.A., 1999. Where is my arm? The relative role of vision and proprioception in the neuronal representation of limb position. *Proceedings of the National Academy of Science*, 96: 10418-10421.

Graziano, M.S.A., Cooke, D.F. and Taylor, C.S.R., 2000. Coding the location of the arm by sight. *Sience*, 290: 1782-1786.

Metta, G., 2000. Babybot: a Study on Sensori-motor Development. Ph.D. diss., University Of Genoa.

Metta, G. and Fitzpatrick, P., 2003. Early Integration of Vision and Manipulation. *Adaptive Behavior*, 11(2): 109-128.

Natale, L., 2004. Linking action to perception in a humanoid robot: a developmental approach to grasping. PhD diss., University Of Genoa.

Natale, L., Metta, G. and Sandini, G., 2002. Development of Auditory-evoked Reflexes: Visuo-acoustic Cues Integration in a Binocular Head. *Robotics and Autonomous Systems*, 39(2): 87-106.

Natale, L., Metta, G. and Sandini, G., 2004. Learning haptic representation of objects, International Conference of Intelligent Manipulation and Grasping, Genoa, Italy.

Natale, L., Rao S. and Sandini, G., 2002. Learning to act on objects, Second International Workshop, BMCV 2002. Lecture Notes in Computer Science. Springer, Tubingen, Germany, 567-575.

Panerai, F., Metta, G. and Sandini, G., 2002. Learning Stabilization Reflexes in Robots with Moving Eyes. *Neurocomputing*, 48(1-4): 323-337.

Rochat, P. and Striano, T., 2000. Perceived self in infancy. *Infant Behavior & Development*, 23: 513-530.

von Hofsten, C., 1983. Catching skills in infancy. Experimental Psychology: *Human Perception and Performance*, 9: 75-85.

von Hofsten, C., Vishton, P., Spelke, E.S., Feng, Q. and Rosander, K., 1998. Predictive action in infancy: tracking and reaching for moving objects. *Cognition*, 67(3): 255-285.

Yoshikawa, Y., Hosoda, K. and Asada, M., 2003. Does the invariance in multi-modalities represent the body scheme ? - a case study with vision and proprioception -, 2nd Intelligent Symposium on Adaptive Motion of Animals and Machines, Kyoto, Japan.