Robot hand discovery based on visuomotor coherence

Ryo Saegusa, Giorgio Metta, Giulio Sandini

Abstract—This paper proposes a plausible approach for a robot to discover its own body based on the coherence of two different sensory feedbacks: vision and proprioception. The image cues of a moving region are stored in an image base with a visuo proprioceptional coherence label. The existence of coherence between the vision and proprioception suggests that the visually detected object is correlated to its own motor functions. By making the image base autonomously, a humanoid robot discovers its own hand without any knowledge of the hand appearances such as predefined visual marker. Also, the robot keeps tracking the hand with distinguishing it from other objects. All modules of visual and proprioceptional processing are distributed in the networks, which allow online perception and interaction.

I. INTRODUCTION

How can a robot know its own hand? This is a fundamental question for embodied intelligence and also the early life of primates. We can recognize our body in general; however it would be hard to assume that we are programmed to recognize all of our body elements inherently. Finding our own body and knowing their sensorimotor functions seem more cognitive and developmental processes. Our main interest in this work is to realize a human like cognitive system allowing visuomotor coordination to perceive the self. The autonomous body discovery is considered essential for robots to sense the boundary between the body and the environment, which would be necessary for general object recognition and visuomotor imitation.

The overview of our approach is depicted in Fig.1. The principal idea is to simply move a part of the body, here we assume a robot hand as the target body part, and monitor the coherence of the visual and proprioceptional feedbacks in sensing. Every moment the image patches of moving region are stored in an image base with the visuomotor coherence label, and visually classified into a certain number of clusters online. The most motor correlated image cluster, here, should be the image cluster of the hand. Making use of the image base, a region of interest in the view is recognized as the hand or other objects by referring the metric to the representatives of image clusters.

This paper is organized as follows: Section II describes the related works. Section III describes the proposed framework and details. Section IV describes the experimental results with the humanoid robot James [1]. Section V gives the conclusion and outlines some future tasks.

II. RELATED WORKS

The sensorimotor coordination is well studied in robotics and there are many excellent works relevant to the author’s interests; sensory prediction (Wolpert et al.[2], Kawato et al.[3]) and learning-based motor control (Atkeson et al.[4][5], Schaal et al.[6]). Also, we have been ambitiously studied constructive understanding of human cognition with humanoid robots, such as the mirror system (Metta et al.[7]) and production-perception link (Fitzpatrick et al.[8]) involving neuroscientific aspects and developmental psychology. Literature of imitation learning (Schaal et al.[9], Calinon et al.[10][11]) are also related to this topic.

In studies on sensorimotor coordination, the body detection is often hand coded with predefined detection rules such as visual markers or knowledge of the body structure. The predefined rules gives robustness in detection to the system as well as certain limits. Let us assume a manipulation task using five robot fingers. Probably we need to suppose how the robot hand is visually projected on the view, or set up five color markers to distinguish each finger. Moreover, when we challenge tool manipulation, the hand detection becomes more difficult if the robot still depends on the predefined knowledge.
The developmental approach for body discovery has been studied in [8][12][13][14][15]. The Kemp’s method in [12] made use of the mutual information between the arm position in joint space and attracted object in task space to evaluate both spatial dependency. The spatially joint dependent object is detected as the hand. The methods in [13][14][15] focuses on the temporal dependency rather than space. These methods are based on the image differentiation by the periodic hand movements, and detecting with a known motor frequency.

Our new approach is also in the direction of motion based detection approach, but characterized by the simplified framework using only the visuomotor coherence without motor assumption, and online image base construction allowing real time body detection.

III. METHOD

The body discovery system is composed of three processing modules: visual process, proprioceptional process, and coordination process. In this section, we describe the function of each module.

A. Vision processing

The overview of the visual processing is shown in Fig.2. The visual processing is modularized as a set of cascaded image filters, and distributed in the networks to allow real time processing. All modules are dually structured for two image streams from the left and right eye cameras.

The receptor modules decompose a left/right input image into the basic features as intensity, colors, and edge. An input RGB color image is down sampled and converted to the YUV color format (Y: intensity component, U,V: color components) by linear transformation as follows:

$$y_j = w_{ij}x_j + v_j,$$

where $y = (Y, U, V)^T \in [0, 255]^3$, $x = (R, G, B) \in [0, 255]^3$. The value of the coefficient $w_{ij}$ and $v_i$ are given in TABLE I. The edge intensity is extracted from the intensity component Y using the second order edge filter (3x3 Laplacian filter) with the convolution kernel $k_{ij}$ in TABLE I. The low edge intensity less than a threshold is filtered for noise reduction.

The motion modules make a smoothed motion image to extract several moving regions. The motion detection starts from simple subtraction of the current and previous edge image. Then image is smoothed by convoluting with the Mexican hat kernel as described:

$$k(x, y) = (1 - 2\xi) \exp(-\xi),$$

$$\xi = (x^2 + y^2)/2\sigma^2$$

where the $\sigma$ is a constant parameter. The $M$ positions of the locally high intensity pixel in the smoothed motion image are detected as moving regions at the frame. These points are used as the candidates of attention point to find the hand and objects. This motion-based detection is suppressed when the robot is turning the head. The motion modules monitor the proprioceptional feedback of the head movement, and suppress motion detection when the head is turning.

The track modules track the attracted point in the image. Here, the attracted point is the hand position detected by the image base module (described in Section III-C). The image patches of the previous tracking point is used for local template matching on the intensity Y, colors U and V, and edge E image. The intensity of matching results, denoted saliency, are normalized in each channel and unified into an image by multiplication. The highest salient position is assigned as the next tracking point.

B. Proprioception process

The proprioception feedback is extracted from the velocity sensing of the motor encoders. In general the velocity profile is affected by many factors such as the motor torque, trajectory, gravity force, and external force applied by contact...
The absolute value of the velocity feedback from the encoder goes over the threshold, the proprioceptional motion signal is activated as a constant value 1.0, while the magnitude is lower than the threshold, the proprioceptional motion signal decays in one order delay. The asymmetry of the activate and deactivate helps to check coherence of more frequent proprioceptional signals to less frequent visual signals, allowing to start up the proprioception firstly and keep activation to wait for the delayed visual process.

To define a simple proprioceptional motion feedback from the velocity profile, we reshaped velocity profile like a on-off signal with decay, characterized by the equation:

\[ p(t) = 1, \quad \text{if} \quad |v(t)| \geq v_0, \quad (4) \]

\[ p(t) = r_p \cdot p(t - \delta t), \quad \text{if} \quad |v(t)| < v_0, \quad (5) \]

where \( v(t), p(t), v_0, \) and \( 0 < r_p < 1 \) are the velocity, proprioceptional motion feedback, velocity threshold, and decay rate, respectively. The velocity profile and proprioceptional motion feedback are compared in Fig 3. The modification of the fast rising and slow decay makes it easy to check the coherence of more frequent proprioceptional signals to less frequent visual signals. The proprioceptional feedbacks of joints of interest are unified as one dimensional feedback by taking the maximum value of them.

C. Visuo proprioceptional coordination

Visual motion feedback and proprioceptional motion feedback are synthesized as images cues coupled with a coherence label, and stored in an image base. Overview of the visuo proprioceptional coordination is depicted in Fig.4. The image base accepts the online image registration and the visuo proprioceptional coordination is depicted in Fig.4. Visuo proprioceptional coordination. The set of image cues of the visual moving region is registered in the image base coupled with a coherence label of the proprioceptional motor feedback. The image base updates image clustering online, then assigns the rank of the motor correlation by comparing the average of each cluster’s coherence label. The highest ranked cluster is regarded as the motor correlated image cluster. The others are regarded as the motor noncorrelated clusters. Independent from the registration process, the motor correlation of query image cues are recognized by referring the centroid of the image clusters.

To allow online updating, Kmeans is a classical clustering method with a given cluster number. All data are labeled randomly in the beginning. Then, centroids of the clusters are calculated, and reassigned the nearest centroid cluster’s label. This update is repeated until any label does not change.

In our image clustering, the new member is registered one by one until the total member number reaches the limit number. Let the limit number and cluster number denote \( N \) and \( K \), respectively. After it reaches the limit, when a new member is registered, a member is removed from the image base. The query (new member) is initially assigned the label of the nearest centroid’s cluster. Then, the clusters are updated in the Kmeans manner online. After the update, a member is randomly selected from the cluster of the new member and removed. The last operation functions for keeping the total number constant and avoiding reducing a member from the minor cluster. The metric of the images is measured as the multiplication of the canonical correlation value \( C(I_a, I_b) \) of each image cue of \( e, y, u, \) and \( v \) patches as defined:

\[ C(I_a, I_b) = \frac{I_a(x, y) \cdot I_b(x, y)}{|I_a(x, y)| |I_b(x, y)| + \epsilon} \quad (6) \]

where \( I_a \) and \( I_b \) denote the vectors of images, and \( \epsilon \) is a small positive constant to avoid zero division.

IV. EXPERIMENT

We performed the experiment of robot hand discovery using with a humanoid robot. In this experiment, we also challenged the hand tracking based on the hand detection. Through the movements, the robot was autonomously getting the appearances of the motor correlated object, which is its own hand, based on the visuo proprioceptional coherence. The robot also collects the appearances of motor noncorrelated objects and people, which are considered useful for general object recognition.
Fig. 5. Online image clustering. The clustering is based on the K-means method [16], but modified to allow online updating. The query (new member) is initially assigned the label of the nearest centroid’s cluster. Then, the clusters are updated in the K-means manner. After the update, a member is randomly selected from the cluster of the new member and removed. The last operation functions for keeping the total image number constant and avoiding reducing a member from the minor cluster.

Fig. 6. The humanoid robot James. James was used for experimental validation of the proposed body discovery system. The shoulder motors of the pitch and yaw are actuated to generate hand movements in the view field.

A. Experimental setting

The humanoid robot James [1] is a fixed upper-body robotic platform dedicated to vision-based manipulation studies. It is composed of a 7dof arm with a dexterous 9dof hand and a 7dof head as shown in Fig.6. It is equipped with binocular vision, force/torque sensors, tactile sensors, inertial sensors and motor encoders.

In this experiment, we mainly used the shoulder motors of the pitch and yaw to generate hand movements in the view field, while the other motors such as the wrist and elbow were statically used to make a suitable arm posture. The shoulder encoders were used to sense the proprioceptional feedbacks of the arm movement based on the velocity profile. The visual effects were extracted from the image streams of the left eye camera mounted on the head. The neck also moves in small magnitude and less frequently, but the visual effect of the head turning is suppressed in the motion-based attention as presented above.

The time line of the robot arm activation is shown in Fig.7. The robot activates and deactivates the shoulder motors alternatively, collecting the appearances of the moving objects and its coherence to the proprioceptional feedbacks. During the activation period, the robot arm movements were generated by random motor babbling of shoulder motors. Let one cycle of activation and deactivation denote episode $T$. We also included less frequent small random neck movements. During the neck movements motion detection was suppressed, but object tracking was available. Therefore, the robot keeps attention to its own hand by the track module during the suppression and the hand movement pauses.

Independent from the robot arm activation, a human partner randomly shows some moving objects in each episode. In the experiment, the hand trajectories and the object movements were not cared precisely, but magnitude of movements were cared to make motion-based detection robust.

| TABLE II  |
| EXPERIMENTAL PARAMETERS |
| $\sigma$ | 1.0 |
| $M$ | 5 |
| $r_p$ | 0.5 |
| $v_0$ | 0.05 |
| $\epsilon$ | 0.001 |
| $K$ | 2 |
| $N$ | 1000 |
| $T$ | 20 sec |

Fig. 7. Time line of the robot arm activation. The schedule is composed of same periods denoted episode $T$. In the first half of each episode, the motors of the shoulder pitch and yaw were activated to generate random trajectories by its own hand. The shoulder joint motors were deactivated in the next half of the episode. Independent from the robot motor activation, a human partner randomly shows some moving objects in each episode. In the experiment, the hand trajectories and the object movements were not cared precisely, but magnitude of movements were cared to make motion-based detection robust.

The parameters used in the experiment is presented in TABLE II. The outputs from the visual processing modules are shown in Fig.8, while the hand recognition and tracking are shown in Fig.9, respectively. The readers of this article are recommended to refer the attached video clip to see the system behavior.

B. Results

Fig.10 shows some images stored in the image base at the end of Ep.4. Most of the motor correlated images were the robot hand images, in which we can see the texture of the black-arc-like tendon wires of the arm. On the other hand, the motor noncorrelated images included ball images which the human partner presented in interaction, and also the human’s
V. DISCUSSION AND CONCLUSIONS

This paper proposed an approach of robot body discovery without giving knowledge of the body appearances in advance. The visuo proprioceptional coherence suggests correlation between its own body motion and moving objects. Then, the robot can remember the appearances of the motor correlated object and recognize them as the private movables; the self body.

The current body discovery system allows the binocular object recognition with motion detection suppression for the neck turning. However, their availability should be proved experimentally. We are also interested in increasing the number of image clusters and the number limit of the image base.

The system recognizes motor correlation, but if the robot manipulates the object like a tool, this object would be recognized as a motor correlated object. Probably we should assume the level of the mobility such that the moving object with always the robot arm should be a kind of permanently extended body, like a watch, shoes and gloves, while some tools such as pens and forks would be temporally extended body. This discussion might be also connected to how we should design general imitation learning in the human-robot interaction.

REFERENCES


Fig. 10. Registered images in the image base. The right figures show the registered motor correlated and noncorrelated images. In this experiment, the human trainer showed the green ball often, therefore the green ball with the trainer’s hand appeared in many motor noncorrelated images. Motor correlated images are mainly the robot hand, visually characterized with many cables of tendon which look a black arc.

Fig. 11. Recognition rate. Each episode includes the 40 to 50 times of visual recognition. The episode was down sampled into 10 to 12 samples to evaluate the recognition rate. The complexity of motion trajectories and image textures are considered to have influenced the performance of the recognition.