Robotica (2011) volume 29, pp. 1–21. © Cambridge University Press 2011
doi:10.1017/S0263574710000755

Self-detection in robots: a method based on detecting temporal contingencies[†]

Alexander Stoytchev*

Developmental Robotics Laboratory, Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011-2274, USA
http://www.ece.iastate.edu/~alexs/

(Received in Final Form: November 5, 2010)

SUMMARY

This paper addresses the problem of self-detection by a robot. The paper describes a methodology for autonomous learning of the characteristic delay between motor commands (efferent signals) and observed movements of visual stimuli (afferent signals). The robot estimates its own efferent-afferent delay from self-observation data gathered while performing motor babbling, i.e., random rhythmic movements similar to the primary circular reactions described by Piaget. After the efferent-afferent delay is estimated, the robot imprints on that delay and can later use it to successfully classify visual stimuli as either "self" or "other." Results from robot experiments performed in environments with increasing degrees of difficulty are reported.

KEYWORDS: Self-detection; Self/other discrimination; Developmental robotics; Behavior-based robotics; Autonomous robotics.

1. Introduction

An important problem that many organisms have to solve early in their developmental cycles is how to distinguish between themselves and the surrounding environment. In other words, they must learn how to identify which sensory stimuli are produced by their own bodies and which are produced by the external world. Solving this problem is critically important for their normal development. For example, human infants that fail to develop self-detection abilities suffer from debilitating disorders such as infantile autism and Rett syndrome.[33]

This paper explores a method for autonomous self-detection in robots that was inspired by Watson's work on self-detection in humans. Watson tested the hypothesis that infants perform self-detection based on the temporal contingency between efferent and afferent signals. He showed that 3-month-old infants can learn a temporal filter that treats events as self-generated if and only if they are preceded by a motor command within a small temporal window; otherwise they are treated as environment-generated. The filter, which is sensitive to a specific efferent-afferent delay (also called the perfect contingency), plays an important role in bootstrapping human development.

This paper tests the hypothesis that a robot can autonomously learn its own efferent-afferent delay from self-observation data and use it to detect the visual features of its own body. The paper also evaluates if the self-detection method can be used by the robot to classify visual stimuli as either "self" or "other." The effectiveness of this approach is demonstrated with robot experiments in environments with increasing degree of difficulty, culminating with self-detection in a TV monitor.

Why should robots have self-detection abilities? There are two main reasons. First, computational models of self-detection in robots may be used to improve our understanding of how biological species achieve the same task. Self-detection abilities are highly correlated with the intelligence of different species (see Section 2). While the reasons for this connection have not been adequately explained so far it is nevertheless intellectually stimulating to take even small steps toward unraveling this mystery. Our computational model is well grounded in the literature on self-detection in humans and animals. At this time, however, it would be premature to claim that our model can be used to explain the self-detection abilities of biological organisms.

Second, self-detection abilities may facilitate the creation of super-adaptive robots that can easily change their end effectors or even their entire bodies while still keeping track of what belongs to their bodies for control purposes. Self-reconfigurable robots that are constructed from multiple identical nodes can benefit from these abilities as well. For example, if one of the nodes malfunctions, then the robot can easily detect if it is still attached to its body by observing that it moves in a temporally contingent way with the motors of another node. This may prompt operations such as self-healing and self-repair.

It is important to draw a distinction between self-recognition and self-detection as this paper deals only with the latter. According to the developmental literature, it is plausible that the process of self-recognition goes through an initial stage of self-detection based on detecting temporal contingencies. Self-recognition abilities, however, require a much more detailed representation for the body than the one needed for self-detection. The notion of "self" has many

[†] This paper is based on Chapter V of the author's PhD dissertation from Georgia Tech.[28] This material has not been previously published.
* Corresponding author. E-mail: alexs@iastate.edu

other manifestations.[19] Rochat,[27] for example, has identified five levels of self-awareness as they unfold from the moment of birth to approximately 4–5 years of age. Most of these are related to the social aspects of the self and thus are beyond the scope of this paper.

2. Related Work

2.1. Self-detection in humans

Almost every major developmental theory recognizes the fact that normal development requires "an initial investment in the task of differentiating the self from the external world."[33] This is certainly the case for the two most influential theories of the 20th century: Freud's and Piaget's. Their theories disagree about the ways in which self-detection is achieved, but they agree that the "self" emerges from actual experience and is not innately predetermined.[33]

Modern theories of human development also seem to agree that the self is derived from actual experience. Furthermore, they identify the types of experience that are required for that: efferent-afferent loops that are coupled with some sort of probabilistic estimate of repeatability.

Rochat[27] suggests that there are certain events that are *self-specifying*. These events are unique as they can only be experienced by the owner of the body. The self-specifying events are also multimodal as they involve more than one sensory or motor modality. Rochat explicitly lists the following self-specifying events: "When infants experience their own crying, their own touch, or experience the perfect contingency between seen and felt bodily movements (e.g., the arm crossing the field of view), they perceive something that no one but themselves can perceive." [27, p. 723]

According to ref. [19], the self is defined through action-outcome pairings (i.e., efferent-afferent loops) coupled with a probabilistic estimate of their regularity and consistency. Here is how they describe the emergence of what they call the "existential self", i.e., the self as a subject distinct from others and from the world: "[The] existential self is developed from the consistency, regularity, and contingency of the infant's action and outcome in the world. The mechanism of reafferent feedback provides the first contingency information for the child; therefore, the kinesthetic feedback produced by the infant's own actions form the basis for the development of self. [...] These kinesthetic systems provide immediate and regular action-outcome pairings," see ref. [19, p. 9]

Watson[33] proposes that the process of self-detection is achieved by detecting the temporal contingency between efferent and afferent stimuli. The level of contingency that is detected serves as a filter that determines which stimuli are generated by the body and which ones are generated by the external world. In other words, the level of contingency is used as a measure of selfness. In Watson's own words: "Another option is that imperfect contingency between efferent and afferent activity implies out-of-body sources of stimulation, perfect contingency implies in-body sources, and noncontingent stimuli are ambiguous," see ref. [33, p. 134]

All three examples suggest that the self is discovered quite naturally as it is the most predictable and the most consistent part of the environment. Furthermore, all seem to confirm that the self is constructed from self-specifying events which are essentially efferent-afferent loops or action-outcome pairs. There are many other studies that have reached similar conclusions. See ref. [19] and ref. [21] for an extensive overview of the literature.

At least one study has tried to identify the minimum set of perceptual features that are required for self-detection. Flom and Bahrick[7] showed that five-month-old infants can perceive the intermodal proprioceptive-visual relation on the basis of motion alone when all other information about the infants' legs was eliminated. In their experiments, they fitted the infants with socks that contained luminescent dots. The camera image was preprocessed such that only the positions of the markers were projected on the TV monitor. In this way the infants could only observe a point-light display[18] of their feet on the TV monitor placed in front of them. The experimental results showed that 5-month-olds were able to differentiate between self-produced (i.e., contingent) leg motion and pre-recorded (i.e., noncontingent) motion produced by the legs of another infant. These results illustrate that only movement information alone might be sufficient for self-detection since all other features like edges and texture were eliminated in these experiments. The robot experiments described later use a similar experimental design as the robot's visual system has perceptual filters that allow the robot to see only the positions and movements of specific color markers placed on the robot's body. Similar to the infants in the dotted socks experiments, the robot can only see a point-light display of its movements.

2.2. Self-detection in animals

Many studies have focused on the self-detection abilities of animals. Perhaps the most influential study was performed by Gallup[10], which reported for the first time the abilities of chimpanzees to detect a marker placed surreptitiously on their head using a mirror. Gallup's discovery was followed by a large number of studies that have attempted to test which species of animals can pass the mirror test. Somewhat surprisingly, the number turned out to be very small: chimpanzees, orangutans, and bonobos (one of the four great apes, often called the forgotten ape, see ref. [5]). There is also at least one study that has documented similar capabilities in bottlenose dolphins.[26] Another recent study reported that one Asian elephant (out of three that were tested) conclusively passed the mirror test.[24] Attempts to replicate the mirror test with other primate and nonprimates species have failed.[3, 6, 12, 25]

Gallup[11] has argued that the interspecies differences are probably due to different degrees of self-awareness. Another reason for these differences "may be due to the absence of a sufficiently well-integrated self-concept," see ref. [11, p. 334]. Yet another reason according to ref. [11] might be that the species that pass the mirror test can direct their attention both outward (toward the external world) and inwards (toward their own bodies), i.e., they can become "the subject of [their] own attention." Humans, of course, have the most developed self-exploration abilities and have used them to create several branches of science, e.g., medicine, biology, and genetics.

2.3. *Self-detection in robots*

Self-detection experiments with robots are still rare. One of the few published studies on this subject is described in ref. [20]. They implemented an approach to autonomous self-detection similar to the temporal contingency strategy described by Watson.[33] Their robot was successful in identifying movements that were generated by its own body. The robot was also able to identify the movements of its hand reflected in a mirror as self-generated motion because the reflection obeyed the same temporal contingency as the robot's body.

In that study, the self-detection was performed at the pixel level and the results were not carried over to high-level visual features of the robot's body. Thus, there was no permanent trace of which visual features constitute the robot's body. Because of this, the detection could only be performed when the robot was moving. This limitation was removed in a subsequent study,[17] which used probabilistic methods that incorporate the motion history of the features as well as the motor history of the robot. The new method calculates and uses three dynamic Bayesian models that correspond to three different hypotheses ("self," "animate other," or "inanimate") for what caused the motion of an object. Using this method the robot was also able to identify its image in a mirror as "self." The method was not confused when a person tried to mimic the actions of the robot.

The study presented in this paper is similar to the two studies mentioned above. Similar to ref. [20], it employs a method based on detecting temporal contingencies, but also keeps probabilistic estimates over the detected visual features to distinguish whether or not they belong to the robot's body. In this way, the stimuli can be classified as either "self" or "other" even when the robot is not moving. Similar to ref. [17], it estimates whether the features belong to the robot's body, but uses a different model based on ref. [33] to update these estimates.

The main difference between our approach and previous work can be summarized as follows. Self-detection is ultimately about finding a cause–effect relationship between the robot's motor commands and perceptible visual changes in the environment. Causal relationships are different from probabilistic relationships, see ref. [22, p. 25], which have been used in previous models. The only way to really know if something was caused by something else is to take into account both the necessity and the sufficiency,[22,33] which is what our model does. Humans tend to extract and remember causal relationships and not probabilistic relationships as the causal relationships are more stable, see ref. [22, p. 25]. Presumably, robots should do the same.

Another difference is that our approach has very few tunable parameters so presumably it is easier to implement and calibrate. Also, our model was tested on several data sets lasting 45 min each, which is an order of magnitude longer than any previously published results.

Another team of roboticists has attempted to perform self-detection experiments with robots based on a different self-specifying event: the so-called double touch.[34] The double touch is a self-specifying event because it can only be experienced by the robot when it touches its own body. This event cannot be experienced if the robot touches an object or if somebody else touches the robot since both cases would correspond to a single touch event.

3. Problem Statement

For the sake of clarity, the problem of autonomous self-detection by a robot will be stated explicitly using the following notation. Let the robot have a set of joints $J = \{j_1, j_2, \ldots, j_n\}$ with corresponding joint angles $\Theta = \{q_1, q_2, \ldots, q_n\}$. The joints connect a set of rigid bodies $B = \{b_1, b_2, \ldots, b_{n+1}\}$ and impose restrictions on how the bodies can move with respect to one another. For example, each joint, j_i, has lower and upper joint limits, q_i^L and q_i^U, which are either available to the robot's controller or can be inferred by it. Each joint, j_i, can be controlled by a motor command, $move(j_i, q_i, t)$, which takes a target joint angle, q_i, and a start time, t, and moves the joint to the target joint angle. More than one $move$ command can be active at any given time.

Also, let there be a set of visual features $F = \{f_1, f_2, \ldots, f_k\}$ that the robot can detect and track over time. Some of these features belong to the robot's body, i.e., they are located on the outer surfaces of the set of rigid bodies, B. Other features belong to the external environment and the objects in it. The robot can detect the positions of visual features and detect whether or not they are moving at any given point in time. In other words, the robot has a set of perceptual functions $P = \{p_1, p_2, \ldots, p_k\}$, where $p_i(f_i, t) \rightarrow \{0, 1\}$. That is to say, the function p_i returns 1 if feature f_i is moving at time t, and 0 otherwise.

The goal of the robot is to classify the set of features, F, into either "self" or "other." In other words, the robot must split the set of features into two subsets, F_{self} and F_{other}, such that $F = F_{\text{self}} \cup F_{\text{other}}$.

4. Methodology

The problem of self-detection by a robot is divided into two separate problems as follows:

Subproblem 1: How can a robot estimate its own efferent-afferent delay, i.e., the delay between the robot's motor actions and their perceived effects?

Subproblem 2: How can a robot use its efferent-afferent delay to classify the visual features that it can detect into either "self" or "other"?

The methodology for solving these two subproblems is illustrated by two figures. Figure 1 shows how the robot can estimate its efferent-afferent delay (subproblem 1) by measuring the elapsed time from the start of a motor command to the start of visual movement. The approach relies on detecting the temporal contingency between motor commands and observed movements of visual features. To estimate the delay the robot gathers statistical information by executing multiple motor commands over an extended period of time. It will be shown that this approach is reliable even if there are other moving visual features in the environment as their movements are typically not correlated with the robot's motor commands. Once the delay is estimated the robot imprints on it (i.e., remembers it irreversibly) and uses it to solve subproblem 2.

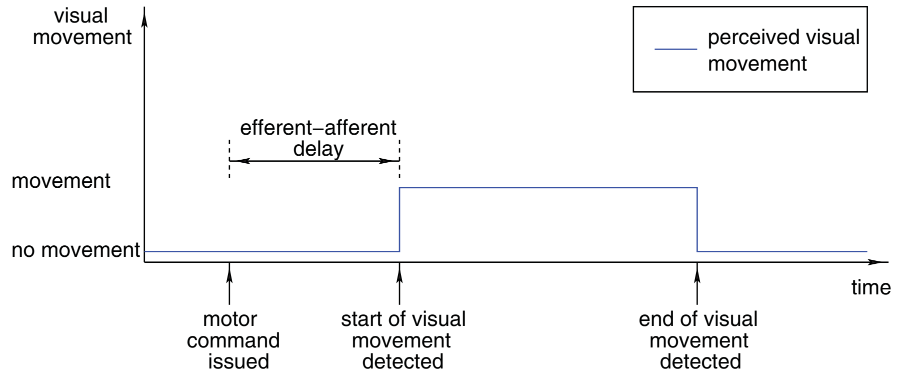

visual movement
perceived visual movement
efferent–afferent delay
movement
no movement
time
motor command issued
start of visual movement detected
end of visual movement detected

Fig. 1. The efferent-afferent delay is defined as the time interval between the start of a motor command (efferent signal) and the detection of visual movement (afferent signal). The robot can learn this characteristic delay (also called the perfect contingency) from self-observation data.

Figure 2 shows how the estimated efferent-afferent delay can be used to classify visual features as either "self" or "other" (subproblem 2). The figure shows three visual features and their detected movements over time represented by red, green, and blue lines. Out of these three features only feature 3 (blue) can be classified as "self" as it is the only one that conforms to the perfect contingency. Feature 1 (red) begins to move too late after the motor command is issued and feature 2 (green) begins to move too soon after the movement command is issued.

A classification based on a single observation can be unreliable due to sensory noise or a lucky coincidence in the movements of the features relative to the robot's motor command. Therefore, the robot maintains a probabilistic estimate for each feature as to whether or not it is a part of the robot's body. The probabilistic estimate is based on the sufficiency and necessity indices proposed by Watson.[33] The sufficiency index measures the probability that the stimulus (visual movement) will occur during some specified period of time after the action (motor command). The necessity index, on the other hand, measures the probability that the action (motor command) was performed during some specified period of time before the stimulus (visual movement) was observed. The robot continuously updates these two indexes for each feature as new evidence becomes available. Features for which both indexes are above a certain threshold are classified as "self." All others are classified as "other." Section 7 provides more details about this procedure.

5. Experimental Setup

5.1. Detecting visual features

All experiments in this paper were performed using the CRS Plus robot arm shown in Fig. 3. The movements of the robot were restricted to the vertical plane. In other words, only joints 2, 3, and 4 (i.e., shoulder pitch, elbow pitch, and wrist pitch) were allowed to move. Joints 1 and 5 (i.e., waist roll and wrist roll) were disabled and their joint angles were set to 0.

Six color markers (also called body markers) were placed on the body of the robot as shown in Fig. 3. Each marker is assigned a number which is used to refer to this marker in the text and figures that follow. From the shoulder to the wrist the markers have the following IDs and colors: (0) dark orange; (1) dark red; (2) dark green; (3) dark blue; (4)

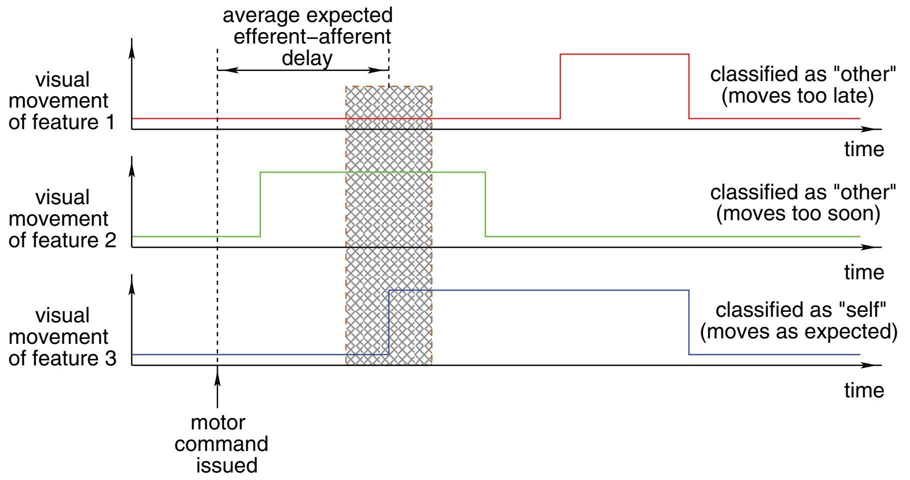

average expected efferent–afferent delay
visual movement of feature 1
classified as "other" (moves too late)
time
visual movement of feature 2
classified as "other" (moves too soon)
time
visual movement of feature 3
classified as "self" (moves as expected)
time
motor command issued

Fig. 2. "Self" versus "Other" discrimination. Once the robot has learned its efferent-afferent delay it can use its value to classify the visual features that it can detect into "self" and "other." In this figure, only feature 3 (blue) can be classified as self as it starts to move after the expected efferent-afferent delay plus or minus some tolerance (shown as the brown region). Features 1 and 2 are both classified as "other" since they start to move either too late (feature 1) or too soon (feature 2) after the motor command is issued.

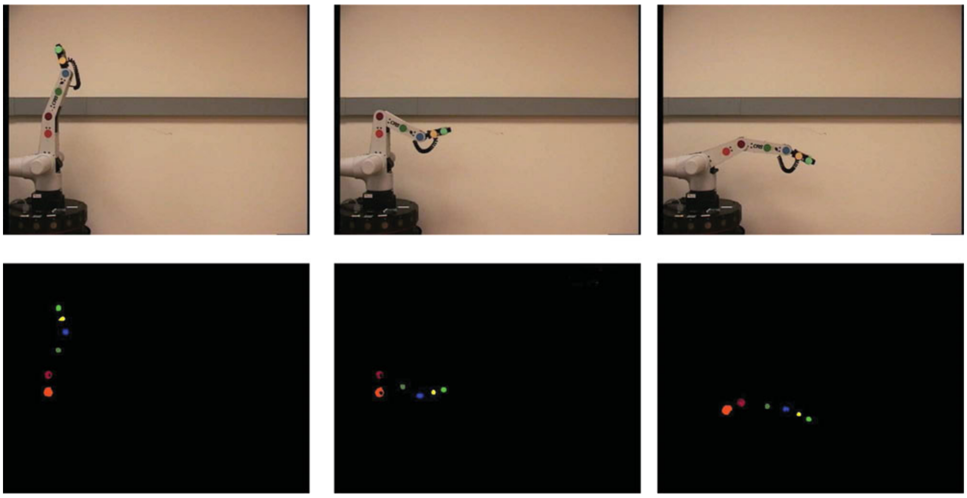


Fig. 3. (Top row) Several of the robot poses selected by the motor babbling procedure. (Bottom row) Color segmentation results for the same robot poses.

yellow; (5) light green. The body markers were located and tracked using color segmentation (see Fig. 3). The position of each marker was determined by the centroid of the largest blob that matched the specific color. The color segmentation was performed using a computer vision code that performs histogram matching in HSV color space with the help of the openCV library (an open source computer vision package). The digital video camera (Sony EVI-D30) was mounted on a tripod and its field of view was adjusted so that it can see all body markers in all possible joint configurations of the robot. The image resolution was set to 640 × 480. For all experiments described in this paper the frames were captured at 30 frames per second.

5.2. *Motor Babbling*

All experiments described in this paper rely on a common motor babbling procedure, which allows the robot to gather self-observation data (both visual and proprioceptive) while performing random joint movements. This procedure consists of random joint movements similar to the primary circular reactions described by Piaget[23] as they are not directed at any object in the environment. Algorithm 1 shows the pseudocode for the motor babbling procedure.

During motor babbling the robot's controller randomly generates a target joint vector and then tries to move the robot to achieve this vector. The movements are performed by adjusting each joint in the direction of the target joint angle. If the target joint vector cannot be achieved within some tolerance (2 degrees per joint was used) then after some timeout period (8 s was used) the attempt is aborted and another random joint vector is chosen for the next iteration. The procedure is repeated for a specified number of iterations (500 iterations were used).

5.3. *Visual movement detection*

For each image frame a color marker was declared to be moving if its position changed by more than 1.5 pixels during the 0.1 s interval immediately preceding the current frame. The timing intervals were calculated from the timestamps of the frames stored in the standard UNIX format.

The result of this tracking technique is a binary 0/1 signal for each of the currently visible markers, similar to the graphs shown in Fig. 2. These signals are still slightly noisy and therefore they were filtered with a box filter (also called averaging filter) of width 5, which corresponds to smoothing each tracking signal over five consecutive frames. The filter changes the values of the movement detection signal to the average for the local neighborhood. For example, if the movement detection signal is 001011100 then the filter will output 000111100. On the other hand, if the sequence is 001000 or 001100 then the filter will output 000000.

Algorithm 2 shows the pseudocode for the movement detector and the box filter.

6. Experimental Results: Learning the Efferent-Afferent Delay

This section describes the procedure used to estimate the efferent-afferent delay of the robot as well as the experimental conditions used to test it. The pseudocode for the procedure is shown in Algorithm 3. The algorithm uses the results from the motor babbling procedure described in Section 5.2, i.e., it uses the array of motor commands and their timestamps. It also uses the results from the movement detection method described in Section 5.3, i.e., it uses the number of captured frames and the *MOVE* array which holds information about what feature was moving during which frame. The algorithm is presented in batch form but it is straightforward to rewrite it in incremental form.

The algorithm maintains a histogram of the measured delays over the interval [0, 6) s. Delays longer than 6 s are ignored. Each bin of the histogram corresponds to $1/30^{-\text{th}}$ of a second, which is equal to the time interval between two consecutive frames. For each frame the algorithm checks which markers, if any, are starting to move during that frame. This information is already stored in the *MOVE* array, which is returned by the MOVEMENT DETECTOR function in Algorithm 2. If the start of a movement is detected, the algorithm finds the last motor command that was executed prior to the current frame. The timestamp of the last motor

Algorithm 1 Motor Babbling

```
GetRandomJointVector($robot$)
$nJoints \leftarrow robot$.GetNumJoints()
for $j \leftarrow 0$ to $nJoints$ do
    $moveThisJoint \leftarrow$ RandomInt(0,1)
    if $moveThisJoint = 1$ then
        $lowerLimit \leftarrow robot$.GetLowerJointLimit($j$)
        $upperLimit \leftarrow robot$.GetUpperJointLimit($j$)
        $JV[j] \leftarrow$ RandomFloat($lowerLimit, upperLimit$)
    else
        // Keep the the current joint angle for this joint.
        $JV[j] \leftarrow robot$.GetCurrentJointAngle($j$)
    end if
end for
return $JV$

IsRobotAtTargetJointVector($robot, targetJV, tolerance$)
$nJoints \leftarrow robot$.GetNumJoints()
for $j \leftarrow 0$ to $nJoints$ do
    $dist \leftarrow$ ABS($targetJV[j]$ - $robot$.GetCurrentJointAngle($j$))
    if $dist > tolerance$ then
        return $false$
    end if
end for
return $true$

MotorBabbling($robot, nIterations, timeout, tolerance, sleepTime$)
for $i \leftarrow 0$ to $nIterations$ do
    $motor[i].targetJV \leftarrow$ GetRandomJointVector($robot$)
    $motor[i].timestamp \leftarrow$ GetTime()
    repeat
        $robot$.MoveToTargetJointVector($motor[i].targetJV$)
        Sleep($sleepTime$)
        if (GetTime() - $motor[i].timestamp$) $> timeout$ then
            // Can't reach that joint vector. Try another one on the next iteration.
            break
        end if
        $done \leftarrow$ IsRobotAtTargetJointVector($robot$, $motor[i].targetJV$, $tolerance$)
    until $done = true$
end for
return $motor$
```

command is subtracted from the timestamp of the current frame and the resulting delay is used to update the histogram. Only one histogram update per frame is allowed, i.e., the bin count for only one bin is incremented by one. This restriction ensures that if there is a large object with many moving parts in the robot's field of view the object's movements will not bias the histogram and confuse the detection process. The pseudocode for the histogram routines is given in ref. [28].

The bins of the histogram can be viewed as a bank of delay detectors each of which is responsible for detecting only a specific timing delay. It has been shown that biological brains have a large number of neuron-based delay detectors specifically dedicated to measuring timing delays.[8,16] Supposedly, these detectors are fine tuned to detect only specific timing delays, just like the bins of the histogram.

After all delays are measured the algorithm finds the bin with the largest count, which corresponds to the peak of the histogram. To reduce the effect of noisy histogram updates, the histogram is thresholded with an empirically derived threshold equal to 50% of the peak value. For example, if the largest bin count is 200, then the threshold will be set to 100. After thresholding, the mean delay can be estimated by multiplying the bin count of each bin with its corresponding delay, then adding all products and dividing the sum by the total bin count.

The value of the mean delay by itself is not very useful, however, as it is unlikely that other measured delays will have the exact same value. In order to classify the visual features as either "self" or "other" the measured delay for the feature must be within some tolerance interval around the mean. This

Algorithm 2 Movement Detection

```
IsMoving($markerID, treshold, imageA, imageB$)
$posA \leftarrow$ FindMarkerPosition($markerID, imageA$)
$posB \leftarrow$ FindMarkerPosition($markerID, imageB$)
$\Delta x \leftarrow posA.x - posB.x$
$\Delta y \leftarrow posA.y - posB.y$
$dist \leftarrow \sqrt{(\Delta x)^2 + (\Delta y)^2}$
if $dist > threshold$ then
    return 1
else
    return 0
end if

BoxFilter($sequence[\,][\,], index, m$)
$sum \leftarrow 0$
for $i \leftarrow index - 2$ to $index + 2$ do
    $sum \leftarrow sum + sequence[i][m]$
end for
if $sum \geq 3$ then
    return 1
else
    return 0
end if

MovementDetector($nFrames, \Delta t, treshold$)
// Buffer some frames in advance so the BoxFilter can work OK
for $i \leftarrow 0$ to 3 do
    $frame[i].image \leftarrow$ GetNextFrame()
    $frame[i].timestamp \leftarrow$ GetTime()
end for
for $i \leftarrow 4$ to $nFrames$ do
    $frame[i].image \leftarrow$ GetNextFrame()
    $frame[i].timestamp \leftarrow$ GetTime()
    // Find the index, k, of the frame captured $\Delta t$ seconds ago
    $startTS \leftarrow frame[i].timestamp$ - $\Delta t$
    $k \leftarrow index$
    while (($frame[k].timestamp < startTS$) and ($k > 0$)) do
        $k \leftarrow k - 1$
    end while

    // Detect marker movements and filter the data
    for $m \leftarrow 0$ to $nMarkers$ do
        $MOVE[i][m]$       $\leftarrow$ IsMoving($m, treshold, frame[i].image, frame[k].image$)
        $MOVE[i-2][m] \leftarrow$ BoxFilter($MOVE, i-2, m$)
    end for
end for
return $MOVE$
```

interval was shown as the brown region in Fig. 2. One way to determine this tolerance interval is to calculate the standard deviation of the measured delays, σ, and then classify a feature as "self" if its movement delay, d, lies within one standard deviation of the mean, μ. In other words, the feature is classified as "self" if $\mu - \sigma \leq d \leq \mu + \sigma$.

The standard deviation can be calculated from the histogram. Because the histogram is thresholded, however, this estimate will not be very reliable as some delays that are not outliers will be eliminated. In this case, the standard deviation will be too small to be useful. On the other hand, if the histogram is not thresholded the estimate for the standard deviation will be too large to be useful as it will be calculated over the entire data sample which includes the outliers as well. Thus, the correct estimation of the standard deviation is not a trivial task. This is especially true when the robot is not the only moving object in the environment.

Fortunately, the psychophysics literature provides an elegant solution to this problem. It is well know that, the discrimination abilities for timing delays in both animals and

Algorithm 3 Learning the efferent-afferent delay

```
CALCULATE EFFERENT AFFERENT DELAY($nFrames$, $frame[\,]$, $MOVE[\,][\,]$, $motor[\,]$)
// Skip the frames that were captured prior to the first motor command.
$start \leftarrow 1$
while $frame[start].timestamp < motor[0].timestamp$ do
    $start \leftarrow start + 1$
end while

// Create a histogram with bin size=$1/30^{-th}$ of a second
// for the time interval $[0, 6)$ seconds.
$hist \leftarrow$INITHISTOGRAM(0.0, 6.0, 180)

$idx \leftarrow 0$ // Index into the array of motor commands
for $k \leftarrow start$ to $nFrames - 1$ do
    // Check if a new motor command has been issued.
    if $frame[k].timestamp > motor[idx+1].timestamp$ then
        $idx \leftarrow idx + 1$
    end if

    for $i \leftarrow 0$ to $nMarkers - 1$ do
        // Is this a $0 \rightarrow 1$ transition, i.e., start of movement?
        if $((MOVE[k-1][i] = 0)$ and $(MOVE[k][i] = 1))$ then
            $delay \leftarrow frame[k].timestamp - motor[idx].timestamp$
            $hist$.ADDVALUE($delay$)
            break // only one histogram update per frame is allowed
        end if
    end for
end for

// Threshold the histogram at 50% of the peak value.
$maxCount \leftarrow hist$.GETMAXBINCOUNT()
$threshold \leftarrow maxCount/2.0$
$hist$.THRESHOLD($threshold$)

efferent-afferent-delay $\leftarrow hist$.GETMEAN()
return efferent-afferent-delay
```

humans obey Weber's law.[13,30,31] This law is named after the German physician Ernst Heinrich Weber (1795–1878) who was one of the first experimental psychologists. Weber observed that the sensory discrimination abilities of humans depend on the magnitude of the stimulus that they are trying to discriminate against. The law can be stated as $|\frac{\Delta I}{I}| = c$, where I represents the magnitude of some stimulus, ΔI is the value of the just noticeable difference (JND), and c is a constant that does not depend on the value of I. The fraction $\frac{\Delta I}{I}$ is known as the Weber fraction. The law implies that the difference between two signals is not detected if that difference is less than the Weber fraction.

Weber's law can also be used to predict if the difference between two stimuli I and I' will be detected. The stimuli will be indistinguishable if the following inequality holds $|\frac{I-I'}{I}| < c$, where c is a constant that does not depend on the values of I and I'.

A similar discrimination rule is used in the robot experiments: $|\frac{\mu-d}{\mu}| < \beta$, where μ is the mean efferent-afferent delay, d is the currently measured delay between a motor command and perceived visual movement, and β is a constant that does not depend on μ.

Weber's law applies to virtually all sensory discrimination tasks in both animals and humans, e.g., distinction between colors and brightness,[1] distances, sounds, weights, and time.[13,30,31] Furthermore, in timing discrimination tasks the just noticeable difference is approximately equal to the standard deviation of the underlying timing delay, i.e., $\frac{\sigma}{\mu} = \beta$. Distributions with this property are know as scalar distributions because the standard deviation is a scalar multiple of the mean.[13] This result has been used in some of the most prominent theories of timing interval learning, e.g., refs. [9, 13–15].

Thus, the problem of how to reliably estimate the standard deviation of the measured efferent-afferent delay becomes trivial. The standard deviation is simply equal to a constant multiplied by the mean efferent-afferent delay, i.e., $\sigma = \beta\mu$. The value of the parameter β can be determined empirically. For timing discrimination tasks in pigeons its value has been estimated at 30%, i.e., $\frac{\sigma}{\mu} = 0.3$, see ref. [4, p. 22]. Other

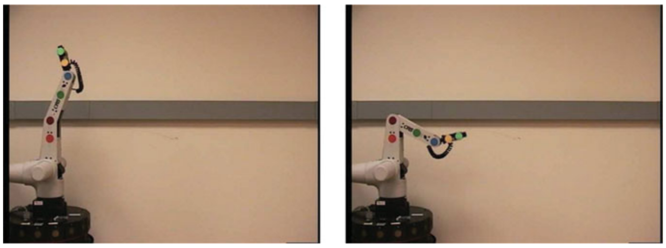


Fig. 4. Frames from a test sequence in which the robot is the only moving object.

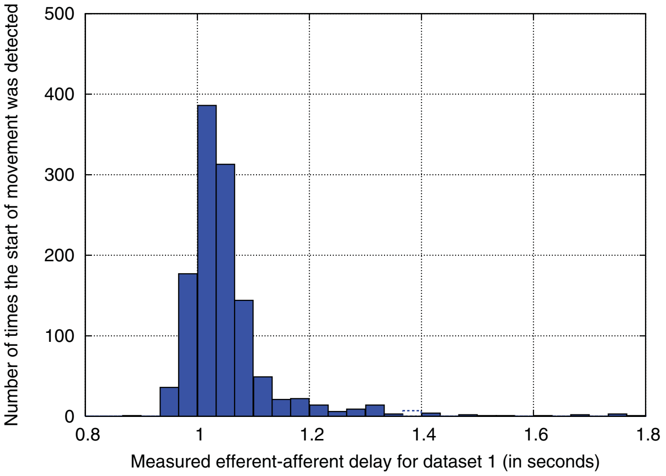


Fig. 5. Histogram for the measured efferent-afferent delays in data set 1.

estimates for different animals range from 10% to 25% see ref. [31, p. 328]. In the robot experiments described below the value of β was set to 25%.

6.1. *Test case with a single robot*

The first set of experiments tested the algorithm under ideal conditions when the robot is the only moving object in the environment (see Fig. 4). The experimental data consists of two data sets, which were collected by running the motor babbling procedure for 500 iterations. For each data set the entire sequence of frames captured by the camera were converted to JPG files and saved to disk. The frames were recorded at 30 frames per second at a resolution of 640×480 pixels and processed offline. Each data set corresponds roughly to 45 min of wall clock time. This time limit was selected so that the data for one data set can fit on a single DVD with storage capacity of 4.7 GB. Each frame also has a timestamp denoting the time at which the frame was captured. The motor commands (along with their timestamps) were also saved as a part of the data set.

Figure 5 shows a histogram for the measured efferent-afferent delays in data set 1 (the results for data set 2 are similar). Each bin of the histogram corresponds to $1/30^{-\text{th}}$ of a second, which is equal to the time between two consecutive frames. As can be seen from the histogram, the average measured delay is approximately 1 s. This delay may seem relatively large but is unavoidable due to the slowness of the robot's controller. A robot with a faster controller may have a shorter delay. For comparison, the average efferent-afferent delay reported in ref. [20] for a more advanced robot was 0.5 s.

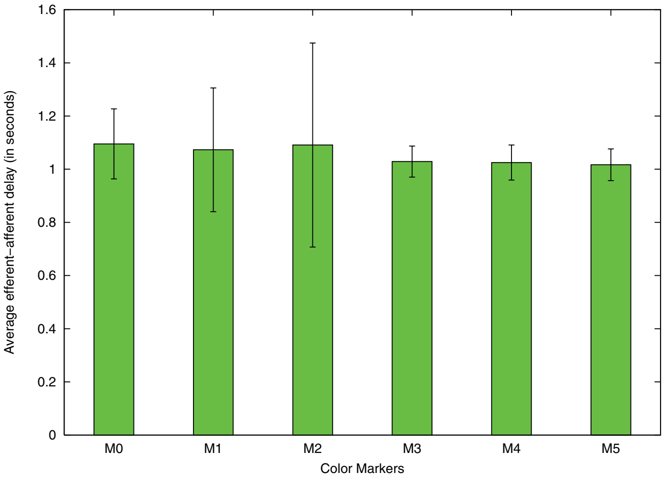


Fig. 6. The average efferent-afferent delay and its corresponding standard deviation for each of the six body markers calculated using data set 1. In this figure only, the standard deviation was calculated using the raw data without using the Weber fraction.

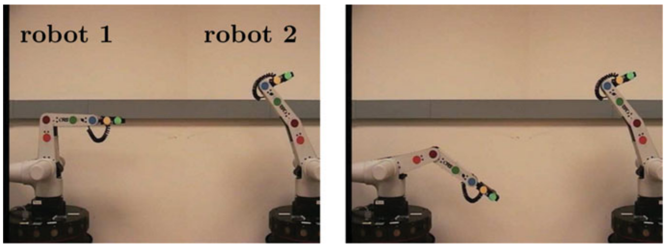


Fig. 7. Frames from a test sequence with two robots in which the movements of the robots are uncorrelated. Each robot is controlled by a separate motor babbling routine. The robot on the left (robot 1) is the one trying to estimate its own efferent-afferent delay.

The measured delays are also very consistent across different body markers. Figure 6 shows the average measured delays for each of the six body markers as well as their corresponding standard deviations in data set 1. As expected, all markers have similar delays and the small variations between them are not statistically significant.

Algorithm 3 estimated the following efferent-afferent delays for each of the two data sets: 1.02945 s (for data set 1) and 1.04474 s (for data set 2). The two estimates are very close to each other. The difference is less than $1/60^{-\text{th}}$ of a second, or half a frame.

6.2. *Test case with two robots: uncorrelated movements*

This experiment was designed to test whether the robot can learn its efferent-afferent delay in situations in which the robot is not the only moving object in the environment. In this case, another robot arm was placed in the field of view of the first robot (see Fig. 7). A new data set with 500 motor commands was generated.

Because there was only one robot available to perform this experiment the second robot was generated using a digital video special effect. Each video frame containing two robots is a composite of two other frames with only one robot in each (these frames were taken from the two data sets described in Section 6.1). The robot on the left (robot 1) is in the same position as in the previous data sets. To get the robot on the

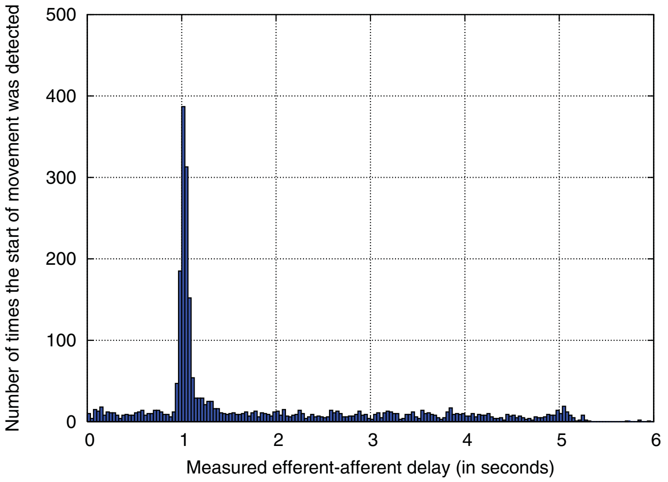


Fig. 8. Histogram for the measured delays between motor commands and observed visual movements in the test sequence with two robots whose movements are uncorrelated (see Fig. 7).

right (robot 2), the left part of the second frame was cropped, flipped horizontally, translated and pasted on top of the right part of the first frame.

Similar experimental designs are quite common in self-detection experiments with infants (e.g., ref. 2, 33). In these studies the infants are placed in front of two TV screens. On the first screen the infants can see their own leg movements captured by a camera. On the second screen they can see the movements of another infant recorded during a previous experiment.

Under this test condition the movements of the two robots are uncorrelated. The frames for this test sequence were generated by combining the frames from data set 1 and data set 2 (described in Section 6.1). The motor commands and all frames for robot 1 come from data set 1; the frames for robot 2 come from data set 2. Because the two motor babbling sequences have different random seed values the movements of the two robots are uncorrelated. In this test, robot 1 is the one that is trying to estimate its efferent-afferent delay.

Figure 8 shows a histogram for the measured delays in this sequence. As can be seen from the figure, the histogram has some values for almost all of its bins. Nevertheless, there is still a clearly defined peak that has the same shape and position as in the previous test cases, which were conducted under ideal conditions. The algorithm estimated the efferent-afferent delay at 1.02941 s after the histogram was thresholded with a threshold equal to 50% of the peak value.

Because the movements of robot 2 are uncorrelated with the motor commands of robot 1 the detected movements for the body markers of robot 2 are scattered over all bins of the histogram. Thus, the movements of the second robot could not confuse the algorithm into picking a wrong value for the mean efferent-afferent delay. The histogram shows that these movements exhibit almost an uniform distribution over the interval from 0 to 5 s. The drop off after 5 s is due to the fact that robot 1 performs a new movement approximately every 5 s. Therefore, any movements performed by robot 2 after the 5-s interval will be associated with the next motor command of robot 1.

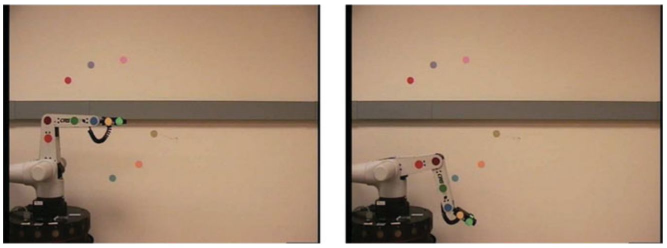


Fig. 9. Frames form a test sequence with six static background markers.

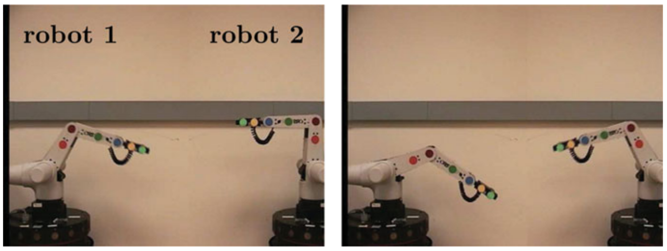


Fig. 10. Frames from a test sequence with two robots in which the robot on the right mimics the robot on the left. The mimicking delay is 20 frames (0.66 s).

6.3. Test case with a single robot and static background features

This experimental setup tested Algorithm 3 in the presence of static visual features placed in the environment. In addition to the robot's body markers, six other markers were placed on the background wall (see Fig. 9). All background markers remained static during the experiment, but it was possible for them to be occluded temporarily by the robot's arm. Once again, the robot was controlled using the motor babbling procedure. A new data set with 500 motor commands was collected using the procedure described in Section 6.1.

The histogram for this data set, which is not shown here due to space limitations but is given in ref. [28], is similar to the histograms shown in the previous subsection. Once again almost all bins have some values. This is due to the detection of false positive movements for the background markers due to partial occlusions that could not be filtered out by the box filter.

These false positive movements exhibit an almost uniform distribution over the interval from 0 to 5 s. This is to be expected as they are not correlated with the motor commands of the robot. As described in the previous section, there is a drop off after 5 s, which is due to the fact that the robot executes a new motor command approximately every 5 s. Therefore, any false positive movements of the background markers that are detected after the 5 s interval will be associated with the next motor command.

In this case the average efferent-afferent delay was estimated at 1.03559 s.

6.4. Test case with two robots: mimicking movements

Under this test condition the robot on the right (robot 2) is mimicking the robot on the left (robot 1). The mimicking robot starts to move 20 frames (0.66 s) after the first robot. As in Section 6.2, the second robot was generated using a

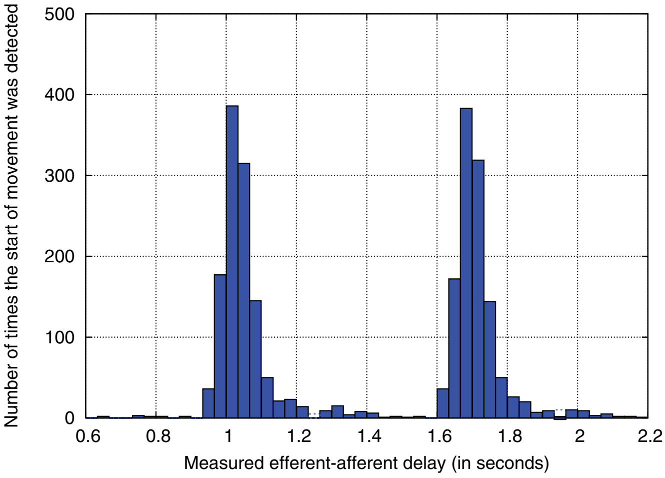


Fig. 11. Histogram for the measured delays between motor commands and observed visual movements in the mimicking test sequence with two robots (see Fig. 10). The left peak is produced by the movements of the body markers of the first robot. The right peak is produced by the movements of the body markers of the second/mimicking robot.

digital video special effect. Another data set of 500 motor commands was constructed using the frames of data set 1 (described in Section 6.1) and offsetting the left and right parts of the image by 20 frames.

Because the mimicking delay is always the same, the resulting histogram (see Fig. 11) is bimodal. The left peak, centered around 1 s, is produced by the body markers of the first robot. The right peak, centered around 1.7 s, is produced by the body markers of the second robot. Algorithm 3 cannot deal with situations like this and therefore it selects a delay that is between the two peaks (Mean = 1.36363 s, Stdev = 0.334185). Calculating the mean delay from the raw data produces an estimate that is between the two peak values as well (Mean = 1.44883 sec, Stdev = 0.52535).

It is possible to modify Algorithm 3 to avoid this problem by choosing the peak that corresponds to the shorter delay, for example. Evidence from animal studies, however, shows that when multiple time delays (associated with food rewards) are reinforced the animals learn "the mean of the reinforced distribution, not its lower limit," see ref. [13, p. 293], i.e., if the reinforced delays are generated from different underlying distributions the animals learn the mean associated with the mixture model of these distributions. Therefore, the algorithm was left unmodified.

Another reason to leave the algorithm intact exists: the mimicking test condition is a degenerate case that is highly unlikely to occur in any real situation, in which the two robots are independent. Therefore, this negative result should not undermine the usefulness of Algorithm 3 for learning the efferent-afferent delay. The probability that two independent robots will perform the same sequence of movements over *an extended period of time* is effectively zero. Continuous mimicking for extended periods of time is certainly a situation that humans and animals never encounter in the real world.

The results of the mimicking robot experiments suggest an interesting study that can be conducted with monkeys provided that a brain implant for detecting and interpreting the signals from the motor neurons of an infant monkey were available. The decoded signals could then be used to send movement commands to a robot arm, which would begin to move shortly after the monkey's arm. If there is indeed an imprinting period, as Watson[33] suggests, during which the efferent-afferent delay must be learned then the monkey should not be able to function properly after the imprinting occurs and the implant is removed.

7. Experimental Results: "Self" versus "Other" Discrimination

The basic methodology for performing this discrimination was already shown in Fig. 2. In the concrete implementation, the visual field of view of the robot is first segmented into features and then their movements are detected using the method described in Section 5.3. For each feature the robot maintains two independent probabilistic estimates that jointly determine how likely it is for the feature to belong to the robot's own body.

The two probabilistic estimates are the *necessity index* and the *sufficiency index* as described by Watson.[32,33] Fig. 12 shows an example with three visual features and their calculated necessity and sufficiency indexes. The necessity index measures whether the feature moves consistently after every motor command. The sufficiency index measures whether for every movement of the feature there is a corresponding motor command that precedes it. In other words:

$$\begin{aligned}&\text{Necessity index}\\&\quad=\frac{\text{Number of temporally contingent movements}}{\text{Number of motor commands}},\\&\text{Sufficiency index}\\&\quad=\frac{\text{Number of temporally contingent movements}}{\text{Number of observed movements for this feature}}.\end{aligned}$$

For each feature, f_i, the robot maintains a necessity index, N_i, and a sufficiency index, S_i. The values of these indexes at time t are given by $N_i(t)$ and $S_i(t)$. Following Fig. 12, the values of these indexes can be calculated by maintaining three counters: $C_i(t)$, $M_i(t)$, and $T_i(t)$. Their definitions are as follows: $C_i(t)$ represents the number of motor commands executed by the robot from some start time t_0 up to the current time t. $M_i(t)$ is the number of observed movements for feature f_i from time t_0 to time t; and $T_i(t)$ is the number of temporally contingent movements observed for feature f_i up to time t. The first two counters are trivial to calculate. The third counter, $T_i(t)$, is incremented every time the feature f_i is detected to move (i.e., when $M_i(t) = 1$ and $M_i(t-1) = 0$) and the movement delay relative to the last motor command is approximately equal to the mean efferent-afferent delay plus or minus some tolerance interval. In other words,

$$T_i(t) = \begin{cases} T_i(t-1) + 1 : \text{if } \ M_i(t) = 1 \ \textbf{ and} \\ \qquad\qquad\qquad\quad M_i(t-1) = 0 \ \textbf{ and } \ \left|\dfrac{\mu - d_i}{\mu}\right| < \beta, \\ T_i(t-1) \qquad\quad : \text{otherwise,} \end{cases}$$

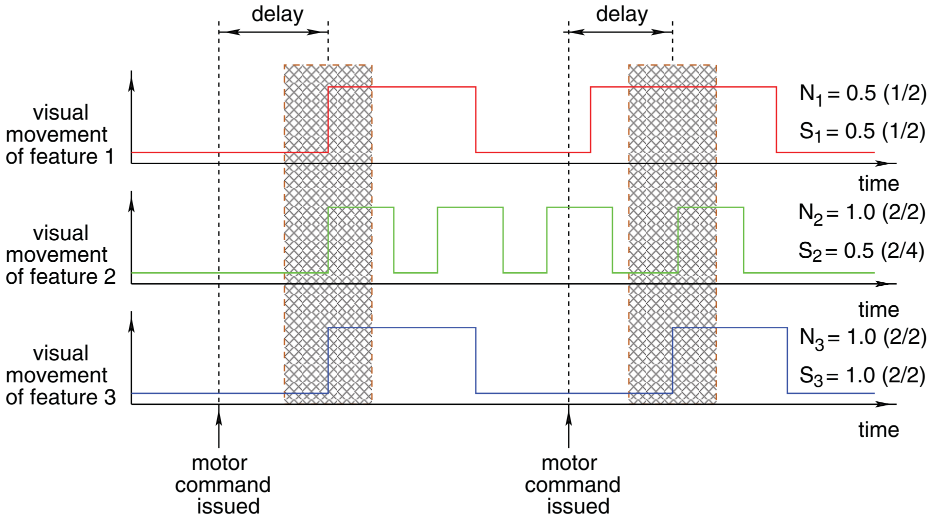


Fig. 12. The figure shows the calculated values of the necessity (N_i) and sufficiency (S_i) indexes for three visual features. After two motor commands, feature 1 is observed to move twice but only one of these movements is contingent upon the robot's motor commands. Thus, feature 1 has a necessity $N_1 = 0.5$ and a sufficiency index $S_1 = 0.5$. The movements of feature 2 are contingent upon both motor commands (thus $N_2 = 1.0$) but only two out of four movements are temporally contingent (thus $S_2 = 0.5$). Finally, feature 3 has both N_3 and S_3 equal to 1.0 as all of its movements are contingent upon the robot's motor commands.

where μ is the estimate for the mean efferent-afferent delay; d_i is the delay between the currently detected movement of feature f_i and the last motor command; and β is a constant. The value of β is independent from both μ and d_i and is equal to Weber's fraction (see Section 6). The inequality in this formula essentially defines the width of the temporal contingency regions (see the brown regions in Fig. 12).

The necessity and sufficiency indexes at time t can be calculated as follows:

$$N_i(t) = \frac{T_i(t)}{C_i(t)},$$

$$S_i(t) = \frac{T_i(t)}{M_i(t)}.$$

Both of these indexes are updated over time as new evidence becomes available, i.e., after a new motor command is issued or after the feature is observed to move. The belief of the robot that f_i is part of its body at time t is given jointly by $N_i(t)$ and $S_i(t)$. If the robot has classify feature f_i it can threshold these values; if both are greater than the threshold value, α, the feature f_i is classified as "self." In other words,

$$f_i \in \begin{cases} F_{\text{self}} & : \text{ if and only if } N_i(t) > \alpha \text{ and } S_i(t) > \alpha, \\ F_{\text{other}} & : \text{ otherwise.} \end{cases}$$

Ideally, both $N_i(t)$ and $S_i(t)$ should be 1. In practice, however, this is rarely the case as there is always some sensory noise that cannot be filtered out. Therefore, for all robot experiments the threshold value, α, was set to 0.75, which was empirically derived.[†]

The subsections that follow test this approach for "self" versus "other" discrimination in a number of experimental situations. In this set of experiments, however, it is assumed that the robot has already estimated its own efferent-afferent delay and is only required to classify the features as either "self" or "other" using this delay.

These test conditions are the same as the ones described in the previous section. For all experiments that follow, the value of the mean efferent-afferent delay was set to 1.035 and the value of β was set to 0.25. Thus, a visual movement will be classified as temporally contingent to the last motor command if the measured delay is between 0.776 and 1.294 s.

7.1. Test case with a single robot

The test condition here is the same as the one described in Section 6.1 and uses the same two data sets with 500 motor babbling commands in each. In this case, however, the robot already has an estimate for its efferent-afferent delay and is only required to classify the markers as either "self" or "other." Because the two data sets don't contain any background markers, the robot should classify all markers as "self." The experiments show that this was indeed the case.

Figure 13 shows the value of the sufficiency index calculated over time for each of the six body markers in data set 1 (the results are similar for data set 2). As mentioned above, these values can never be equal to 1.0 for a long period of time due to sensory noise. In this case, the sufficiency indexes for all six markers are greater than 0.75 (which is the value of the threshold α).

An interesting observation about this plot is that after the initial adaptation period (approximately 5 min) the values for the indexes stabilize and do not change much. This suggests that these indexes can be calculated over a running window instead of over the entire data set with very similar results.

[†] It is worth mentioning that $N_i(t)$ is the maximum likelihood estimate of Pr(feature i moves | motor command executed) and also that $S_i(t)$ is the maximum likelihood estimate of Pr(motor command executed | feature i moves). The comparison of the two indexes with a constant ensures that the strength of the causal connection in both directions meets a certain minimum threshold.

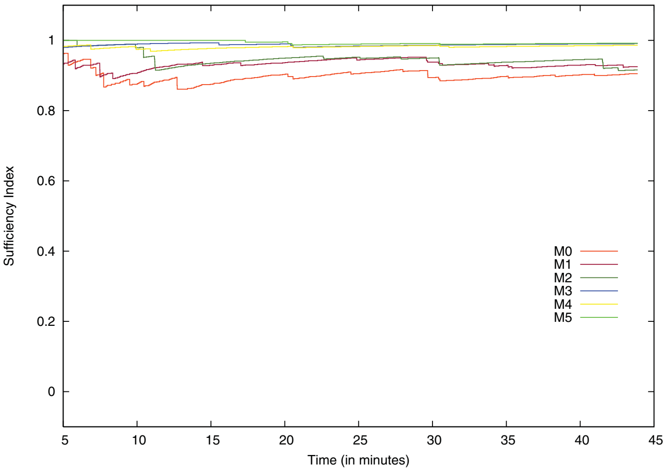


Fig. 13. The figure shows the value of the sufficiency index calculated over time for the six body markers. The index value for all six markers is above the threshold $\alpha = 0.75$. The values were calculated using data set 1.

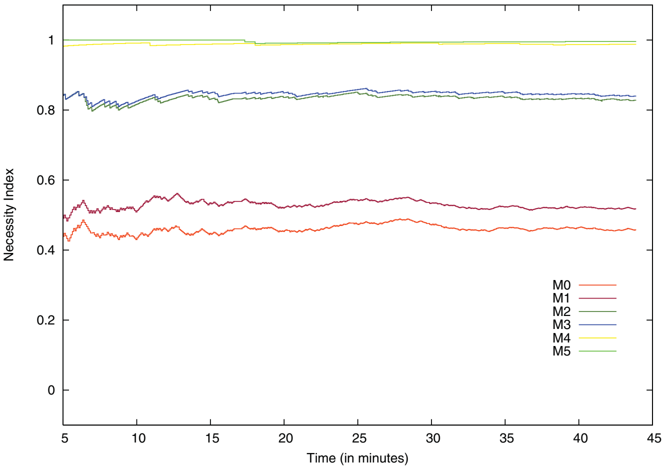


Fig. 14. The value of the necessity index calculated over time for each of the six body markers in data set 1. This calculation does not differentiate between the type of motor command that was performed. Therefore, not all markers can be classified as "self" as their index values are less than the threshold $\alpha = 0.75$ (e.g., M0 and M1). The solution to this problem is shown in Fig. 15 (see text for more details).

The oscillations in the first 5 min of each trial (not shown) are due to the fact that all counters and index values initially start from zero. Also, when the values of the counters are relatively small (e.g., 1–10) a single noisy update for any counter results in large changes for the value of the fraction that is used to calculate a specific index (e.g., the difference between 1/2 and 1/3 is large but the difference between 1/49 and 1/50 is not).

Figure 14 shows the value of the necessity index calculated over time for each of the six markers in data set 1 (the results are similar for data set 2). The figure shows that the necessity indexes are consistently above the 0.75 threshold only for body markers 4 and 5 (yellow and green). At first this may seem surprising; after all, the six markers are part of the robot's body and, therefore, should have similar values for their necessity indexes. The reason for this result is that the robot has three different joints which can be affected by the motor babbling routine (see Algorithm 1). Each motor command moves one of the three joints independently of the other joints. Furthermore, one or more of these motor commands can be executed concurrently.

Thus, the robot has a total of seven different types of motor commands. Using binary notation these commands can be labeled as: 001, 010, 011, 100, 101, 110, and 111. In this notation, 001 corresponds to a motor command that moves only the wrist joint; 010 moves only the elbow joint; and 111 moves all three joints at the same time. Note that 000 is not a valid command since it does not move any of the joints. Because markers 4 and 5 are located on the wrist they move for every motor command. Markers 0 and 1, however, are located on the shoulder and thus they can be observed to move only for four out of seven motor commands: 100, 101, 110, and 111. Markers 2 and 3 can be observed to move for 6 out of 7 motor commands (all except 001), i.e., they will have a necessity index close to of 6/7 which is approximately 0.85 (see Fig. 14).

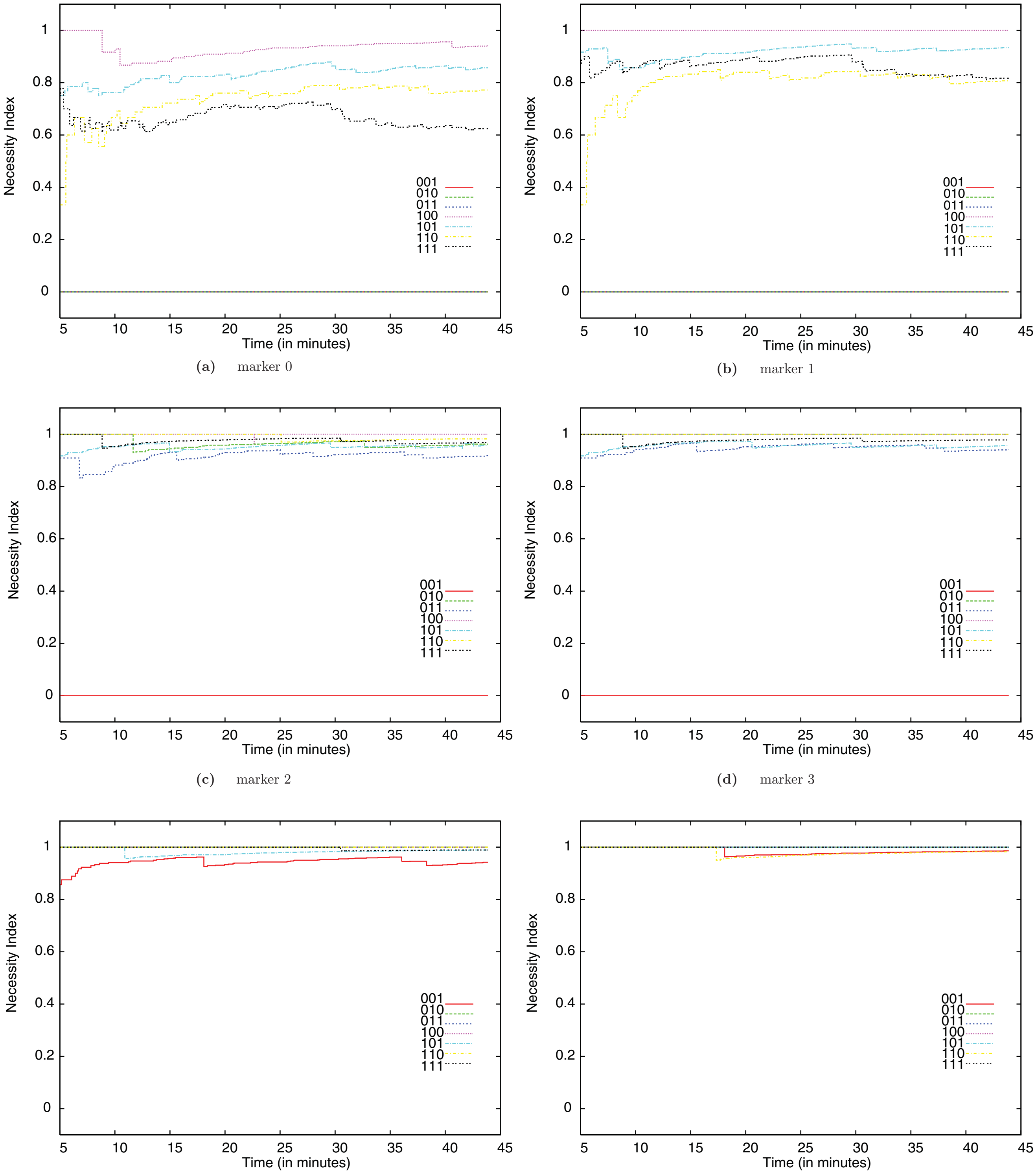
This example shows that the probability of necessity may not always be computed correctly as there may be several competing causes. In fact, this observation is well supported fact in the statistical inference literature, see ref. [22, p. 285]. "Necessity causation is a concept tailored to a specific event under consideration (singular causation), whereas sufficient causation is based on the general tendency of certain event *types* to produce other event types," see ref. [22, p. 285]. This distinction was not made by Watson[32,33] as he was only concerned with discrete motor actions (e.g., kicking or no kicking) and it was tacitly assumed that the infants always kick with both legs simultaneously.

While the probability of necessity may not be identifiable in the general case, it is possible to calculate it for each of the possible motor commands. To accommodate for the fact that the necessity indexes, $N_i(t)$, are conditioned upon the motor commands the notation is augmented with a superscript, m, which stands for one of the possible types of motor commands. Thus, $N_i^m(t)$ is the necessity index associated with feature f_i and calculated only for the m^{th} motor command at time t. The values of the necessity index for each feature f_i can now be calculated for each of the m possible motor commands as $N_i^m(t) = \frac{T_i^m(t)}{C_i^m(t)}$, where $C_i^m(t)$ is the total number of motor commands of type m performed up to time t; and $T_i^m(t)$ is the number of movements for feature f_i that are temporally contingent to motor commands of type m. The calculation for the sufficiency indexes remains the same as before.

Using this notation, a marker can be classified as "self" at time t if the value of its sufficiency index $S_i(t)$ is greater than α and there exists at least one type of motor command, m, such that $N_i^m(t) > \alpha$. In other words,

$$f_i \in \begin{cases} F_{\text{self}} & : \text{if and only if } \exists m : N_i^m(t) > \alpha \text{ and } S_i(t) > \alpha, \\ F_{\text{other}} & : \text{otherwise.} \end{cases}$$

Figure 15 shows the values of the necessity index for each of the six body markers calculated over time using data set 1 and the new notation. Each graph in this figure shows



(a) marker 0

(b) marker 1

(c) marker 2

(d) marker 3

(e) marker 4

(f) marker 5

Fig. 15. The figures shows the values of the necessity index, $N_i^m(t)$, for each of the six body markers (in data set 1). Each figure shows seven lines that correspond to one of the seven possible types of motor commands: 001, . . . , 111. To be considered for classification as "self" each marker must have a necessity index $N_i^m(t) > 0.75$ for at least one motor command, m, at the end of the trial. All markers are classified as "self" in this data set.

seven lines, which correspond to one of the seven possible motor commands. As can be seen from the figure, for each marker there is at least one motor command, m, for which the necessity index $N_i^m(t)$ is greater than the threshold, $\alpha = 0.75$. Thus, all six markers are correctly classified as "self." The results are similar for data set 2.

It is worth noting that the approach described here relies only on identifying which joints participate in any given motor command and which markers are observed to start moving shortly after this motor command. The type of robot movement (e.g., fast, slow, fixed speed, variable speed) and how long a marker is moving as a result of it does not

Table I. Values of the necessity and sufficiency indexes at the end of the trial. All markers are classified correctly as “self” or “other”.

Marker	$\max_m(N_i^m(t))$	$S_i(t)$	Threshold α	Classification	Actual
M0	0.941	0.905	0.75	“self”	“self”
M1	1.000	0.925	0.75	“self”	“self”
M2	1.000	0.912	0.75	“self”	“self”
M3	1.000	0.995	0.75	“self”	“self”
M4	1.000	0.988	0.75	“self”	“self”
M5	1.000	0.994	0.75	“self”	“self”
M6	0.066	0.102	0.75	“other”	“other”
M7	0.094	0.100	0.75	“other”	“other”
M8	0.158	0.110	0.75	“other”	“other”
M9	0.151	0.107	0.75	“other”	“other”
M10	0.189	0.119	0.75	“other”	“other”
M11	0.226	0.124	0.75	“other”	“other”

affect the results produced by this approach. The following subsections test this approach under different experimental conditions.

7.2. *Test case with two robots: Uncorrelated movements*

This experimental condition is the same as the one described in Section 6.2. The data set recorded for the purposes of Section 6.2 was used here as well. If the self-detection algorithm works as expected only 6 of the 12 markers should be classified as “self” (markers M0–M5). The other six markers (M6–M11) should be classified as “other.” Table I shows that this is indeed the case.

Figure 16 shows the sufficiency indexes for the six body markers of the first robot (i.e., the one trying to perform the self versus other discrimination—left robot in Fig. 7). As expected, the index values are very close to 1. Figure 17 shows the sufficiency indexes for the body markers of the second robot. Since the movements of the second robot are not correlated with the motor commands of the first robot these values are close to zero.

The necessity indexes for each of the 6 body markers of the first robot for each of the seven motor commands are very similar to the plots shown in the previous subsection. As expected, these indexes (not shown) are greater than 0.75 for at least one motor command. Figure 18 shows the necessity indexes for the markers of the second robot. In this case, the necessity indexes are close to zero. Thus, these markers are correctly classified as “other.”

7.3. *Test case with a single robot and static background features*

This test condition is the same as the one described in Section 6.3. In addition to the robot’s body markers, six additional markers were placed on the background wall (see Fig. 9). Again, the robot performed motor babbling for 500 motor commands. The data set recorded for the purposes of Section 6.3 was used here as well.

Table II shows the classification results at the end of the test. The results demonstrate that there is a clear distinction between the two sets of markers: markers M0–M5

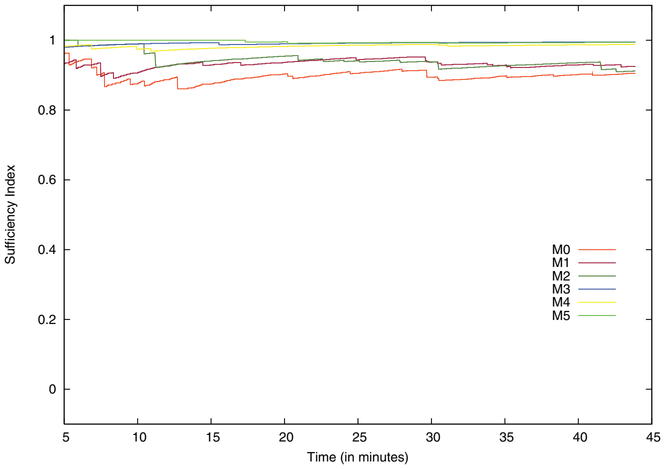


Fig. 16. The figure shows the sufficiency indexes for each of the six body markers of the first robot (left robot in Fig. 7). As expected, these values are close to 1, and thus, above the threshold $\alpha = 0.75$. The same is true for the necessity indexes (not shown). Thus, all markers of the first robot are classified as “self.”

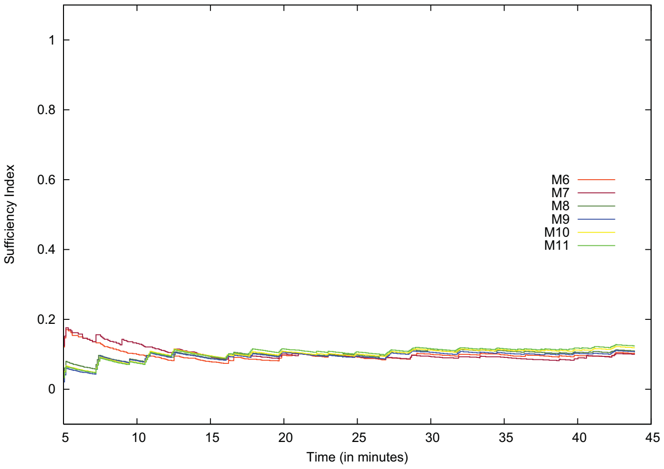


Fig. 17. The figure shows the sufficiency indexes for each of the six body markers of the second robot (right robot in Fig. 7). As expected, these values are close to 0, and thus, below the threshold $\alpha = 0.75$. The same is true for the necessity indexes as shown in Fig. 18. Thus, the markers of the second robot are classified as “other.”

Table II. Values of the necessity and sufficiency indexes at the end of the trial. The classification for each marker is shown in the last column.

Marker	$\max_m(N_i^m(t))$	$S_i(t)$	Threshold α	Classification	Actual
M0	0.977	0.887	0.75	"self"	"self"
M1	1.000	0.943	0.75	"self"	"self"
M2	1.000	0.998	0.75	"self"	"self"
M3	1.000	0.995	0.75	"self"	"self"
M4	1.000	0.840	0.75	"self"	"self"
M5	1.000	0.996	0.75	"self"	"self"
M6	0.000	0.000	0.75	"other"	"other"
M7	0.068	0.140	0.75	"other"	"other"
M8	0.017	0.100	0.75	"other"	"other"
M9	0.057	0.184	0.75	"other"	"other"
M10	0.147	0.112	0.75	"other"	"other"
M11	0.185	0.126	0.75	"other"	"other"

are classified correctly as "self." All background markers, M6–M11, are classified correctly as "other." The background markers are labeled clockwise starting from the upper left marker (red) in Fig. 9. Their colors are: red (M6), violet (M7), pink (M8), tan (M9), orange (M10), light blue (M11).

All background markers (except marker 8) can be temporarily occluded by the robot's arm, which increases their position tracking noise. This results in the detection of occasional false positive movements for these markers. Therefore, their necessity indexes are not necessarily equal to zero. Nevertheless, by the end of the trial the maximum necessity index for all background markers is well below 0.75 and, thus, they are correctly classified as "other." Due to space limitations the necessity and sufficiency plots are not shown here. They are given in ref. [28].

7.4. *Test case with two robots: Mimicking movements*

This test condition is the same as the one described in Section 6.4. The mean efferent-afferent delay for this experiment was also set to 1.035 s. Note that this value is different from the wrong value (1.36363 s) estimated for this degenerate case in Section 6.4.

Table III shows the values for the necessity and sufficiency indexes at the end of the 45 min interval. As expected, the sufficiency indexes for all body markers of the first robot are close to 1. Similarly, the necessity indexes are close to 1 for at least one motor command. For the body markers of the second robot the situation is just the opposite. Due to space limitations the necessity and sufficiency plots are not shown here, but they are given in ref. [28].

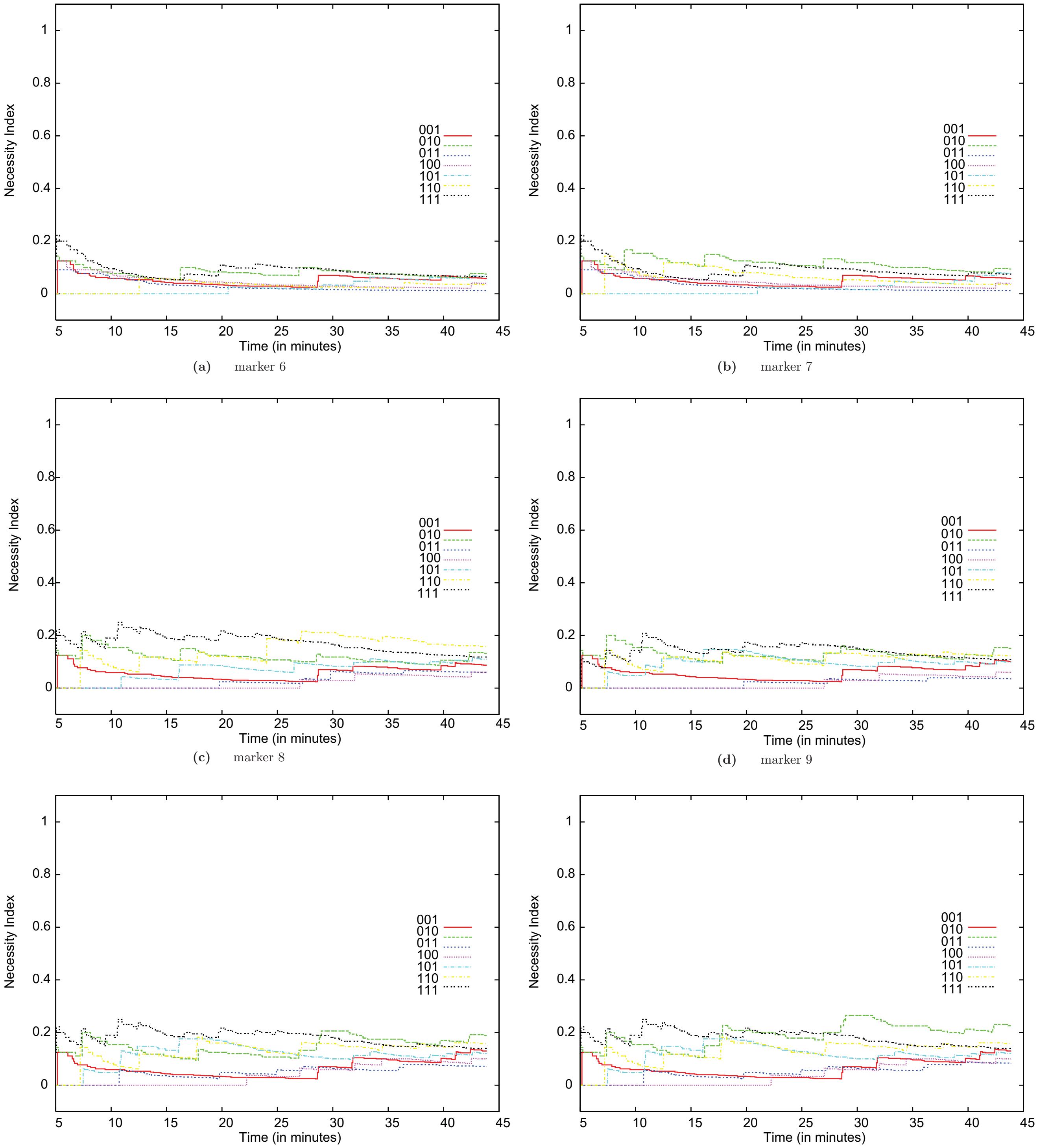
Somewhat surprisingly, the mimicking test condition turned out to be the easiest one to classify. Because the second robot always starts to move a fixed interval of time after the first robot, almost no temporally contingent movements are detected for its body markers. Thus, both the necessity and sufficiency indexes for most markers of the second robot are equal to zero. Marker 8 is an exception because it is the counterpart to marker 2 which has the noisiest position detection.

8. Self-Detection in a TV monitor

The experiment described in this section adds a TV monitor to the existing setup as shown in Fig. 19. The TV image displays the movements of the robot in real time as they are captured by a second camera that is different from the robot's camera. This experiment was inspired by similar setups used by Watson[33] in his self-detection experiments with infants.

Table III. Values of the necessity and sufficiency indexes at the end of the trial. All markers are classified correctly as "self" or "other" in this case.

Marker	$\max_m(N_i^m(t))$	$S_i(t)$	Threshold α	Classification	Actual
M0	0.941	0.905	0.75	"self"	"self"
M1	1.000	0.925	0.75	"self"	"self"
M2	1.000	0.918	0.75	"self"	"self"
M3	1.000	0.995	0.75	"self"	"self"
M4	1.000	0.988	0.75	"self"	"self"
M5	1.000	0.994	0.75	"self"	"self"
M6	0.000	0.000	0.75	"other"	"other"
M7	0.022	0.007	0.75	"other"	"other"
M8	0.059	0.011	0.75	"other"	"other"
M9	0.000	0.000	0.75	"other"	"other"
M10	0.000	0.000	0.75	"other"	"other"
M11	0.000	0.000	0.75	"other"	"other"



(a) marker 6

(b) marker 7

(c) marker 8

(d) marker 9

(e) marker 10

(f) marker 11

Fig. 18. The necessity index, $N_i^m(t)$, for each of the six body markers of the second robot. Each figure shows seven lines that correspond to one of the seven possible types of motor commands: 001, . . . , 111. To be considered for classification as "self," each marker must have a necessity index $N_i^m(t) > 0.75$ for at least one motor command, m, at the end of the trial. This is not true for the body markers of the second robot shown in this figure. Thus, they are correctly classified as "other" in this case.

The experiment tests whether a robot can use its estimated efferent-afferent delay to detect that an image shown in a TV monitor is an image of its own body.

A new data set with 500 movement commands was gathered for this experiment. Similarly to previous experiments, the robot was under the control of the motor babbling procedure. The data set was analyzed in the same way as described in the previous sections. The only difference was that the position detection for the TV markers was slightly more noisy than in previous data sets. Therefore, the raw marker position data was averaged over three consecutive frames (the smallest number required for proper averaging). Also, detected marker movements shorter than six frames in duration were ignored.

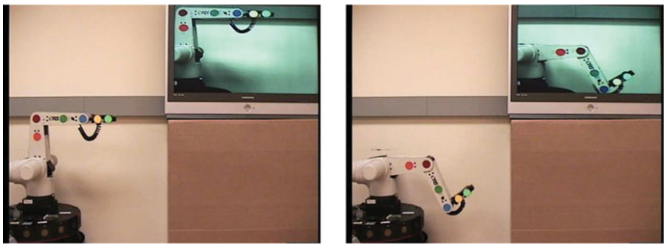


Fig. 19. Frames from the TV sequence. The TV image shows in real time the movements of the robot captured from a video camera that is different from the robot's camera.

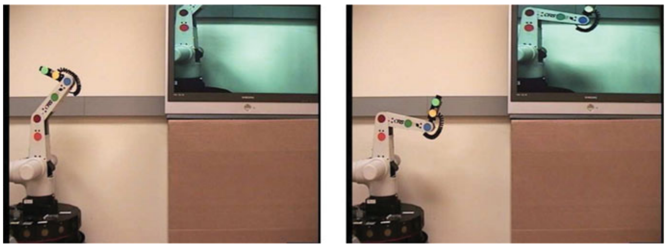


Fig. 20. Frames from the TV sequence in which some body markers are not visible in the TV image due to the limited size of the TV screen.

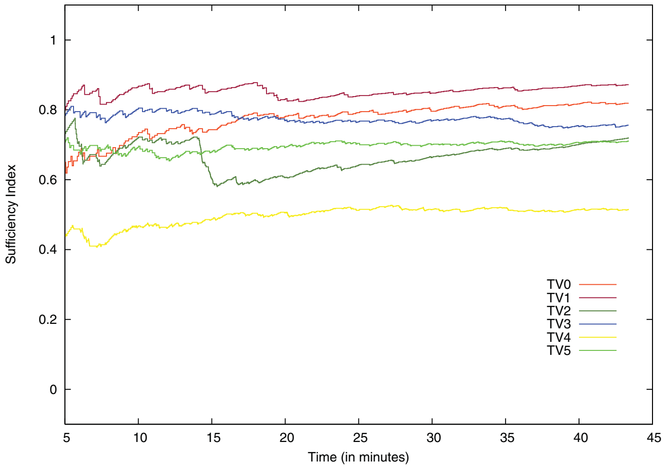


Fig. 21. The sufficiency indexes calculated over time for the six TV markers. These results are calculated *before* taking the visibility of the markers into account.

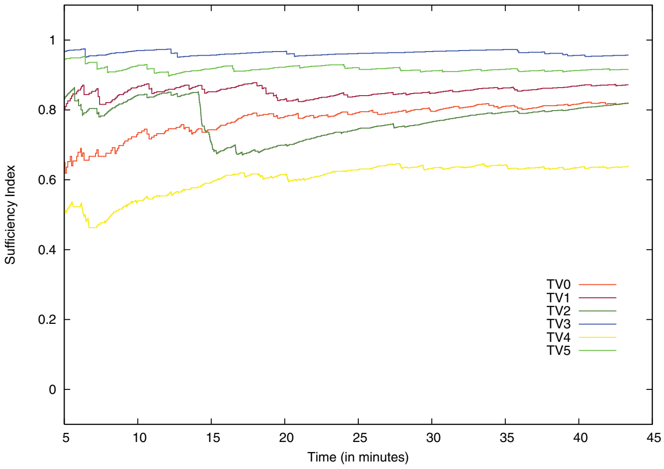


Fig. 22. The sufficiency indexes calculated over time for the six TV markers. These results are calculated *after* taking the visibility of the markers into account.

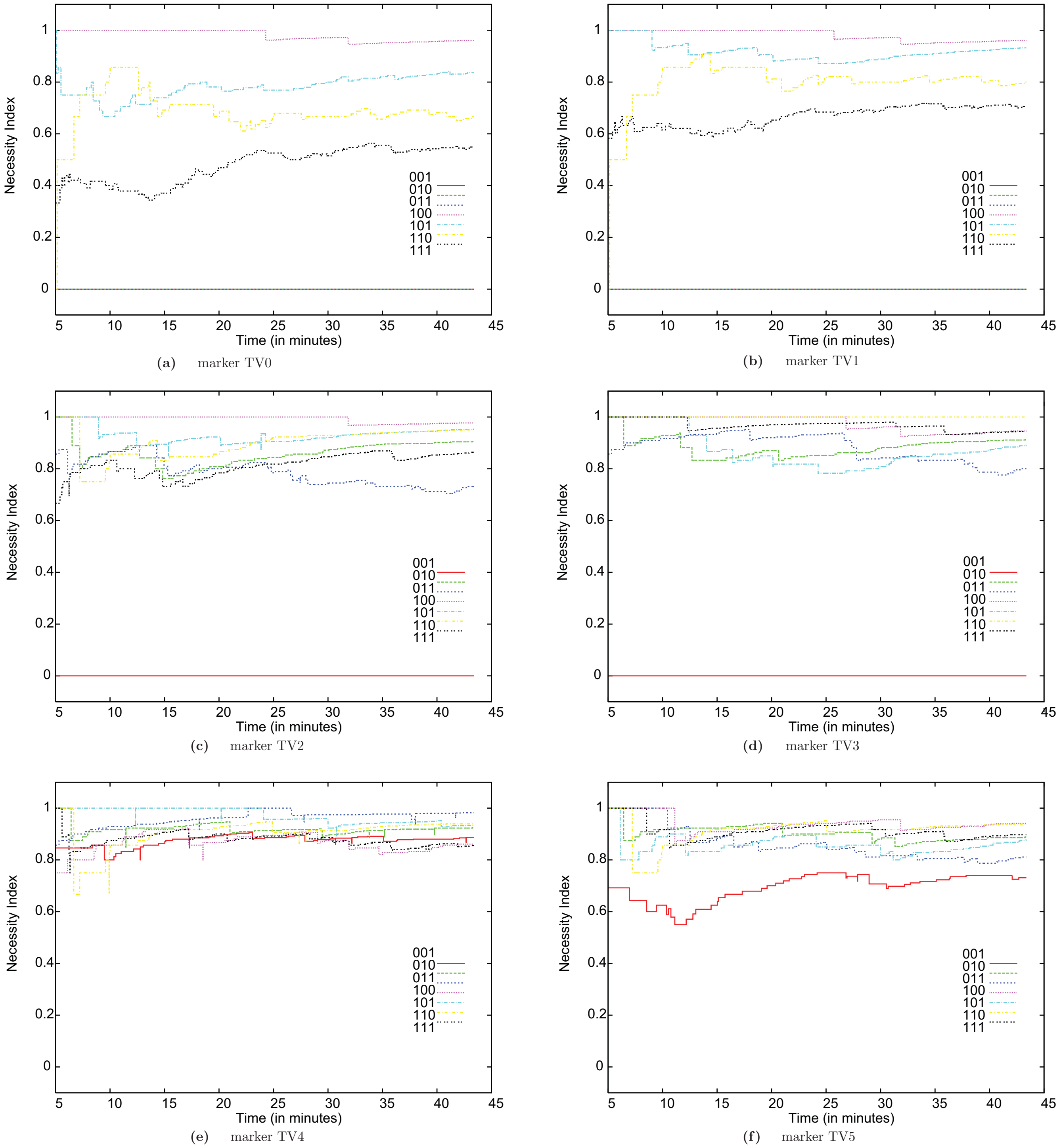
The results for the sufficiency and necessity indexes for the robot's six body markers are similar to those described in the previous sections and thus will not be discussed any further. This section will only describe the results for the images of the six body markers in the TV monitor, which will be refereed to as TV markers (or TV0, TV1, . . . , TV5).

Figure 21 shows the sufficiency indexes calculated for the six TV markers. Somewhat surprisingly, the sufficiency indexes for half of the markers do not exceed the threshold value of 0.75 even though these markers belong to the robot's body and they are projected in real time on the TV monitor. The reason for this, however, is simple and it has to do with the size of the TV image. Unlike the real body markers, which can be seen by the robot's camera for all body poses, the projections of the body markers in the TV image can only be seen when the robot is in specific body poses. For some body poses the robot's arm is either too high or too low and thus the markers cannot be observed in the TV monitor. Figure 20 shows several frames from the TV sequence to demonstrate this more clearly. The actual visibility values for the six TV markers are as follows: 99.9% for TV0, 99.9% for TV1, 86.6% for TV2, 72.1% for TV3, 68.5% for TV4, and 61.7% for TV5. In contrast, the robot's markers (M0–M5) are visible 99.9% of the time.

This result prompted a modification of the formulas for calculating the necessity and sufficiency indexes. In addition to taking into account the specific motor command, the self-detection algorithm must also take into account the visibility of the markers. In all previous test cases, all body markers were visible for all body configurations (subject to the occasional transient sensory noise). Because of that, visibility was never considered even though it was implicitly included in the detection of marker movements. For more complicated robots (e.g., humanoids) the visibility of the markers should be taken into account as well. These robots have many body poses for which they may not be able to see some of their body parts (e.g., hand behind the back).

To address the visibility issue, the following changes were made to the way the necessity and sufficiency indexes are calculated. The robot checks the visibility of each marker for all frames in the time interval immediately following a motor command. Let the kth motor command be issued at time T_k and the (k+1)-st command be issued at time T_{k+1}. Let $\hat{T}_k \in [T_k, T_{k+1})$ be the time at which the kth motor command is no longer considered contingent upon any visual movements. In other words, $\hat{T}_k = T_k + \mu + \beta\mu$, where μ is the average efferent-afferent delay and $\beta\mu$ is the estimate for the standard deviation calculated using Weber's law (see Section 6). If the ith marker was visible during less than 80% of the frames in the interval $[T_k, \hat{T}_k)$, then the movements of this marker (if any) are ignored for the time interval $[T_k, T_{k+1})$ between the two motor commands. In other words, none of the three counters ($T_i(t)$, $C_i(t)$, and $M_i(t)$) associated with this marker and used to calculate its necessity and sufficiency indexes are updated until the next motor command.

Figure 22 shows the sufficiency indexes for the six TV markers after correcting for visibility. Now their values are



(a) marker TV0

(b) marker TV1

(c) marker TV2

(d) marker TV3

(e) marker TV4

(f) marker TV5

Fig. 23. Values of the necessity index, $N_i^m(t)$, for each of the six TV markers. Each figure shows seven lines that correspond to one of the seven possible types of motor commands: 001, . . . , 111. To be considered for classification as "self" each marker must have at the end of the trial a necessity index, $N_i^m(t) > 0.75$ for at least one motor command, m. These graphs are calculated *after* taking the visibility of the TV markers into account.

all above the 0.75 threshold. The only exception is the yellow marker (TV4) which has a sufficiency index of 0.64 even after correcting for visibility. The reason for this is the distortion of the marker's color, which appears very similar to the background wall in the TV image. As a result, its position tracking is noisier than before.

Figure 23 shows the necessity indexes calculated for each type of motor command for each TV marker (after taking the visibility of the markers into account). As can be seen from Fig. 22 and Fig. 23, five out of six TV markers are correctly classified as "self" because at the end of the trial they all have a sufficiency index greater than 0.75 and a necessity index greater than 0.75 for at least one motor command. The only marker that was not classified as "self" was the yellow marker for reasons explained above.

The results of this section demonstrate, to the best of my knowledge, the first-ever experiment of self-detection in a TV monitor by a robot. Furthermore, it is possible to build upon these results to achieve video-guided robot behaviors (another first). In other words, it is possible for a robot to detect its own image in a TV monitor and use that image to guide its own reaching movements in order to grasp an object that can only be seen in the TV image. See ref. [28] and [29] for more details.

9. Summary and Conclusions

This paper described a methodology for autonomous self-detection by a robot. The methodology is based on the detection of the temporal contingency between motor commands (efferent signals) and visual movements (afferent signals) to estimate the efferent-afferent delay of the robot. It was shown how the robot could estimate its own efferent-afferent delay from self-observation data gathered while the robot performs motor babbling, i.e., random joint movements similar to the primary circular reactions described by Piaget.[23] The results demonstrate that the self-detection algorithm performs well for the experimental conditions described in this paper.

This paper also introduced a method for feature-level self-detection based on the ideas proposed by Watson.[33] The method maintains a probabilistic estimate across all features as to whether or not they belong to the robot's body. The probabilities are estimated based on estimates of necessity and sufficiency. The sufficiency index measures the probability of the stimulus (visual movement) occurring some specified amount of time after the behavior (motor command). The necessity index estimates the probability that the a behavior (motor command) was performed in a temporal window before the stimulus (visual movement) was observed. By using these two indexes the robot can overcome the problems associated with the detection of false positives and false negatives, which are bound to occur due to lucky coincidences.

The experimental results show that a robot can successfully distinguish between its own body and the external environment. The robot was able to correctly classify different visual stimuli as either "self" or "other." This was possible even when there were other moving objects in the environment because the movements of environmental features (including other robots) were not perfectly correlated with the motor commands of the robot. Also, the method proposed here was successfully used by the robot to detect its self-image in a TV monitor, which is an original contribution of this research.

The results show that Watson's ideas are suitable for robotic applications. There are some implementation details, however, that Watson did not foresee (or maybe they were not applicable to his experimental setups with infants). For example, the size of the TV image imposes a restriction on which body markers can be seen and for which body poses. Previous studies with infants (e.g., ref. [2, 33]) have tacitly assumed that the required features are visible at all times. Without correcting for visibility the values of the necessity and sufficiency indexes can exhibit at most medium levels of contingency. Another factor that is not mentioned by Watson is that the self-detection algorithm must take into account the types of motor commands that are issued as not all body markers are moved by a given motor command. Without this correction the necessity indexes cannot reach the near perfect values required for successful self-detection. Both of these modifications were implemented and tested successfully on the robot.

One limitation of the current implementation is that it makes the assumption that visually distinct features exist on the surface of the robot's body and that these features can be identified and tracked reliably. The color markers were chosen in order to solve the tracking and correspondence problems in a computationally efficient way. Future work should focus on either eliminating the need for distinct perceptual features or adding the ability to learn these features from actual experience.

Future work can also focus on extending the computational methods and ideas presented here to other sensory domains. For example, it should be possible to couple the robot's motor commands with auditory data in order to implement auditory self-detection. In other words, the robot should be able to detect its own mechanical noises. Furthermore, it should be possible to classify auditory events as either caused by the robot or produced by the environment. Applications to other sensory modalities such as touch should also be straightforward.

References

1. M. Aguilar and W. Stiles, "Saturation of the rod mechanism of the retina at high levels of stimulation," *Optica Acta.* **1**, 59–65 (1954).
2. L. Bahrick and J. Watson, "Detection of intermodal proprioceptive-visual contingency as a basis of self perception in infancy," *Dev. Psychol.* **21**, 963–973 (1985).
3. J. Barth, D. Povinelli and J. Cant, "Bodily Origins of Self," **In**: *The Self and Memory* (D. Beike, J. Lampinen and D. Behrend, eds.) (Psychology Press, New York, 2004), pp. 11–43.
4. A. Catania, "Reinforcement Schedules and Psychophysical Judgments: A Study of Some Temporal Sroperties of Behavior," **In**: *The Theory of Reinforcement Schedules* (W. Schoenfeld, ed.) (Appleton-Century-Crofts, New York, 1970), pp. 1–42.
5. F. de Waal, *Bonobo: The Forgotten Ape* (University of California Press, Berkeley, 1997).
6. F. de Waal, M. Dindo, C. Freeman and M. Hall, "The monkey in the mirror: Hardly a stranger," *Proc. Nation. Acad. Sci. (PNAS)* **102**(32), 11140–11147 (2005).
7. R. Flom and L. Bahrick, "Is featural information necessary for 5-month-olds' perception of motion specifying the self?," *Society for Research in Child Development*, Minneapolis, MN (Apr. 2001).
8. C. Gallistel, "Time has come," *Neuron* **38**(2), 149–150 (2003).
9. R. Gallistel and J. Gibbon, "Time, rate and conditioning," *Psychol. Rev.* **107**, 289–344 (2000).
10. G. Gallup, "Chimpanzees: Self-recognition," *Science* **167** (3914), 86–87, (1970).
11. G. Gallup, "Self recognition in primates: A comparative approach to the bidirectional properties of consciousness," *Am. Psychol.* **32**, 329–338 (1977).
12. G. Gallup, J. Anderson and D. Shillito, "The Mirror Test," **In**: *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition* (M. Bekoff, C. Allen and G. Burghardt, eds.) (MIT Press, Cambridge, MA, 2002).
13. J. Gibbon, "Scalar expectancy theory and Weber's law in animal timing," *Psychol. Rev.* **84**(3), 279–325 (1977).
14. J. Gibbon, "Origins of scalar timing," *Learn. Motivation* **22**(1–2), 3–38 (1991).

15. J. Gibbon and R. Church, "Sources of Variance in an Information Processing Theory of Timing," **In**: *Animal Cognition* (H. Roitblat, T. Bever and H. Terrace, eds.) (Lawrence Erlbaum Associates, Hillsdale, NJ, 1984), pp. 465–488.
16. J. Gibbon, C. Malapani, C. Dale and C. Gallistel, "Toward a neurobiology of temporal cognition: Advances and challenges," *Curr. Opin. Neurobiol.* **7**(2), 170–184 (1997).
17. K. Gold and B. Scassellati, "Using probabilistic reasoning over time to self-recognize," *Robot. Auton. Syst.* **57**(4), 384–392 (2009).
18. G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception Psychophys.* **14**, 201–11 (1973).
19. M. Lewis and J. Brooks-Gunn, *Social Cognition and the Acquisition of Self* (Plenum Press, New York, 1979).
20. P. Michel, K. Gold and B. Scassellati, "Motion-Based Robotic Self-Recognition," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sendai, Japan (2004).
21. S. Parker, R. Mitchell and M. Boccia (eds.), *Self-Awareness in Animals and Humans: Developmental Perspectives* (Cambridge University Press, Cambridge, 1994).
22. J. Pearl, *Causality: Models, Reasoning and Inference* (Cambridge University Press, New York, 2000).
23. J. Piaget, *The Origins of Intelligence in Children* (International Universities Press, New York, 1952).
24. J. Plotnik, F. de Waal and D. Reiss, "Self-recognition in an Asian Elephant," *Proc. Natl. Acad. Sci. USA (PNAS)* **103**(45), 17053–17057 (2006).
25. D. Povinelli and J. Cant, "Arboreal clambering and the evolution of self-conception," *Q. Rev. Biol.* **70**(4), 393–421 (1995).
26. D. Reiss and L. Marino, "Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence," *Proc. Nation. Acad. Sci. (PNAS)* **98**(10), 5937–5942 (2001).
27. P. Rochat, "Five levels of self-awareness as they unfold early in life," *Consciousness Cognit.* **12**, 717–731 (2003).
28. A. Stoytchev, Robot Tool Behavior: A Developmental Approach to Autonomous Tool Use, *PhD Thesis* (College of Computing, Georgia Institute of Technology, Atlanta, GA, Aug. 2007).
29. A. Stoytchev, "Toward Video-Guided Robot Behaviors," *Proceedings of the Seventh International Conference on Epigenetic Robotics (EpiRob)*, Rutgers University, NJ (Nov. 5–7, 2007), pp. 165–172.
30. M. Triesman, "Temporal discrimination and the indifference interval: Implications for a model of the "internal clock," *Psychol. Monogr.* **77** (13) (1963).
31. M. Triesman, "Noise and Weber's law: The discrimination of brightness and other dimensions," *Psychol. Rev.* **71**, 314–330 (1964).
32. J. Watson, "Contingency Perception in Early Social Development," **In**: *Social Perception in Infants* (T. Field and N. Fox, eds.) (Ablex Pub. Corp., Norwood, NJ, 1985), pp. 157–176.
33. J. Watson, "Detection of self: The perfect algorithm," **In**: *Self-Awareness in Animals and Humans: Developmental Perspectives* (S. Parker, R. Mitchell and M. Boccia, eds.) (Cambridge University Press, Cambridge, 1994), pp. 131–148.
34. Y. Yoshikawa, K. Hosoda and M. Asada, "Cross-Anchoring for Binding Tactile and Visual Sensations via Unique Association Through Self-Perception," *Proceedings of the Third International Conference on Development and Learning (ICDL)*, La Jolla, CA (Oct. 22–22, 2004).