# Learning of Object Identification by Natural Language Controlled Robots

Chandimal Jayawardena, Keigo Watanabe and Kiyotaka Izumi
Department of Advanced Systems Control Engineering,
Graduate School of Science and Engineering,
Saga University, Saga 840-8502, Japan.
chandimal@ieee.org, {watanabe, izumi}@me.saga-u.ac.jp

***Abstract*— Natural language communication is very important in Human-Robot cooperative work. This paper presents an object sorting robotic system which is controlled by natural language commands. A PA-10 robot manipulator is issued commands like *"pick the big red cube"* to pick objects placed on a table. The robot learns to interpret the meaning of this type of natural commands by learning individual lexical symbols in the grammar and their corresponding object features.**

**Keywords:** Object identification, Natural language commands, Object perception, Lexical symbols, Object features.

## I. Introduction

Human-robot interaction is one of the most important developments in the field of robotics. The effectiveness of a human-robot cooperative systems would be enhanced by improving the naturalness of the human-robot interface. In achieving this, the ability to communicate as peers using natural languages is of utmost importance [1][2][3].

On the other hand, object identification is one of the important features in intelligent robotic systems. Most research on vision based object identification systems have concentrated on identifying known objects in a scene (e.g. [4] [5]). In addition, there have been some research on learning and identifying unknown objects too (e.g. [6]).

In the experiment presented in this paper, a human user can command a robot manipulator verbally to pick objects placed on a table. The user can refer to objects naturally using references such as "*small red cube*." Learning to identify objects referred to in this manner is important for natural language understanding robots.

### A. Learning object identification

The object identification method employed in this work is different from the existing systems pointed out above. In this method, instead of learning an objects as it is, different object features and lexical symbols which represent those features in English language are learned. Then, that knowledge is applied to identify new objects which are characterized by combinations of learned features.

In natural languages object references are composed of combinations of lexical symbols representing shapes, colors, sizes, etc. In order to infer the meaning of such a combination, one should know the meaning of each lexical symbol. For example, to identify a "*large green car*," one should know what is meant by *large*, *green*, and *car*. In the human learning process, once the grounded meaning of a lexical symbol is learned, humans are capable of interpreting it with relation to different scenarios. This is true for childhood learning as well as for new language learning by adults. Our objective is to apply a similar strategy for learning object identification by robots.

Object perception by any robot is only via sensors. If the camera images are used it is possible to extract various features of the objects presented in a scene. This is a completely automated process where there is no consideration as to how these objects are represented in the domain of natural languages.

Although the robot perception is limited to sensory data, a human user may refer to objects with combinations of lexical symbols. "*red cube*", "*blue cylinder*", or "*big yellow sphere*" are some examples. In order to execute user commands which consist of such references, there should be a method to learn the meanings of these lexical symbols.

There have been many important work related to this problem [7] [8]. However, those work considered the problem as a fundamental cognitive problem. In this paper, learning the meaning of lexical symbols with the help of a human user is studied.

On the other hand, it is not limited to acquiring knowledge of some symbols; rather it uses independently learned lexical symbols to understand the meaning of a composite lexical item: i.e. a complete reference to an object. For example, after the meaning of the lexical symbol "*red*" is learned, it is meaningful for any red object; "*red cube*", "*red cylinder*", etc.

Here, we define two terms: relative features and non-relative features. An object feature whose meaning can be inferred without comparing with other objects is called a non-relative features. The object color is an example. In contrast, meaning of a relative feature can be inferred only after comparing with other objects. For example, the meanings of "*small red cube*" and "*big red cube*" are understood only by comparison between all "*red cubes*."

## II. Object Perception by a Robot

Let the number of objects presented in a scene be $N$. Assume that any object possesses $K$ number of non-relative

| | |
|---|---|
| 1 | Get the user command. |
| 2 | Extract non-relative lexical symbols, $l^{(1)}, l^{(2)}, ..., l^{(j)}, ..., l^{(K)}; l^{(j)} \in L^{(j)}$ |
| 3 | IF $l^{(1)}$ is known |
| | ⋮ |
| $2+K$ | IF $l^{(K)}$ is known<br>• Identify the object(s). |
| $3+K$ | ELSE<br>• Consult the user:<br>"*What do you mean by $l^{(K)}$?*".<br>• User points to any object with the feature described by $l^{(K)}$.<br>• Learn $l^{(K)}$: Map $l^{(K)}$ to the cluster to which the pointed object belongs.<br>• Go to step $2+K$. |
| | ⋮ |
| $2+2K$ | ELSE<br>• Consult the user:<br>"*What do you mean by $l^{(1)}$?*".<br>• User points to any object with the feature described by $l^{(1)}$.<br>• Learn $l^{(1)}$: Map $l^{(1)}$ to the cluster to which the pointed object belongs.<br>• Go to step 3. |

Fig. 1. Learning non-relative lexical symbols.

feature values which belong to $K$ number of mutually exclusive non-relative feature categories. For example, color may be a feature category. Depending on the application, vector of $RGB$ color components may be a feature value in the category color. Therefore, object $i$ can be represented with a vector $\boldsymbol{r}_i$.

$$\boldsymbol{r}_i = \left\{ \boldsymbol{f}_i^{(1)}, \boldsymbol{f}_i^{(2)}, ..., \boldsymbol{f}_i^{(j)}, ..., \boldsymbol{f}_i^{(K)} \right\} \tag{1}$$

where $\boldsymbol{f}_i^{(j)}$ is a feature value of the feature category $j$. It is a value obtained from raw sensory data.

$$R = \{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_N\} \tag{2}$$

is the set of all object representations.

Assuming that the number of distinguishable features within each non-relative feature category is finite, it should be possible to identify feature clusters within sensory data pertaining to any feature category; i.e. it should be possible to identify clusters among $\sum_{i=1}^{i=N} \boldsymbol{f}_i^{(j)}$ for the $j$th feature category. Let the number of clusters identified within sensory data pertaining to the $j$th feature category be $C^{(j)}$.

Set of objects that belong to any cluster in the $j$th non-relative feature category is given by:

$$a_{p_j}^{(j)} \subset R \tag{3}$$

for $p_j = 1, ..., C_j$. Here $a_{p_1}^{(j)} \cap a_{p_2}^{(j)} = 0$ where $p_1 \neq p_2$ for any $p_1$ and $p_2$.

Let the set of all $a_p^{(j)}$ be $A^{(j)}$.

All the objects in cluster $a_{p_j}^{(j)}$ have one common feature that belongs to the $j$th non-relative feature category. Let that feature be $b_{p_j}^{(j)}$.

TABLE I

GRAMMAR

| Action | Article | Size | Color | Shape |
|---|---|---|---|---|
| pick<br>grab<br>take | (the) | small<br>medium<br>big | red<br>green<br>blue | cube<br>cylinder |

Therefore, the sets of objects which have features $b_{p_1}^{(1)}, b_{p_2}^{(2)}, ..., b_{p_K}^{(K)}$ are given by:

$$t_{1_{p_1},...,K_{p_K}} = a_{p_1}^{(1)} \cap a_{p_2}^{(2)} \cap ... \cap a_{p_K}^{(K)} \tag{4}$$

where $p_1 = 1, ..., C_1$, $p_2 = 1, ..., C_2$ and so on. Let $t$ be any $t_{1_{p_1},..,K_{p_K}}$.

## III. Lexical Representations

Suppose, in the user lexicon, the set of lexical symbols corresponding to non-relative feature category $j$ is $L^{(j)}$. These non-relative lexical symbol learning is described by the bijective functions $g_j$ such that

$$g_j : L^{(j)} \rightarrow A^{(j)} \tag{5}$$

Learning described by $g_j$ is achieved by the algorithm shown in Fig. 1.

However, the above relationship is not valid for learning relative lexical symbols such as "*big*" or "*small*" which are associated with relative features. Let the set of relative feature categories be $Q = \{q_1, q_2, ..., q_m, ..., q_M\}$. Assume that each $q_m$ relative feature category is associated with an ordering relation $O^{(m)}$. For example, relative feature category *size* may be associated with the ordering relation *number of pixels in an object*.

Let the set of lexical symbols corresponding to the relative feature category $m$ be $S^{(m)}$. If $t^{(m)}$ is a well-ordered set whose elements consist of the elements of $t$ which are ordered according to the ordering relation $O^{(m)}$, the learning of lexical symbols is described by the bijective function $g_s$ such that

$$g_s : S^{(m)} \rightarrow t^{(m)} \tag{6}$$

The algorithm shown in Fig. 1 would identify objects considering only non-relative features. Therefore, it will identify all the objects which differ only in relative features. For example, it will identify all "*red cubes*" irrespective of the presence of a "*small red cube*" and a "*big red cube*." The set of such objects form the set $t$. Therefore, learning described by $g_s$ is achieved through the algorithm shown in Fig. 2.

## IV. Overview of the System

Overview of the experimental system developed to demonstrate the above concept is shown in the Fig. 3. On the object table, objects of different colors, shapes and sizes are placed. Observing the objects a user may ask the robot to pick any one of the objects. For example, user may say "*pick the small red cube*". Valid grammar for this experiment is given in the Table I. Any combination of the lexical symbols size, color, and shape would form a valid reference to an object.

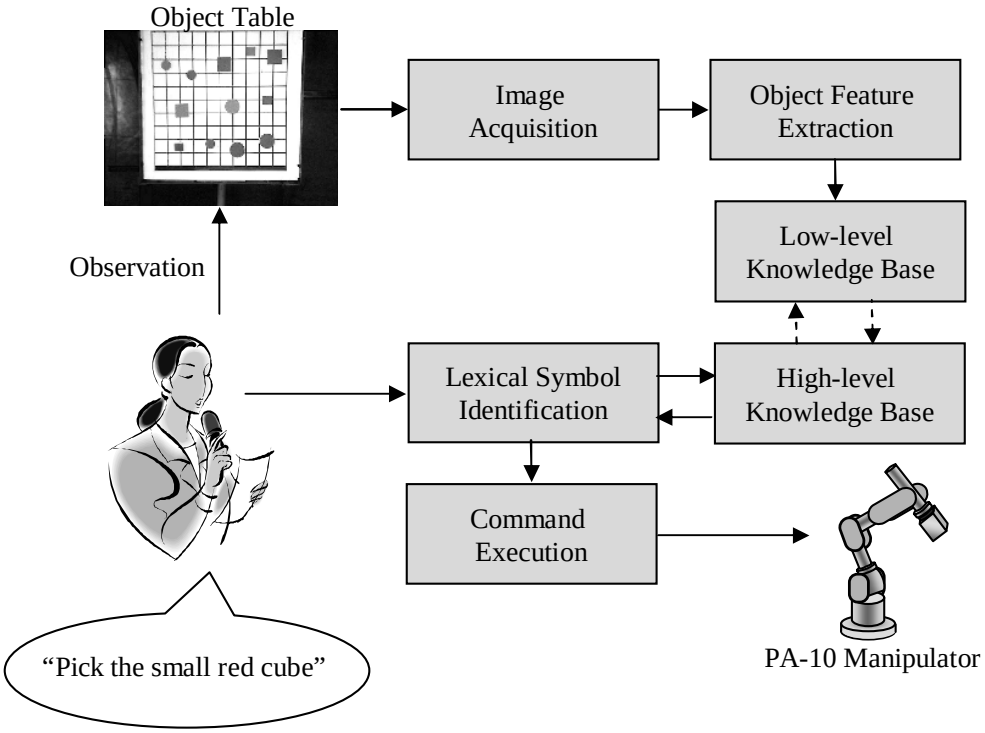


Fig. 3. Overview of object identification system by a robot.

## V. IMPLEMENTATION

### A. Experimental Setup

The experimental setup consists of a PA-10 industrial manipulator and controller, object table, three USB cameras, a microphone and a PC running WindowsXP. The three cameras are placed over, in front of, and on left of the object table. For the image acquisition DirectX technology is used. Voice recognition is performed using IBM ViaVoice SDK.

### B. Image acquisition

For image acquisition, three webcams are used. The camera placed right above the table provides a calibrated image and it is further processed in order to extract object features. All three images are displayed on the users computer monitor in order to provide three dimensional details of the workspace.

### C. Object feature extraction

Object feature extraction module in Fig. 3 extracts shape, color and size representations of each object.

**Shape representation:** Shape representation of an object should be invariant to change in size, to change in location and to rotation. Although there are various descriptors such as thinness ratio, shape elongation, spreading, compactness, etc. Hu descriptors has the particularity of being invariant to scale, translation and rotation [9].

For a 2 dimensional function $f(x,y)$, the moment of order $(p+q)$ is defined as:

$$m_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^p y^q f(x,y) dxdy \qquad (7)$$

for $p, q = 0, 1, 2, ...$

If $f(x,y)$ is piecewise continuous and has nonzero values only in a finite part of the $xy$-plane, moments of all orders exist, and the moment sequence $(m_{pq})$ is uniquely determined by $f(x,y)$. Conversely, $m_{pq}$ uniquely determines $f(x,y)$.

The central moments are defined as:

$$\mu_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-\overline{x})^p (y-\overline{y})^q f(x,y) dxdy \qquad (8)$$

where $\overline{x} = m_{10}/m_{00}$ and $\overline{y} = m_{01}/m_{00}$.

If $f(x,y)$ is a digital image, the Eq. 8 becomes:

$$\mu_{pq} = \sum_x \sum_y (x-\overline{x})^p (y-\overline{y})^q f(x,y) \qquad (9)$$

The normalized central moments are defined as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \qquad (10)$$

where $\gamma = \frac{p+q}{2} + 1$.

From the normalized moments of order up to three, it is possible to derive seven invariant moments or Hu descriptors. In this work, only first Hu descriptor, $\phi_1$ was used as the shape representation,

$$\phi_1 = \eta_{20} + \eta_{02} \qquad (11)$$

If more complicated and diverse shapes are used, more descriptors may be used to increase the representing accuracy.

**Size representation:** Number of pixels of an object is used as the size representation.

| | |
|---|---|
| 1 | Get the user command. |
| 2 | Extract relative lexical symbols, $s^{(1)}, s^{(2)}, ..., s^{(j)}, ..., s^{(M)}$; $s^{(j)} \in S^{(j)}$ |
| 3 | Get the set of objects identified by the non-relative lexical symbols: $t$ |
| 4 | Find the ordered sets: $t^{(1)}, t^{(2)}, ..., t^{(m)}, ..., t^{(M)}$ |
| 5 | IF $s^{(1)}$ is known |
| | $\vdots$ |
| $4+M$ | IF $s^{(M)}$ is known<br>• Identify the object. |
| $5+M$ | ELSE<br>• Consult the user: "*What do you mean by* $s^{(M)}$*?*".<br>• User points to any object with the feature described by $s^{(M)}$.<br>• Learn $s^{(M)}$: Map $s^{(M)}$ to the position of the pointed object in the ordering relation $O^{(M)}$.<br>• Go to step $4+M$. |
| | $\vdots$ |
| $4+2M$ | ELSE<br>• Consult the user: "*What do you mean by* $s^{(1)}$*?*".<br>• User points to any object with the feature described by $s^{(1)}$.<br>• Learn $s^{(1)}$: Map $s^{(1)}$ to the position of the pointed object in the ordering relation $O^{(1)}$.<br>• Go to step 5. |

Fig. 2. Learning relative lexical symbols.

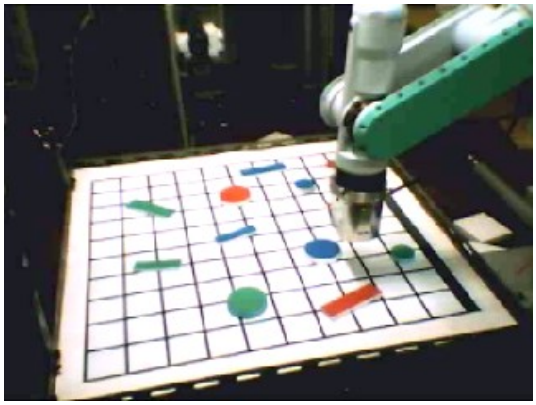

Fig. 4. A view of the experimental setup.

**Color representation:** Normalized red ($r$), green ($g$), and blue ($b$) components are used as the color representations.

$$r = \frac{R}{R+G+B} \tag{12}$$

$$g = \frac{G}{R+G+B} \tag{13}$$

$$b = \frac{B}{R+G+B} \tag{14}$$

where $R$, $G$, and $B$ are the components of an RGB color pixel.

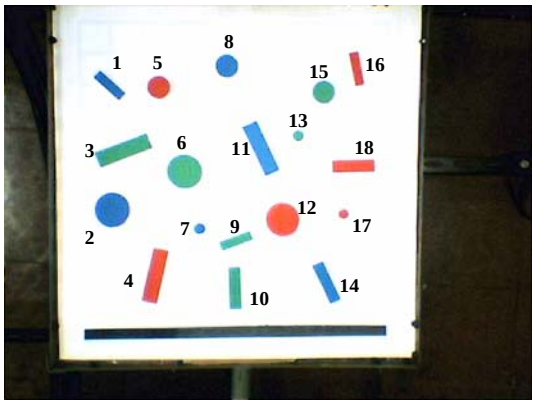


Fig. 5. Object table.

### D. Low level knowledge base

Three kind of representations found above are the elements of $\boldsymbol{r}_i$ in Eq. (1). All $r_i$'s (or $R$) are stored in the low level knowledge base. It is *low level* in the sense that it contains only sensory data without any lexical information. A portion of the content in the low-level knowledge base is shown in the Table II.

According to the discussion in the section II, color and shape are non-relative lexical symbols while size is a relative lexical symbol. Since these objects belong to finite number of colors and shapes, it should be possible to identify color and shape clusters within sensory data shown in the Table II as explained in the section II. If these clusters are correctly identified, the number of color clusters should be equal to the number of object colors and the number of shape clusters should be equal to the number of object shapes.

*1) Object clustering:* Object clustering is performed according to the non-relative features described in the section II. The number of clusters is not a priori known for both shape and color. Therefore, we have used a leader-follower algorithm to find clusters because it need not know the number of clusters in advance [10].

When defining

$w_i$ = current center for cluster $i$,
$\theta$ = threshold,
$x$ = a sample,

the algorithm is as follows:

**begin** **initialize** $\nu$, $\theta$
  $w_i \leftarrow x$
  **do** accept new $x$
    $j \leftarrow \arg\min_{j'} \|x - w_{j'}\|$ (find nearest cluster)
    **if** $\|x - w_j\| < \theta$
      **then** $w_j \leftarrow 0.5(w_j + x)$
      **else** add new $w \leftarrow x$
  **until** no more patterns
  **return** $w_1, w_2, ...$
**end**

TABLE II

OBJECT REPRESENTATIONS.

| Object No. | Pixels | First Hu Descriptor | Color ($r$, $g$, $b$) | Center ($x, y$)pixels |
|---|---|---|---|---|
| 1 | 2081 | 496 | 0.1067, 0.3281, 0.5652 | 255, 765 |
| 2 | 5744 | 206 | 0.1181, 0.3465, 0.5354 | 262, 461 |
| 3 | 5296 | 498 | 0.1897, 0.4545, 0.3557 | 288, 606 |
| 4 | 5067 | 494 | 0.6102, 0.2165, 0.1732 | 366, 301 |
| 5 | 2458 | 207 | 0.6024, 0.2362, 0.1614 | 375, 759 |
| 6 | 5467 | 207 | 0.2047, 0.4685, 0.3268 | 437, 555 |
| 7 | 540 | 209 | 0.1344, 0.3360, 0.5296 | 473, 415 |
| 8 | 2445 | 207 | 0.1660, 0.3241, 0.5099 | 540, 811 |
| 9 | 1698 | 508 | 0.1850, 0.4488, 0.3661 | 562, 386 |
| 10 | 2951 | 498 | 0.1732, 0.4331, 0.3937 | 559, 270 |
| 11 | 4917 | 505 | 0.1462, 0.3557, 0.4980 | 622, 611 |
| 12 | 5171 | 206 | 0.6181, 0.2362, 0.1457 | 675, 438 |
| 13 | 507 | 207 | 0.1700, 0.4348, 0.3953 | 713, 642 |
| 14 | 2910 | 517 | 0.1024, 0.3386, 0.5591 | 781, 284 |
| 15 | 2333 | 206 | 0.2087, 0.4409, 0.3504 | 774, 748 |
| 16 | 2990 | 503 | 0.6126, 0.2292, 0.1581 | 847, 568 |
| 17 | 418 | 209 | 0.4980, 0.2451, 0.2569 | 822, 451 |
| 18 | 1912 | 499 | 0.6220, 0.2165, 0.1614 | 854, 805 |

TABLE III

CLUSTERED OBJECTS.

| Shape Cluster | Object No. |
|---|---|
| 1 | 1, 3, 4, 9, 10, 11, 14,16, 18 |
| 2 | 2, 5, 6, 7, 8, 12, 13, 15, 17 |

| Color Cluster | Object No. |
|---|---|
| 1 | 1, 2, 7, 8, 11, 14 |
| 2 | 3, 6, 9, 10, 13, 15 |
| 3 | 4, 5, 12, 16, 17, 18 |

TABLE IV

LEXICAL SYMBOLS TO CLUSTER MAPPING

| lexical symbol | Shape cluster | Color cluster |
|---|---|---|
| cube | 1 | – |
| cylinder | 2 | – |
| red | – | 3 |
| green | – | 2 |
| blue | – | 1 |

For shape clustering, $\theta$ is taken to be 100. $x$ are the Hu moments given in the third column of the Table II. For color clustering, $\theta$ is taken to be 0.1. $x$ are the normalized $r, g, b$ vectors given by the fourth column.

### *E. High level knowledge base*

This is high-level in the sense that it contains lexical knowledge. This contains the mappings between lexical symbols and corresponding object features obtained from sensory data. Initially, this is empty. It is filled using the algorithms in the Fig. 1 and 2 as discussed in the section III.

TABLE V

OBJECTS OF SAME COLOR AND SHAPE

| Colored Object | Object Nos. |
|---|---|
| red cube | 4, 16, 18 |
| red cylinder | 5, 12, 17 |
| green cube | 3, 9, 10 |
| green cylinder | 6, 13, 15 |
| blue cube | 1, 11, 14 |
| blue cylinder | 2, 7, 8 |

## VI. RESULTS AND CONCLUSION

An image of the object table taken from the top camera is shown in the Fig. 5. The Table II shows a portion of object representations corresponding to object 1 to 18 contained in the low-level knowledge base. Table III shows objects clustered according to non-relative features, shape and color.

After learning with the algorithm given in Fig. 1, mapping between non-relative lexical symbols and the clusters mentioned above is shown in the Table IV. This mapping provides the result shown in the Table V. We can see that there are three objects of the same color and the shape. They should be identified with their relative features as explained in the section III. In this experiment there is one relative feature, size.

Final object identification result is shown in the Table VI.

In this paper, we have discussed the possibility of learning of object identification by robots commanded by natural language. The proposed concept was demonstrated with an object identification experiment using a PA-10 redundant manipulator. Users could command the robot to pick objects placed on a table using natural references like "*big red cube*," "*small blue cylinder*," etc.

To identify the referred objects, composite lexical item understanding system based on individual lexical symbol learning was presented.

In this implementation, relative small set of lexical symbols

TABLE VI

FINAL OBJECT IDENTIFICATION.

| Object No. | Lexical Representation |
|---|---|
| 1 | small blue cube |
| 2 | big blue cylinder |
| 3 | big green cube |
| 4 | big red cube |
| 5 | medium red cylinder |
| 6 | big green cylinder |
| 7 | small blue cylinder |
| 8 | medium blue cylinder |
| 9 | small green cube |
| 10 | medium green cube |
| 11 | big blue cube |
| 12 | big red cylinder |
| 13 | small green cylinder |
| 14 | medium blue cube |
| 15 | medium green cylinder |
| 16 | small red cube |
| 17 | small red cylinder |
| 18 | medium red cube |

was used. Incorporating more lexical symbols and study about their interpretation is a future work. On the other hand, here we have not considered the learning of actions. That too is a possible improvement that can be included in a future work.

## REFERENCES


[1] P. Menzel and F. D'Aluisio, *Robosapiens*, The MIT Press, England, 2000.

[2] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: challenges and results," *Robotics and Autonomous Systems*, vol. 42, no.3–4, pp. 271–281, 2003.

[3] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, pp. 143–166, 2003.

[4] A. J. Baerveldt, A vision system for object verification and localization based on local features, *Robotics and Autonomous Systems*, vol. 34, 83–92, 2001.

[5] D. Kragic, M. Bjorkman, H. I. Christensen, and J. O. Eklundh, Vision for robotic object manipulation in domestic setting, *Robotics and Autonomous Systems*, vol. 52, 85–100, 2005.

[6] N. Bredeche, Y. Chevaleyre, J. D. Zucker, A. Drogoul, and G. Sabah, A meta-learning approach to ground symbols from visual percepts, *Robotics and Autonomous Systems*, vol. 43, 149–162, 2003.

[7] P. Vogt, The physical symbol grounding problem, *Cognitive Systems Research*, vol. 3, 429–457, 2002.

[8] D. Roy, K. Y. Hsiao, N. Mavridis, Mental imagery for a conversational robot, *IEEE Trans. on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 34, No. 3, 1374–1383, 2004.

[9] M. K. Hu, Visual pattern recognition by moment invariants, *IRE Trans. of Info. Theory*, vol. 8, 179–187, 1962.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd edition. New York, NY: Wiley, 2004.