

# Towards Natural Human-Robot Interaction using Multimodal Cues

Technical Report, December 2011

Ranjini Swaminathan  
ranjinis@gmail.com

Mohan Sridharan  
mohan.sridharan@ttu.edu

## Abstract

Robust human-robot interaction in dynamic domains requires that the robot autonomously learn from sensory cues and adapt to unforeseen changes. However, the uncertainty associated with sensing and actuation on mobile robots makes autonomous operation a formidable challenge. This paper describes a novel framework for robots to incrementally learn object models and categorize objects based on multimodal sensory cues. The robot learns multimodal models composed of visual and verbal vocabularies to describe domain objects. The visual vocabulary consists of learned probabilistic models of object properties such as color, shape and size. Probabilistic graphical models and lexical tools are used to learn the verbal vocabulary consisting of object property labels and category labels that specify the relative importance of the objects. The robot also learns association models that enable the description of visual observations with words allowing for more natural human robot interaction. Furthermore, the robot uses the multimodal models to identify novel objects and augment existing object descriptions by posing natural language queries for human feedback.

## 1 Introduction

Enabling natural and robust human-robot interaction (HRI) in dynamic domains is an open research problem [13, 28]. Sophisticated sensory input processing algorithms have enabled the use of mobile robots in many real-world application domains [6, 15, 23, 29]. However, in domains characterized by partial observability, non-deterministic action outcomes and unforeseen changes, a robot equipped with multiple high-fidelity sensors still lacks the ability to robustly sense the environment and naturally interact with human participants. The sensory cues (e.g., vision and speech) are sensitive to environmental factors (e.g., illumination and background noise) and the information extracted by the processing algorithms is unreliable. Sustaining natural HRI requires the robot to understand and learn from context-dependent cues provided by human participants. Furthermore, the lack of time and expertise frequently makes it infeasible for humans to provide elaborate and accurate feedback—the robot needs to use human feedback judiciously as needed. Robust and natural HRI in dynamic domains

hence poses formidable challenges related to adaptive sensory processing, contextual information understanding and learning from human cues.

This paper describes a framework for natural HRI that enables a robot to use multimodal sensory cues to incrementally learn object models, categorize objects and acquire human inputs as needed. Specifically, the framework enables the robot to learn multimodal models of real-world objects based on visual and verbal vocabularies. The robot processes images to incrementally learn a vocabulary of object properties that is used to represent domain objects. The robot also analyzes verbal human inputs describing specific aspects of the scene to learn a more natural verbal vocabulary to describe domain objects. A learned association between the visual and verbal vocabularies enables the robot to provide natural descriptions of subsequent visual inputs and acquire human feedback as needed. The following key contributions are made in this paper:

- A probabilistic bootstrap learning approach is developed to enable the robot to incrementally populate the visual vocabulary with learned stochastic models of object properties such as color, shape and size.
- Probabilistic graphical models and lexical tools are used to learn a verbal vocabulary of object property labels and category labels that specify the relative importance of the objects.
- An association is learned between the visual and verbal vocabularies to enable the robot to map visual properties to words, identify ambiguities and augment existing object descriptions by posing natural queries for human feedback.

All algorithms are implemented and evaluated on a robot interacting with a human to describe objects in a tabletop scenario. The remainder of the paper is organized as follows. Section 2 summarizes related work to motivate the work described in this paper, while Section 3 describes the proposed multimodal learning framework. Experimental results are discussed in Section 4, followed by the conclusions in Section 5.

## 2 Related work

Sophisticated sensory input processing and decision-making algorithms have enabled the deployment of robots and software agents to interact with humans in different application domains [5, 15, 24]. For instance, the HUMAINE project aims to develop an integrated framework for emotion-oriented computing, and describe the emotional responses in human-machine interaction. Pineau et al. [24] developed a hierarchy of partially observable Markov decision processes (POMDPs) for behavior control on a robot operating as a nursing assistant at a hospital. Hoey et al. [15] used a similar POMDP hierarchy to develop a vision-based automatic monitoring and prompting system for people with dementia engaged in hand-washing. However, the hierarchy underlying these systems had to be manually specified.

There has been considerable work on cognitive architectures [1, 9, 20, 23] that build computational models to study (and understand) human-level reasoning, and to enable knowledge acquisition and reasoning on virtual agents and mobile robots. Large research consortia are focusing on cognitive HRI (e.g., CogX) [8, 9], where information obtained from different sensory cues (e.g., vision and speech) are bound together based on predetermined rules. However, many of these schemes are computationally expen-

sive, require manual encoding of a significant amount of domain knowledge, and lack proper schemes for inference when dealing with information associated with varying levels of uncertainty.

Many HRI schemes have focused on enabling robots to operate autonomously based on sensory inputs [4, 11], or to learn from extensive manual training and domain knowledge [2, 14]. Since dynamic domains make it challenging for a robot to operate without any human input and make it difficult for a human observer to provide elaborate feedback, researchers are focusing on enabling robots to acquire and use limited human input based on need and availability [26]. However, these methods do not model the unreliability of human inputs and require elaborate knowledge of the task and domain, limiting their use to simple simulated domains or specific tasks. There has also been considerable work on integrating multimodal cues within an appropriate architecture for HRI. For instance, Perzanowski et al. [23] modeled human-level communication to integrate gesture recognition and speech understanding within an architecture that enables multimodal HRI. More recently, Kennedy et al. [17] integrated computational cognitive models, multiple spatial representations and sensory cues (gestures and speech) to enable human-robot collaboration in a reconnaissance task. However, these approaches require manual encoding of considerable domain knowledge. In addition, adaptive sensory processing, contextual speech understanding and learning from human cues continue to be challenges for robust and natural HRI. The next section describes our framework that aims to address these challenges by learning multimodal associations between sensory cues.

### 3 Proposed Framework

The section describes the experimental scenario and the proposed framework that learns object descriptions from multimodal sensory cues, as shown in Figure 1.

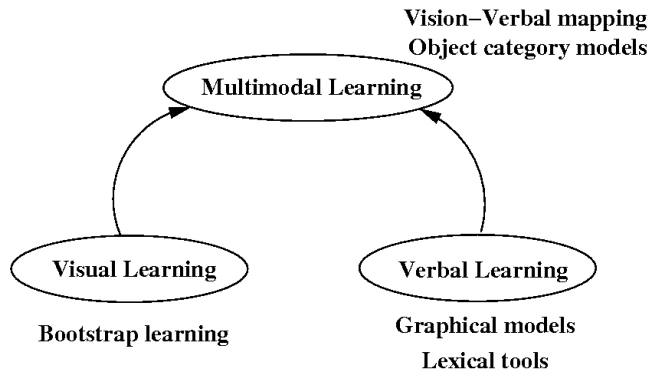


Figure 1: Overview of the multimodal learning framework.

### 3.1 Tabletop Scenario

The algorithms described in this paper are illustrated in a scenario where a human and a robot observe and describe objects placed on a tabletop. The scenario, though simplistic, encapsulates the challenges that we seek to address. Figures 2(a)–2(d) show some examples of the candidate objects. The objects are characterized by different properties such as *color*, *shape* and *size*.

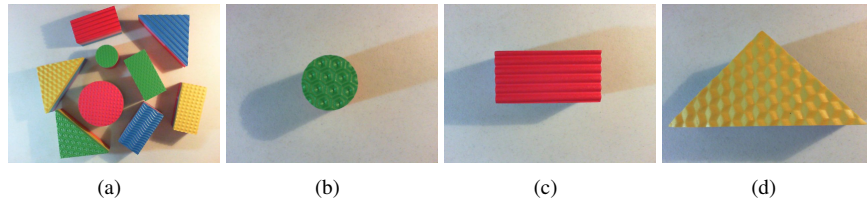


Figure 2: Images of the tabletop scenario and sample objects.

The robot learns incrementally and simultaneously from the sensory cues that consist of images from a color camera and verbal cues from a human observer. Section 3.2 describes the probabilistic bootstrap learning algorithm that enables the robot to incrementally learn a visual vocabulary of relevant object properties such as color, shape and size. Regions of interest (ROIs) in subsequent images result in probability distributions over the property class labels. A human participant observing the scene in Figure 2(b) may provide a verbal input of the form: “*the small green circle is not typical*”. Section 3.3 describes the graphical models and lexical tools that process the text corresponding to this input to extract candidate property labels (e.g., *small*, *green* and *circle*) and category labels (e.g., *not standard*). Finally, Section 3.4 describes a scheme to associate a natural verbal vocabulary with the visual vocabulary, enabling the robot to provide natural labels to subsequent sensory cues. In addition, feature vectors consisting of probability distributions over visual and verbal vocabularies (i.e., class labels) are used to learn a mapping to object category labels. In this paper, objects are characterized by three properties: *color*, *shape* and *size*; and two categories: *normal*, i.e., typical, and *suspicious*, i.e., needs investigation. For ease of explanation, the description below assumes that the robot first learns the vocabularies and then classifies test objects—learning can however be continuous. We also assume that objects are viewed sequentially during learning, which makes it easier to establish correspondence between multimodal cues.

### 3.2 Visual features

This section describes the use of visual cues to learn the visual vocabulary composed of models of object properties.

### 3.2.1 Visual Property Descriptions

An object in the tabletop scenario is characterized by color, shape and size properties. In order to learn a visual vocabulary to describe objects, the robot hence needs an approach to use images of different objects to learn models of object properties.

Consider the situation where salient regions of interest (ROIs) corresponding to objects have been extracted from an image, and consider a single ROI. Each ROI pixel  $\mathbf{m}$  is a point in a three-dimensional color space, i.e., a vector  $\langle m_1, m_2, m_3 \rangle$  of values along the color channels (e.g., RGB). In this paper, the color property of an object is modeled as a distribution of the corresponding image pixels in the RGB color space. A disjunctive representation is used to model color distributions as a Gaussian mixture model (GMM) [3] or a 3D histogram:

$$p(\mathbf{m}) \sim \sum_{i=1}^K w_i \mathcal{N}(\mu_i, \Sigma_i) \quad \text{or} \quad \equiv \frac{\text{hist}(b_1, b_2, b_3)}{\text{hist}} \quad (1)$$

where the color distribution is modeled as the weighted sum of  $K$  Gaussians or a 3D histogram in color space. The parameters of the GMM, i.e.,  $w_i, \mu_i, \Sigma_i$  and  $K$ , are computed by Expectation-Maximization [3] on the image pixels being considered. To build a 3D histogram, the range of pixels values along the three channels are grouped into bins, and  $(b_1, b_2, b_3)$  are the bin indices corresponding to the color values  $\mathbf{m} = (m_1, m_2, m_3)$ . The histogram is normalized to obtain a probability distribution. The disjunctive representation provides a good trade-off between ease of representation and computational efficiency. A GMM or a histogram can be learned from one or more images (of the same object) and represents an entry in the visual vocabulary.

The shape property seeks to capture the external contour of the object in the ROI. The pixels corresponding to the boundary of the object in the ROI are therefore extracted and the contour is modeled using the seven Hu invariant moments [9, 16]:

$$\begin{aligned} sm_1 &= \eta_{20} + \eta_{02} \\ sm_2 &= (\eta_{20} + \eta_{02})^2 + (2\eta_{11})^2 \\ sm_3 &= (\eta_{30} + 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ sm_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ sm_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ &\quad (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ sm_6 &= (\eta_{20} - \eta_{02})[\eta_{30} + \eta_{12}^2 - (\eta_{21} + \eta_{03})^2] + \\ &\quad 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ sm_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - \\ &\quad (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \quad (2)$$

where  $\eta_{ij} = \mu_{ij} / \mu_{00}^{1 + \frac{i+j}{2}}$ . These moments  $\{sm_i, i \in [1, 7]\}$  are robust to changes in image scale, rotation, translation and reflection, e.g., the first moment ( $sm_1$ ) is similar to the moment of inertia around the ROI's centroid and the seventh moment is skew

invariant. Each unique shape description represents an entry in the visual vocabulary, which can be used to model different shapes.

Finally, the size property measures the relative size of the object in the image. This is represented by computing the number of pixels within the ROI under consideration and dividing it by the total number of pixels in the image, creating a unique size description entry in the visual vocabulary.

### 3.2.2 Bootstrap Learning and Matching

Given the visual feature descriptions described above, this section describes the bootstrap learning approach for autonomously and incrementally learning the unique models of visual properties to populate the visual vocabulary.

Consider the incremental learning of color-based entries in the visual vocabulary. For ease of explanation, assume that  $N$  unique entries have been learned for color distribution-based descriptions, i.e., there are  $N$  color property classes:  $C_i, i \in [1, N]$ . Let the robot now process the ROI in a new image. As described in Section 3.2.1, Equation 1 is used to learn a model  $p_{new}(\mathbf{m})$  of the color distribution in the ROI, as a GMM or a histogram. The robot compares this learned color distribution model with the existing  $N$  unique color description models. For the GMM, this comparison measures the degree of overlap between  $p_{new}(\mathbf{m})$  and  $p_j(\mathbf{m}), j \in [1, N]$ . For histogram models, the distance between the new distribution and the existing distributions can be measured using the Jensen-Shannon measure:

$$JS(\mathbf{a}, \mathbf{b}) = \frac{KL(\mathbf{a}, \mathbf{m}) + KL(\mathbf{b}, \mathbf{m})}{2} \quad (3)$$

$$KL(\mathbf{a}, \mathbf{b}) = \sum_i (\mathbf{a}_i \cdot \ln \frac{\mathbf{a}_i}{\mathbf{b}_i}), \quad \mathbf{m} = \frac{\mathbf{a} + \mathbf{b}}{2}$$

where  $\mathbf{a}$  and  $\mathbf{b}$  are the distributions to be compared and  $KL(\mathbf{a}, \mathbf{m})$  is the KL-divergence measure between distributions. This measure is robust to sudden spikes in the distributions [27]. If the new distribution is a close match with one of the existing color property descriptions, it is merged with the existing description using GMM-merging or histogram-merging techniques [27]. If a close match is not found, a new color property class created, i.e., a new entry is created in the visual vocabulary.

Next, assume that  $M$  shape property descriptions:  $Sh_j, j \in [1, M]$  have been learned. The shape description corresponding to the ROI extracted in a new image  $sm_{i,new}, i \in [1, 7]$  is compared with the existing  $M$  shape descriptions using a distance measure in the seven-dimensional space of moments—Equation 3. A new shape property class is created, i.e., a new entry is created in the visual vocabulary, if the observed shape is not a close match with any of the existing descriptions. Given that the ROI-size just measures the area of the image covered by the ROI, new size descriptions are created if the observed ROI size is more than 1.5 times larger or smaller than  $L$  existing size models ( $Sz_j, j \in [1, L]$ ). All size values within a range are merged together.

Assume that the robot has learned models for  $N, M,$  and  $L$  color, shape and size property classes in its visual vocabulary, using the bootstrap learning approach. *Any object is now described using the vocabulary entries.* A strategy is developed to match

the property distributions in an image with the learned property models, to obtain a feature vector that describes objects as a probability distribution over the vocabulary.

Consider the color distribution  $p_{test}$  extracted from a ROI in the test image. The similarity between this distribution and the learned color property models is computed as the degree of overlap with the existing GMMs or the JS distance (Equation 3) to the existing histograms. This distance is used to obtain a probability distribution over the color-based entries in the vocabulary, i.e., a  $N$ -dimensional vector of match probabilities:  $\langle mp_{c_1}, \dots, mp_{c_N} \rangle$ . Similar match probabilities are obtained by measuring the similarity between the shape property description  $sm_{i,test}$ ,  $i \in [1, 7]$  of the test image ROI with the  $M$  learned shape property models. The shape-based comparison provides an  $M$ -dimensional vector:  $\langle mp_{sh_1}, \dots, mp_{sh_M} \rangle$ . To obtain match probabilities based on ROI-size, the test image ROI size is compared on a linear scale between the largest and smallest learned size property models. The net visual feature vector for the test object is a combination of the individual match probability vectors over the visual vocabulary:

$$mp_{c_1}, \dots, mp_{c_N}, mp_{sh_1}, \dots, mp_{sh_M}, mp_{sz_1}, \dots, mp_{sz_L} \quad (4)$$

The robot learns associations between this feature vector and the verbal feature vectors of Section 3.3. The robot also uses the entropy in the match probability vectors to identify ambiguous test objects, e.g., for the color-based match probability vector:

$$\mathcal{H}(C) = - \sum_{i=1}^N mp_{c_i} \log(mp_{c_i}) \quad (5)$$

If the match between color properties of the object in the test image ROI and the learned color property models is ambiguous, this entropy measure will have a large value. Similar entropy measures are computed based on the shape and size properties, and the maximum of these entropies is used to determine the need for human inputs.

### 3.3 Verbal Features

As stated in Section 3.1, the verbal cues consist of transcripts of human descriptions of image ROIs, i.e., sentences of the form: “*the small red triangle looks quite standard*”. The robot is also given the dictionary of object category labels: *normal* and *suspicious*. The verbal vocabulary for describing objects is the set of labels for object properties such as *color*, *shape* and *size*, i.e., a dictionary of labels such as *red*, *green*, *circle*, *triangle* and *large*. This vocabulary is learned by isolating words or phrases in the verbal inputs which are in the form of annotations. Verbal features are then learned by computing the semantic interpretation of each of these verbal properties. Next, associations are learned between visual and verbal vocabularies to describe each object as a feature vector. A set of feature vectors, along with the category labels, are used to build a joint model that classifies objects as *normal* or *suspicious*.

#### 3.3.1 Verbal Property Descriptions

Extracting verbal vocabulary from verbal cues involves tagging individual words or phrases with object properties. The property labels are *COL*, *SIZ*, *SHA*, *COM* for *color*,

*size*, *shape* and *comment* where *comment* refers to the category labels in the text. The tagging is done according to the IOB2 convention [25] where the *B*, *I* and *O* are used to indicate that a word is at the beginning, inside or outside of a property label. For instance, the three words *looks quite normal* that represent the comment property in a sentence, are tagged *B\_COM*, *I\_COM*, *I\_COM* and all words that do not correspond to any specific property are tagged *O*.

Part of Speech Tag	<b>DT</b>	<b>JJ</b>	<b>JJ</b>	<b>NN</b>	<b>VBZ</b>	<b>RB</b>	<b>JJ</b>
Sentence	<i>This</i>	<i>small</i>	<i>red</i>	<i>triangle</i>	<i>looks</i>	<i>quite</i>	<i>standard</i>
Object Property Tag	<b>O</b>	<b>B_SIZ</b>	<b>B_COL</b>	<b>B_SHA</b>	<b>B_COM</b>	<b>I_COM</b>	<b>I_COM</b>

Figure 3: Part of speech and object property tags for words in a sentence.

In addition to the property tags, *Part of Speech* tags (POS tags) of the individual words in the annotated data are generated automatically using the Stanford Log-linear POS Tagger [30] with the tags belonging to the Penn Treebank tag set [21]. Common POS tags are *noun*, *adjective*, *verb*, *adverb*, *determiner*, which are denoted by: *NN*, *JJ*, *VBZ*, *RB*, *DT*. Figure 3 illustrates the assignment of POS and object property tags for a sample sentence.

The property tags and POS tags are used to learn a Conditional Random Field (CRF) [19, 3, 18] that can tag new annotations with their verbal property tags. A CRF is a partially directed graph whose nodes correspond to  $\mathbf{Y} \cup \mathbf{X}$  where  $\mathbf{Y}$  is a set of target variables and  $\mathbf{X}$  is a set of observed variables. The graph is parameterized as a set of factors,  $\phi_1(\mathbf{D}_1), \dots, \phi_m(\mathbf{D}_m)$  in the same way as a Markov network. However, rather than encoding the distribution  $P(\mathbf{Y}, \mathbf{X})$ , the network encodes a conditional distribution as follows:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \tilde{P}(\mathbf{Y}, \mathbf{X}) \quad (6)$$

$$\tilde{P}(\mathbf{Y}, \mathbf{X}) = \prod_{i=1}^m \phi_i(\mathbf{D}_i)$$

$$Z(\mathbf{X}) = \sum_{\mathbf{Y}} \tilde{P}(\mathbf{Y}, \mathbf{X})$$

The POS tags and words are modeled as the observed variables ( $\mathbf{X}$ ) and the object property tags *SIZ*, *SHA*, *COL*, *COM* are the target variables ( $\mathbf{Y}$ ) used for tagging the words in the annotation. The CRF model is learned using the CRF Toolkit [10]. A small set of annotated sentences are manually labeled with verbal property tags and provided as input to the toolkit along with the corresponding words and POS tags. However, the CRF model bootstraps off of the available information (similar to the visual vocabulary). As property tags are identified in subsequent sentences, they are used as automatically labeled training data. The candidate verbal tags are processed



using a lexical tool (WordNet[22], see below) to identify verbal vocabulary entries, i.e., items to be placed in the verbal dictionary. The verbal dictionary will therefore consist of entries (i.e., labels) for color, shape and size.

### 3.3.2 Learning Semantic Interpretations

Given the verbal vocabulary entries, the semantic content of the words is then extracted to generate verbal features corresponding to any object, using a lexical database. Consider the color property. As stated above, the verbal vocabulary for color consists of entries such as *red*, *blue*, *green*, *yellow*. WordNet [22, 12], is a large lexical database of English with words grouped into *synsets* or cognitive synonyms. Each synset expresses a distinct concept and the words in the synsets are connected through different relationships such as *synonyms*, *antonyms*, *hypernyms* and *hyponyms*.

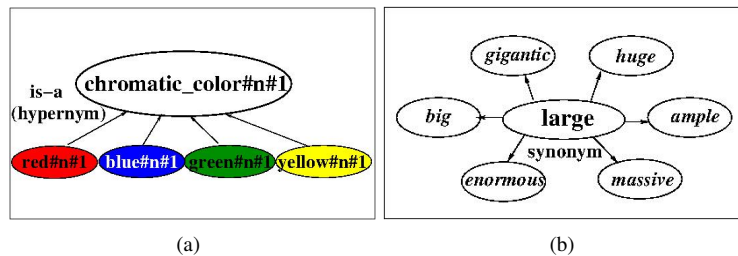


Figure 4: Using WordNet to compute semantic distances. The colors red, blue, green and yellow all have the same hypernym chromatic color and this is an example of a *is-a* relationship. The adjective large does not have a hypernym but instead has several synonyms and similar words, a few of which are shown in this figure.

When a word is tagged with an object property, the meaning of the word is expressed as a semantic distance with the different possible dictionary values for the property. Color values such as red, blue and green have the same hypernym *chromatic\_color* and this is an instance of a *is-a* relationship where red *is a* chromatic color. Once a word is identified as a color property in a new sentence, the semantic distance between the word and the possible entries for color in the dictionary are computed and this distribution represents the color-based verbal feature for this sentence. Similar semantic distances are computed by matching size and shape property tags in the sentence with the entries in the size and shape dictionaries (vocabulary). For instance, given the sentence: “this is a gigantic object”, a distribution will be obtained over learned size-based vocabulary entries (e.g., *small*, *medium*, *large*), and the best match is likely to be with: *large*. Figures 4(a)–4(b) show how WordNet is used to semantically link words in the annotation with synsets of words in the dictionary. Thus words in the annotated text are semantically interpreted using WordNet and the distributions over possible values of color, shape and size are computed. Similar to Equation 4, a verbal feature vector is generated as a distribution over verbal dictionary entries. However, unlike the visual features, a match is also obtained between words in the annotated text with the

given category labels, i.e., the verbal feature represents a sentence using a vocabulary composed of object properties and category labels.

### 3.4 Algorithm

---

#### Algorithm 1 Multimodal Learning and Inference

---

```

1: Learning Scheme:
   {Visual Learning}
2: for  $i = 1$  to  $N_{img}$  do
3:   Extract  $N_{sr,i}$  salient regions in image  $I_i$ .
4:   for  $j = 1$  to  $N_{sr,i}$  do
5:     Extract visual properties (color, shape, size) of  $ROI_j$ .
6:     Compute probability distribution of match with existing property classes.
7:     if detectNewObject() then
8:       Populate new vocabulary entries and obtain human input if necessary.
9:     else
10:      Merge with appropriate object property distributions.
11:    end if
12:  end for
13: end for
   {Verbal Learning}
14: Get  $N_{ver}$  sentences corresponding to human verbal descriptions.
15: for  $i = 1$  to  $N_{ver}$  do
16:   Extract verbal property tags and comment tags in  $ST_i$ .
17:   Compute distribution over dictionary entries for properties and comments.
18: end for
   {Multimodal mapping}
19: Extract co-occurrence patterns of visual and verbal descriptions of object properties.
20: Learn multimodal models of object properties and categories.

21: Classification/Inference:
22: for  $i = 1$  to  $N_{test}$  do
23:   Extract salient regions from  $I_i$ .
24:   for  $j = 1$  to  $N$  do
25:     Extract visual property distributions from  $ROI_j$ .
26:     Compute match probabilities with learned property classes.
27:     Use visual class probability distributions to obtain verbal class distributions.
28:     Classify feature vectors of visual and verbal class distributions to obtain object category labels.
29:     Draw attention or acquire human inputs for objects with ambiguous labels.
30:   end for
31: end for

```

---

We propose that combining visual and verbal descriptions of objects leads to more

natural human-robot interaction. Visual and verbal features are hence used to develop a joint model for classifying objects into one of two categories. Algorithm 1 shows the different stages involved in the learning and classification process. As stated earlier, the learning and classification can be continuous processes after an initial learning phase.

The visual vocabulary is learned as described in Section 3.2 from images of various objects (lines 2-13) in Algorithm 1. The vocabulary is populated incrementally by learning models of distributions of object properties such as color, shape and size. As property descriptions are learned from image ROIs, they are matched against existing descriptions, resulting in the creation of new vocabulary entries (line 8) or merging with the existing property descriptions. After some vocabulary entries are populated for the object properties, any object can be characterized as a feature vector, i.e., a distribution over the learned vocabulary entries—Equation 4.

The verbal vocabulary entries are learned next based on the verbal descriptions corresponding to the images used to learn the visual vocabulary (lines 14-18) in Algorithm 1. Some of the annotated text is generated manually to train a CRF model (Equation 6) that is used to identify candidate verbal dictionary items. A lexical tool is then used to identify the vocabulary entries corresponding to object properties. In addition to the property tags, the object category tags (corresponding to dictionary items: *normal* and *suspicious*) are also extracted.

Next, an association is learned between visual and verbal features (line 19 in Algorithm 1), i.e., the distributions over the visual vocabulary entries are mapped to the corresponding distributions over the verbal vocabulary entries. An object is now characterized using feature representations over visual and verbal vocabularies. If an annotation for an object has the word *red* in it, the corresponding visual feature distribution over the visual vocabulary’s color entries is associated with the word *red*. Each entry in the verbal dictionary corresponding to *color words* (described in Section 3.3.1) is hence mapped to one or more visual feature distributions. In other words, every property entry in the verbal dictionary is associated with a set of class distributions over the visual vocabulary. The visual-verbal association is very useful because different objects of the same property (e.g., color) may not have identical distributions over the visual classes. For instance, color distributions can vary based on illumination changes. However, different verbal property tags for the same color are likely to be clustered together, and hence the different visual feature distributions are likely to be annotated with the same label from the verbal dictionary. As far as the visual features are concerned, the object simply has a distribution over the set of property classes. It is only when computing the verbal features, that the actual name of the property and its semantic meaning are known.

The feature vectors (considered individually and together) and category tags are also used to learn a Support Vector Machine [3] classifier that can categorize subsequent visual and/or verbal features (line 20 in Algorithm 1).

The learned models and associations are used for classification (lines 21-30) in Algorithm 1. When a new object with no corresponding annotation is seen, its visual features are computed as before and mapped to appropriate labels from the verbal dictionaries for each property. This mapping is done by computing the distance between the visual feature distribution for the new object and the distributions associated with each verbal label, based on the Jensen-Shannon distance measure in Equ-

tion 3. For instance, the distance between the color-specific entries of the visual feature (of the test object) and the set of feature distributions associated with each color *red, blue, green, etc.* in the verbal space is computed. The object is then assigned the label of whichever set of distributions is closest to it. At the same time, if a good match is not found with any of the existing verbal classes (e.g., with Equation 5), human input can be requested by posing more natural queries. Once the verbal label is determined, the visual and verbal features are used in the joint classification model to determine object category labels, which can (once again) be used to pose natural language queries to solicit human feedback over “suspicious” objects.

## 4 Experimental Setup and Results

This section describes the experimental framework and results from the different experiments conducted. The data for the experiments consists of 40 objects in the tabletop scenario described in Section 3. Each input image consists of one object to simulate the extraction of ROIs from an image. The  $640 \times 480$  images are captured by a monocular color camera mounted on a wheeled robot. The objects are characterized using color, shape and size properties, and each object is unique in terms of at least one of the three properties. The objects used in the experiments had colors that mapped to four possible verbal dictionary terms: *red, blue, green* or *yellow*—the object colors included different shades of these colors. Object shapes (similarly) mapped to three verbal dictionary terms: *rectangle, circle* and *triangle* and sizes mapped to *small, medium* or *large*. As with colors, there were minor differences between objects that mapped to the same verbal shape. Note that there is mostly only one instance of each possible combination of the object property labels (e.g., large, green circle). Any characterization of an object (during learning or testing) hence has to occur by combining feature vectors learned independently. This approach simplifies the learning process but presents a considerable challenge when novel objects have to be labeled and classified during evaluation (see below).

Each object is also described by a sentence (verbal cue provided by a human) that provides information about each of the object properties (color, shape and size). In addition, the sentence includes an object category description (normal or suspicious). These object category labels are used for learning the SVM classifier and serve as ground truth during the evaluation of the learned classifier.

The following hypotheses were evaluated within the experimental framework: (I) visual vocabulary entries and visual features are learned successfully from objects in images; (II) semantic interpretations of object properties are learned successfully from verbal cues; (III) the association between visual and verbal vocabularies results in successful labeling of novel visual features; and (IV) multimodal models of object properties and categories are learned successfully, and used to label and categorize objects in test images.

All images and sentences in the dataset were considered to evaluate hypotheses I and II. Visual property descriptions were extracted from images as described in Section 3.2.1. New entries were added to the visual vocabulary for each of the object properties by bootstrapping off of the available information, as described in Section 3.2.2.

Multiple (20) experimental trials were conducted by presenting the objects in different sequential order. In each trial, the robot was able to successfully acquire the different color, shape and size class models to populate the visual vocabulary—in many cases, the robot learned a good model of each property class after observing just one instance of it. The learned vocabulary terms were also used to successfully generate visual features for objects as distributions over the visual vocabulary entries. Correct visual features were learned for an object even if its property labels had never been observed together, e.g., visual feature was computed for a *red triangle* even though the color (*red*) and shape (*triangle*) had never been observed together.

To evaluate hypothesis II, verbal vocabulary entries were learned from all the annotated sentences as described in Section 3.3. The CRF model was learned (and incrementally revised) and used in combination with WordNet to arrive at the correct verbal vocabulary entries. As with the visual vocabulary, the learning was successful when the sentences were presented in different sequential order. In addition, semantically similar cues were successfully grouped under the same verbal vocabulary entry.

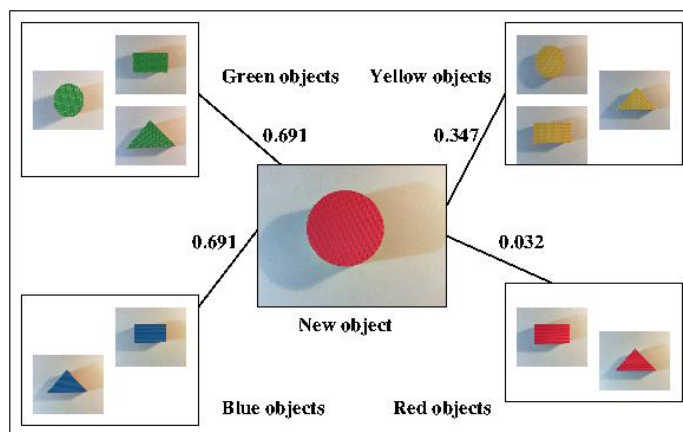


Figure 5: Mapping of visual features to verbal labels for the object’s color property. The distance is computed between the distribution over all color classes for the object in the center (a red circle) and those of the already labeled objects. The closest match is to the objects with color label *red*.

To evaluate hypothesis III, the mapping between visual and verbal features was done after learning the individual vocabularies. In this case, 50% of the available objects were presented sequentially to populate the vocabularies and generate some visual feature distributions corresponding to different verbal vocabulary entries, as described in Section 3.4. The other 50% of the images were used to generate the visual features that need verbal property labels. As stated in Algorithm 1, the distance is measured between the visual feature corresponding to an object with the visual feature distributions associated with the existing verbal property classes. The visual feature is assigned the label obtained by grouping the labels corresponding to the closest match obtained with each object property. Figure 5 shows an example of this labeling process based on just

Features used	Classification accuracy percentage
Visual+ Verbal + Category	97.5
Visual	75
Verbal	72.5
Visual + Verbal	77.5

Table 1: Accuracy results for object category classification. The results shown are for five-fold cross validation.

the color property. The feature distribution over the color classes for the object under consideration is closest in distance to the group of objects labeled with the color *red*. Similar performance was obtained over repeated trials. The feature vectors were also used successfully to identify instances where the test objects were not a close match with any of the existing property labels.

To evaluate hypothesis IV, the visual and verbal features for different objects, along with the verbal category labels, were used to learn a SVM classifier [7]. The learning and classification was done on the 40 objects with five-fold cross validation, and results are summarized in Table 1.

In these experiments, certain object property combinations were considered to be *suspicious* in the training data, e.g., *red rectangle* and *red circle* and evaluated the classifier’s ability to accurately detect it. The results show that the category features play an important role in the classification with the best accuracy results obtained when they are included. At this time, the text in the annotations consists of simple representations for category either directly implying a category or negating a category. For instance, the phrase *looks standard* has the word *standard* which indicates it is normal and the phrase *does not look standard* is an example of negation, which implies that the object is suspicious. We believe that more complex structures in the annotations may make inference and classification more difficult. Also, the category features provide absolute information, whereas the visual and verbal features have more variations since they are more complex representations of the object and are harder to learn. Yet another reason for the drop in performance when only the visual and verbal features are considered is that object property labels are independent of each other. This means that the object property labels for size, color and shape seen in objects in the training data need not be repeated in the exact combinations in the test data. For instance, a *small red rectangle* seen during training is not necessarily a part of the test set. The test set will contain objects that have one or even two of the same properties but not all (*large red rectangle*, *small red circle etc.*). The classifier thus has to learn the category labels for combinations based on the occurrences of the individual object property labels. This makes the classification task more challenging and is the reason for most of the classification errors reported in Table 1. The presence of comment features helps to improve classification performance by explicitly identifying the object property label combinations that are normal or suspicious. However, we do observe that combining the visual and verbal features results in better classification. Therefore a unified framework combining both visual and verbal cues gives better results for object description, modeling and classification.

## 5 Conclusions and Future Work

This paper describes a framework for robots to use multimodal cues to describe objects in the domain, resulting in more natural human-robot interaction. Images and human descriptions of objects were used to learn visual and verbal vocabularies of object properties. The learned vocabulary was then used to generate visual and verbal features as probability distributions over the vocabulary entries. The learning process is incremental, i.e, it allows for new object properties to be identified and modeled. Objects characterized by the multimodal features were also assigned labels corresponding to one of two categories. A set of (category) labeled multimodal feature vectors were used to learn a classifier that predicted category labels of novel objects. In addition, the proposed approach provided a mechanism to generate candidate verbal labels corresponding to novel visual features, which can be used to formulate natural language queries for human input, especially when the object is deemed suspicious.

Future work will include characterization of more complex objects with a richer, more sophisticated annotation vocabulary. Furthermore, natural language processing algorithms will be developed to formulate queries using the verbal feature descriptions of objects.

## References

- [1] V. Andronache and M. Scheutz. An Architecture Development Environment for Virtual and Robotic Agents. *Artificial Intelligence Tools*, 15(2):251–286, 2006.
- [2] B. Argall, S. Chernova, M. Veloso, and B. Browning. A Survey of Robot Learning from Demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.
- [3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2008.
- [4] M. Cakmak, N. DePalma, R. Arriaga, and A. Thomaz. Exploiting Social Partners in Robot Learning. *Autonomous Robots*, 29:309–329, 2010.
- [5] L. Canemero. The HUMAINE Project, 2010. <http://emotion-research.net/>.
- [6] J. Casper and R. R. Murphy. Human-robot Interactions during the Robot-assisted Urban Search and Rescue Response at the WTC. *IEEE Transactions on Systems, Man and Cybernetics*, 33(3):367–385, 2003.
- [7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [8] CogX. Cognitive Systems that Self-Understand and Self-Extend, 2011. <http://cogx.eu/>.

- [9] CoSy. Cognitive Systems for Cognitive Assistants, 2008. <http://www.cognitivesystems.org/>.
- [10] CRF++. CRF++:Yet Another CRF Tool Kit, 2010. <http://crfpp.sourceforge.net>.
- [11] J. Fasola and M. Mataric. Robot Motivator: Increasing User Enjoyment and Performance on a Physical/Cognitive Task. In *International Conference on Development and Learning*, Ann Arbor, USA, August 2010.
- [12] C. Fellbaum. *WordNet: An electronic lexical database*. The MIT press, 1998.
- [13] M. A. Goodrich and A. C. Schultz. Human-Robot Interaction: A Survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2007.
- [14] D. Grollman. *Teaching Old Dogs New Tricks: Incremental Multimap Regression for Interactive Robot Learning from Demonstration*. PhD thesis, Department of Computer Science, Brown University, 2010.
- [15] J. Hoey, P. Poupart, A. Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis. Automated Handwashing Assistance for Persons with Dementia using Video and a Partially Observable Markov Decision Process. *Computer Vision and Image Understanding*, 114(5):503–519, 2010.
- [16] M. Hu. Visual Pattern Recognition by Moment Invariants. *Transactions on Information Theory*, 8(2):179–187, 1962.
- [17] W. Kennedy, M. Bugajska, M. Marge, W. Adams, B. Fransen, D. Perzanowski, A. Schultz, and G. Trafton. Spatial Representation and Reasoning for Human-Robot Interaction. In *Twenty-Second Conference on Artificial Intelligence*, pages 1554–1559, Toronto, Canada, 2007.
- [18] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2010.
- [19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289. Citeseer, 2001.
- [20] P. Langley and D. Choi. An Unified Cognitive Architecture for Physical Agents. In *The Twenty-first National Conference on Artificial Intelligence (AAAI)*, 2006.
- [21] M. Marcus, M. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
- [22] G. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.



- [23] D. Perzanowski, A. Schultz, W. Adams, E. Marsh, and M. Bugajska. Building a Multimodal Human-Robot Interface. *IEEE Intelligent Systems*, 16(1):16–21, January-February 2001.
- [24] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun. Towards Robotic Assistants in Nursing Homes: Challenges and Results. *Robotics and Autonomous Systems, Special Issue on Socially Interactive Robots*, 42(3-4):271–281, 2003.
- [25] A. Ratnaparkhi. *Maximum entropy models for natural language ambiguity resolution*. PhD thesis, University of Pennsylvania, 1998.
- [26] S. Rosenthal, J. Biswas, and M. Veloso. An Effective Personal Mobile Robot Agent Through Symbiotic Human-Robot Interaction. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Toronto, Canada, May 2010.
- [27] M. Sridharan and P. Stone. Color Learning on a Mobile Robot: Towards Full Autonomy under Changing Illumination. In *The International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.
- [28] A. Tapus, M. Mataric, and B. Scassellati. The Grand Challenges in Socially Assistive Robotics. *Robotics and Automation Magazine, Special Issue on Grand Challenges in Robotics*, 14(1):35–42, March 2007.
- [29] S. Thrun. Stanley: The Robot that Won the DARPA Grand Challenge. *Field Robotics*, 23(9):661–692, 2006.
- [30] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *HLT-NAACL*, pages 252–259, 2003.