In *Human-Agent-Robot Teamwork Workshop at the International Conference on Human-Robot Interaction (HRI)* (HART 12), Boston, USA, March 5, 2012.

Towards Robust Human-Robot Interaction using Multimodal Cues

Ranjini Swaminathan and Mohan Sridharan
Department of Computer Science
Texas Tech University, USA
ranjinis@gmail.com, mohan.sridharan@ttu.edu

ABSTRACT

Real-world domains characterized by partial observability and non-determinism frequently make it difficult for a robot to operate without any human feedback. However, human participants are unlikely to have the time and expertise to provide elaborate and accurate feedback. The deployment of mobile robots to interact with humans in dynamic domains hence requires that the robot learn from multimodal sensory cues and high-level natural-language interactions with human participants. This paper describes a novel framework for robots to incrementally learn 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 applied on human verbal cues to populate the verbal vocabulary with object property labels and category labels that specify the relative importance of 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 object descriptions by posing natural language queries for human feedback.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

General Terms

Algorithms, Experimentation

Keywords

Robotics::Intelligence for human-robot interaction, Robotics:: Machine learning for robotics

1. INTRODUCTION

Enabling robust human-robot interaction (HRI) in dynamic domains is an open research problem [13, 29]. Although sophisticated sensory input processing algorithms have enabled the use of mobile robots in many application domains [6, 15, 23], robots still lack the ability to robustly sense the environment and interact with human participants in domains characterized by partial observability, non-deterministic action outcomes and unforeseen changes. The sensory cues (e.g., vision and speech) are sensitive to environmental factors (e.g., illumination and background noise) and the information extracted by sensory input processing algorithms is unreliable. In addition, the lack of time and expertise frequently makes it infeasible for humans to provide elaborate and accurate feedback. The robot hence needs to judiciously acquire and utilize context-dependent high-level verbal cues from human participants. Robust 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 robust 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 verbal vocabulary to describe domain objects. Learned associations between the visual and verbal vocabularies enable the robot to provide natural descriptions of subsequent visual inputs and acquire human feedback as needed. This paper hence makes the following key contributions:

- A probabilistic bootstrap learning algorithm enables the robot to incrementally learn a visual vocabulary of object properties such as color, shape and size.
- Probabilistic graphical models and lexical tools are used to learn a verbal vocabulary of labels that represent object properties and 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, and resolve identified ambiguities by posing natural queries for human feedback.

All algorithms are 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, while Section 3 describes the proposed multimodal learning framework. Experimental results are discussed in Section 4, followed by conclusions in Section 5.

2. RELATED WORK

Sophisticated sensory input processing and decision-making algorithms have enabled the deployment of robots and agents to interact with humans in many application domains [15, 24, 27]. For

instance, the HUMAINE project [27] seeks 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 nursing assistant at a hospital. Hoey et al. [15] also used a POMDP hierarchy to develop a vision-based 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 [2, 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 [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 expensive, 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 algorithms have focused on enabling robots to operate autonomously based on sensory inputs [5, 11], or extensive manual training and domain knowledge [3, 14]. Since dynamic domains make it difficult for a human observer to provide elaborate feedback, researchers are 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 real-world 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 for multimodal HRI. More recently, Kennedy et al. [17] integrated computational cognitive models, spatial representations and sensory cues (gestures and speech) for human-robot collaboration in a reconnaissance task, while Aboutalib and Veloso [1] used multiple visual and action cues for object recognition. However, adaptive sensory processing, speech understanding and learning from human cues continue to be challenges for robust HRI. Our framework seeks to address these challenges by enables a robot to learn multimodal associations between sensory cues, building rich object (and domain) descriptions that enable high-level object classification and natural-language interactions.

3. PROPOSED FRAMEWORK

This section describes the proposed framework that learns object descriptions from multimodal sensory cues, as shown in Figure 1. Section 3.2 describes the probabilistic bootstrap learning algorithm for incrementally populating the visual vocabulary, while Section 3.3 describes the algorithm based on graphical models and lexical tools to learn a verbal vocabulary of labels for object properties. Section 3.4 describes the algorithm for learning associations between these vocabularies and identifying situations where human input is necessary. Furthermore, we illustrate the use of the associations to learn a high-level classifier that predicts the relative importance of scene objects. We begin with a description of the experimental scenario.

3.1 Tabletop Scenario

The algorithms are illustrated in a scenario where a human and a robot observe and describe simple tabletop objects. The scenario, though simplistic, presents the challenges we seek to address, and

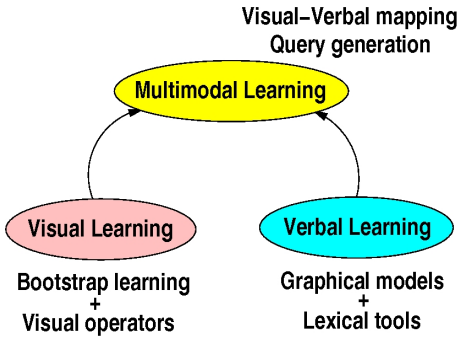


Figure 1: Overview of the multimodal learning framework.

enables us to isolate and analyze the effect of individual factors. The proposed algorithms are applicable to more complex domains. Figures 2(a)–2(d) show examples of the candidate objects, which are characterized by properties such as *color*, *shape* and *size*.

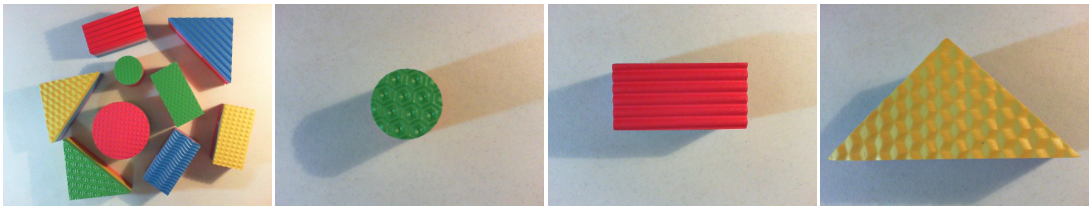
The robot learns incrementally and simultaneously from the sensory cues that consist of images from a color camera and verbal cues from a human observer. Section 3.2 describes the probabilistic bootstrap learning algorithm that enables the robot to incrementally learn a visual vocabulary of relevant 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"*. As described in Section 3.3, graphical models and lexical tools are used to 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 an algorithm to learn associations between the verbal and visual vocabularies, enabling the robot to provide natural-language 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 assumes that the robot first learns the vocabularies and then classifies test objects—learning can however be done continuously. To make it easier to establish correspondence between multimodal cues, we also (currently) assume that objects are viewed sequentially during learning.

3.2 Visual features

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

3.2.1 Visual Property Descriptions

To build a visual vocabulary to describe objects, the robot needs an approach to learn models of object properties from images. Consider an image where the salient regions of interest (ROIs) corresponding to objects have been extracted, 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 distribu-



(a) (b) (c) (d)

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

tions as a Gaussian mixture model (GMM) [4] or a 3D histogram:

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

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, μ_i, $\mathbf{\Sigma_i}$ and K, are computed by Expectation-Maximization [4] 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]. 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 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. 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:

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

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 [28]. 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 [28]. 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 2. 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 significantly (> 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* by matching 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 to the existing histograms (Equation 2). This measure is used to obtain a probability distribution over the color-based entries in the vocabulary, i.e., a N-dimensional vector of match probabilities: $\langle mp_{c_1}, \ldots, mp_{c_N} \rangle$. 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: $\langle mp_{sh_1}, \ldots, mp_{sh_M} \rangle$. 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:

$$mp_{c_1}, \ldots, mp_{c_N}, mp_{sh_1}, \ldots, mp_{sh_M}, mp_{sz_1}, \ldots, mp_{sz_L} \tag{3}$$

The robot learns associations between this feature vector and the

verbal feature vectors, as described in Section 3.4. The robot also uses the entropy in the match probability to identify ambiguous test objects, e.g., for the color-based match probability vector:

$$\mathcal{H}(C) = -\sum_{i=1}^{N} mp_{c_i} log(mp_{c_i}) \quad (4)$$

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 shape and size, 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 annotated. Verbal features are then learned by computing the semantic interpretation of each of these verbal properties.

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 *B*, *I* and *O* are used to indicate that a word is at the beginning, inside or outside 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*.

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) [4, 19, 18] that can tag new annotations with their verbal property tags. A CRF is a partially directed graph whose nodes correspond to $\mathbf{Y} \bigcup \mathbf{X}$ 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, $\phi_1(\mathbf{D}_1), ..., \phi_m(\mathbf{D}_m)$ 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:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})}\tilde{P}(\mathbf{Y}, \mathbf{X}) \quad (5)$$

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

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

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 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 using a lexical database to generate verbal features corresponding to any object. Consider the color property. As stated above, the verbal vocabulary for color will consist 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*.

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 3, 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 ($normal$ and $suspicious$), i.e., the verbal feature represents a sentence using a vocabulary composed of object properties and category labels.

3.4 Algorithm

We propose that combining visual and verbal descriptions of objects leads to more natural and robust human-robot interaction. Associations are hence learned between visual and verbal vocabularies to describe each object as a joint feature vector. In addition, a set of feature vectors, along with the category labels, are used to build a joint model that classifies objects as *normal* or *suspicious*. Algorithm 1 shows the different stages involved in learning and classification. As stated earlier, learning and classification can occur simultaneously and continuously 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. The learned property descriptions are matched with existing descriptions. Depending on the degree of match, the learned descriptions are merged with existing descriptions or new vocabulary entries are

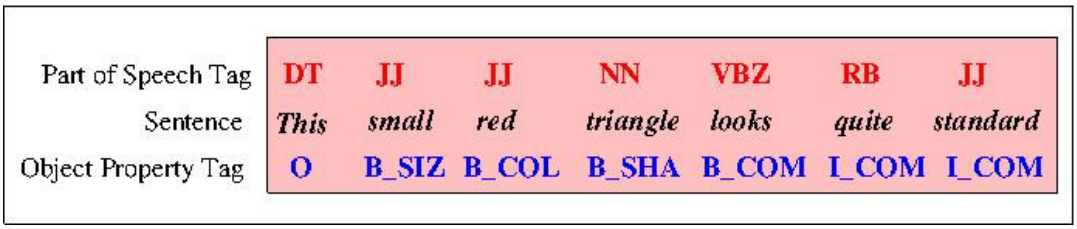
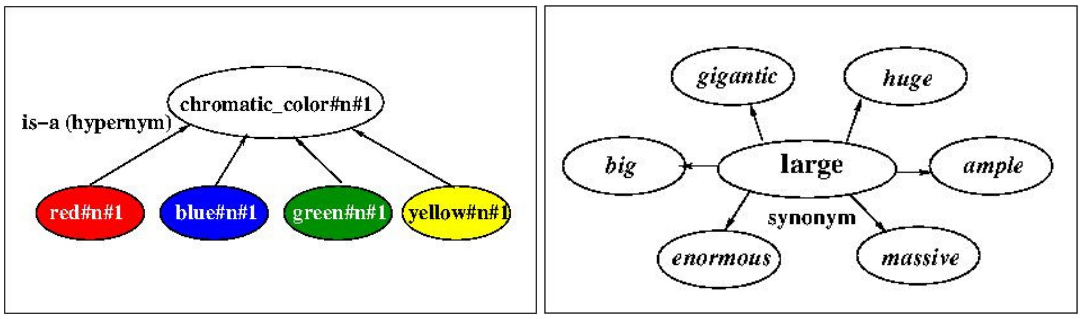


Figure 3: Part of speech and object property tags for words in a sentence.



(a) (b)

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.

created (line 8). After some vocabulary entries are learned for object properties, an object can be characterized as a feature vector, i.e., a distribution over the learned vocabulary entries—Equation 3.

The verbal vocabulary entries are learned based on the verbal descriptions corresponding to the images used to learn the visual vocabulary (lines 14-18). Some annotated text is generated manually to train a CRF model (Equation 5) that is used to identify candidate verbal dictionary items. A lexical tool is used to identify 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* (see Section 3.3.1) is hence mapped to one or more visual feature distributions, i.e., every property entry in the verbal dictionary is associated with a set of class distributions over the visual vocabulary. This visual-verbal association is useful because different objects with the same property (e.g., color) may not have identical distributions over the visual classes, e.g., due to illumination changes. However, different verbal property tags for the same color are likely to be clustered together, resulting in the corresponding visual feature distributions being annotated with the same verbal dictionary label. When the visual features are considered, the object only has a distribution over the visual property classes. The actual name of the property and its semantic meaning are known only when the verbal features are computed.

The feature vectors (considered individually and together) and category tags ($normal$ and $suspicious$) are also used to learn a Support Vector Machine [4] classifier (with radial basis functions) that classifies subsequent visual and/or verbal features (line 20).

The learned models and associations are used for classification (lines 21-30). When a new object without a verbal annotation is seen, its visual features are computed as 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, using the JS distance measure (Equation 2). 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 distributions that are most similar to it. If a good match is not found with any of existing verbal class (e.g., with Equation 4), human input can be requested by posing specific 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 setup and results. The data for the experiments consists of 40 objects in the tabletop scenario described in Section 3. Each input image consists of one object, resulting in the extraction of a single ROI—images in more complex scenes will result in multiple ROIs. The 640×480 images are captured by a monocular color camera on a wheeled robot. The objects are characterized using color, shape and size properties. The objects used in the experiments had colors that mapped to four

Algorithm 1 Multimodal Learning and Inference

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

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

verbal dictionary terms: *red*, *blue*, *green* or *yellow*—objects had 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 and size. There is typically only one instance of each possible combination of object property labels (e.g., large, green, circle). Any object can hence be characterized as a combination of feature vectors learned independently. This approach simplifies learning but presents a considerable challenge when novel objects have to be labeled and classified during evaluation.

Each object is also described by a sentence (human verbal cue) that provides information about object properties. The sentence also includes an object category description ($normal$ or $suspicious$). These category labels are used for learning the SVM classifier and serve as ground truth to evaluate the learned classifier.

The following hypotheses were evaluated: (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 that class. The learned vocabulary terms were also used to generate visual features for objects, i.e., distributions over the visual vocabulary entries. Visual features were learned correctly 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 annotated sentences, as described in Section 3.3. The CRF model was learned (and incrementally revised) and used with WordNet to obtain the correct verbal vocabulary entries. Learning was accomplished successfully when sentences were presented in different sequential order. In addition, semantically similar cues were grouped under the same verbal vocabulary entry.

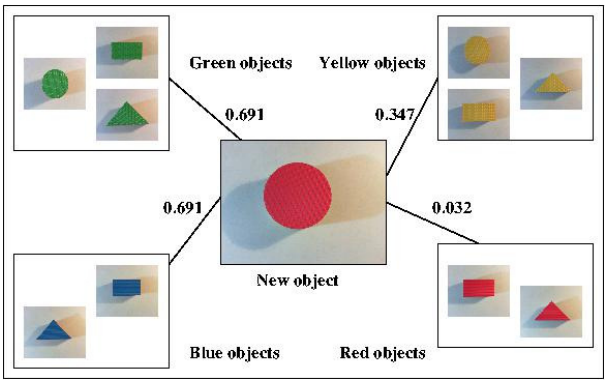


Figure 5: Mapping of visual features to verbal labels for the object's color property. The distance is computed between the distribution over all color classes for the object in the center (a red circle) and those of the already labeled objects. The closest match is to the objects with color label *red*.

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 and the corresponding verbal vocabulary entries. The remaining 50% of the images were used to generate visual features that need verbal property labels. As stated in Algorithm 1, the visual feature is assigned object property labels by grouping the labels corresponding to the closest match obtained for each object

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.

property. Figure 5 shows an example of this labeling process using just the color property. The feature distribution over color classes for the object under consideration is closest in distance to the group of objects labeled with color *red*. Similar performance was obtained over repeated trials. The feature vectors were also used to identify instances where the test objects did not closely match 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*—the classifier's ability to accurately detect these combinations was evaluated. The results show that the category features play an important role in the classification with the best results obtained when they are included. In these experiments, the text in the annotation either implies a category or negates 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 are more complex representations of the object that 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 learned independent of each other. The combinations of property labels (for size, color and shape) of objects in the test data are not necessarily seen in the training data. For instance, a *small red rectangle* seen during training is not necessarily in the test set. The test set may contain objects that have one or even two of the same properties but not all (e.g., *large red rectangle, small red circle* etc.). The classifier thus has to learn category labels for property combinations based on the occurrences of individual property labels. This makes classification more challenging and is the reason for most of the errors reported in Table 1. The presence of comment features helps 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. In addition, learning the individual properties separately helps provide partial labels for novel objects that share some properties with the objects seen before, resulting in more specific queries. Thus, a framework combining visual and verbal cues substantially improves the robot's ability to describe, model and classify domain objects.

5. CONCLUSIONS AND FUTURE WORK

This paper describes a framework for robots to exploit multimodal cues to learn rich descriptions of objects in the domain, resulting in more natural human-robot interaction. Images and human (verbal) descriptions of objects were used to learn visual and verbal vocabularies of object properties. The learned vocabularies were then used to generate visual and verbal feature vectors for objects in the form of probability distributions over the corresponding vocabulary entries. The learning process is incremental and autonomous, i.e, it allows for new object properties to be identified and modeled. In addition, associations are learned between the visual and verbal vocabularies to provide richer object descriptions. To illustrate the use of these associations, 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. The proposed approach also provides a mechanism to generate candidate verbal labels corresponding to novel visual features, which can be used to formulate specific natural language queries for human input.

Future work will include more complex objects with a richer, more sophisticated annotation vocabulary. In addition, natural language processing algorithms will be developed to formulate queries using the verbal feature descriptions of objects. Furthermore, correspondence between visual and verbal cues will be learned automatically. The long-term goal is to enable robust and natural human-robot interaction in complex real-world domains.

6. REFERENCES

[1] S. Aboutalib and M. Veloso. Multiple-Cue Object Recognition in Outside Datasets. In *International Conference on Intelligent Robots and Systems (IROS)*, Taipei, October 2010.

[2] V. Andronache and M. Scheutz. An Architecture Development Environment for Virtual and Robotic Agents. *Artificial Intelligence Tools*, 15(2):251–286, 2006.

[3] 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.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2008.

[5] M. Cakmak, N. DePalma, R. Arriaga, and A. Thomaz. Exploiting Social Partners in Robot Learning. *Autonomous Robots*, 29:309–329, 2010.

[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. Schroder and R. Cowie. *Developing a Consistent View on Emotion-oriented Computing*. Springer LNCS 3869, 2006.

[28] 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.

[29] 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.

[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.